ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

FORECASTING THE PERFORMANCE OF SHALE GAS WELLS USING MACHINE LEARNING

M.Sc. THESIS

Mohammed SHEDAIVA

Department of Petroleum and Natural Gas Engineering

Petroleum and Natural Gas Engineering Programme

AUGUST 2023

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

FORECASTING THE PERFORMANCE OF SHALE GAS WELLS USING MACHINE LEARNING

M.Sc. THESIS

Mohammed SHEDAIVA
(505201511)

Department of Petroleum and Natural Gas Engineering

Petroleum and Natural Gas Engineering Programme

Thesis Advisor: Assoc. Prof. Dr. Emre ARTUN

AUGUST 2023

ISTANBUL TEKNİK ÜNİVERSİTESİ ★ LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

MAKİNE ÖĞRENMESİ KULLANARAK ŞEYL GAZ KUYULARININ PERFORMANSININ TAHMİN EDİLMESİ

YÜKSEK LİSANS TEZİ

Mohammed SHEDAIVA
(505201511)

Petrol ve Doğal Gaz Mühendisliği Anabilim Dalı

Petrol ve Doğal Gaz Mühendisliği Programı

Tez Danışmanı: Doç. Dr. Emre ARTUN

AĞUSTOS 2023

Mohammed SHEDAIVA, a M.Sc. student of ITU Graduate School student ID 505201511, successfully defended the thesis entitled "FORECASTING THE PERFORMANCE OF SHALE GAS WELLS USING MACHINE LEARNING", which he/she prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor :	**Assoc. Prof. Dr. Emre ARTUN** Istanbul Technical University	.....
Jury Members :	**Assis. Prof. Dr. Burak KULGA** Istanbul Technical University	.....
	Assit. Prof. Dr. Betül YILDIRIM Middle East Technical University	.....

Date of Submission : 16 August 2023
Date of Defense : 24 August 2023

Dedicated to my spouse and son,
for their love and support.

FOREWORD

First and foremost, I am deeply grateful to my advisor Assoc. Prof. Dr. Emre ARTUN, for his generous guidance, unwavering support, and immense patience throughout my studies. His vast knowledge, invaluable suggestions, and significant contributions to this work have been instrumental in the success of this research.

I also appreciate my professors, friends, and committee members at the Department of Petroleum and Natural Gas Engineering for their contributions. Furthermore, heartfelt thanks to my mother, wife, and my son for their love and support throughout my studies.

August 2023 Mohammed SHEDAIVA

TABLE OF CONTENTS

Page

FOREWORD ..... **ix**
TABLE OF CONTENTS ..... **xi**
ABBREVIATIONS ..... **xiii**
SYMBOLS ..... **xv**
LIST OF TABLES ..... **xvii**
LIST OF FIGURES ..... **xix**
SUMMARY ..... **xxi**
ÖZET ..... **xxiii**
1. INTRODUCTION **1**
1.1 Thesis Structure 6
2. LITERATURE REVIEW **9**
2.1 Exploratory Data Analysis 10
2.2 Machine Learning Applications 13
2.3 Model Evaluation 16
2.4 Variable Importance 17
2.5 Summary 19
3. STATEMENT OF THE PROBLEM **21**
3.1 Objectives of the study 21
3.2 General Workflow 22
4. METHODOLOGY **25**
4.1 Dataset Collection 26
4.2 Exploratory Data Analysis 27
4.2.1 Univariate analysis 29
4.2.1.1 Measures of center 29
4.2.1.2 Measures of dispersion 31
4.2.1.3 Univariate data graphing 32
4.2.2 Bivariate analysis 35
4.2.2.1 Correlation and rank correlation 36
4.2.2.2 Bivariate data graphing 37
4.2.3 Multivariate analysis 39
4.3 Machine Learning-based Predictive Modeling 41
4.3.1 Simple linear regression 41
4.3.2 Multiple linear regression 44
4.3.2.1 Best subset selection 48
4.3.2.2 Best Model Selection 48
4.3.2.3 Goodness of fit 52
4.3.2.4 Regression assumptions and diagnostics 53
4.3.3 Tree-based methods 56
4.3.3.1 Classification and Regression Trees (CART) 56
4.3.3.2 Bagging 59
4.3.3.3 Random Forest (RF) 60
4.3.3.4 Extreme gradient boosting (XGBoost) 61

5. RESULTS AND DISCUSSION 63
5.1 Descriptive Statistics 63
5.1.1 Reservoir properties 63
5.1.2 Operational properties 64
5.2 Univariate Analysis 67
5.2.1 Reservoir parameters boxplots 67
5.2.2 Operational parameters boxplots 69
5.2.3 Performance metric boxplot 71
5.2.4 Reservoir parameters histograms 72
5.2.5 Performance metric histogram 72
5.2.6 Operational parameters histograms 73
5.3 Bivariate Analysis 76
5.3.1 Reservoir parameters scatterplots 76
5.3.2 Operational parameters scatterplots 78
5.3.3 Correlation coefficients test 80
5.4 Multivariate Analysis 82
5.5 Machine Learning-based Predictive Modeling 83
5.5.1 Multiple linear regression 83
5.5.2 Tree-based Methods 95
5.5.3 Variable Importance 102
5.5.4 Comparison of Results 104
6. CONCLUSIONS 105
6.1 Recommendations 106
REFERENCES 107
APPENDICES 115
APPENDIX A: Steps of developing a predictive model 117
APPENDIX A.1 - Best subset selection 117
APPENDIX A.2 – Extracting best variables 119
APPENDIX A.3 – Multiple linear regression 120
APPENDIX A.4 – Tree-based methods 121
APPENDIX B: Additional 3 Years Result Graphs 125
CURRICULUM VITAE 127

ABBREVIATIONS

AAE : Average Absolute Error
AI **:** Artificial Intelligence
AIC **:** Akaike Information Criterion
ANN : Artificial Neural Networks
BIC **:** Bayesian Information Criterion
CC : Correlation Coefficient
EDA **:** Exploratory Data Analysis
EUR **:** Estimated Ultimate Recovery
GAM **:** Generalized Additive Model
GBM **:** Gradient Boosting Machine
IoT **:** Internet-of-Things
IR 4.0 **:** 4th Industrial Revolution
KNN : K-Nearest Neighbors
ML **:** Machine Learning
MLR **:** Multiple Linear Regression
MMscf : Million Standard Cubic Feet
MSE : Mean Squared Error
OGIN **:** Original Gas in Place
OLS **:** Ordinary-Least-Squares
RF **:** Random Forest
RMSE : Root Mean Squared Error
RSE **:** Residual Standard Error
SVM : Support Vector Machine
Tcf **:** Trillion Cubic Feet
TRR **:** Technically Recoverable Resources
XGBoost **:** eXtreme Gradient Boosting

SYMBOLS

$\boldsymbol{\sigma_x}$	**:** Standard Deviation of X
$\boldsymbol{\sigma_y}$	**:** Standard Deviation of Y
$\boldsymbol{\sigma_{xy}}$	**:** Covariance
$\boldsymbol{\overline{X}}$	**:** Mean of X
$\boldsymbol{\overline{Y}}$	**:** Mean of Y
$\boldsymbol{N-1}$	**:** Degree of Freedom
$\boldsymbol{x_i}$	**:** Single Output for x
$\boldsymbol{y_i}$	**:** Single Output for y
$\boldsymbol{N}$	**:** Number of Observations
$\boldsymbol{d}$	**:** Difference of Two Ranks at Each Observation
$\boldsymbol{V[X]}$, $\boldsymbol{\sigma_x^2}$	**:** Variance
$\boldsymbol{X}$	**:** Random Variable
$\boldsymbol{f_i}$	**:** Relative Frequency

LIST OF TABLES

Page

Table 4. 1 : Dataset parameters 27
Table 5. 1 : Reservoir parameters descriptive statistics 65
Table 5. 2 : Operational parameters descriptive statistics 66
Table 5. 3 : CC of reservoir properties with the cumulative gas production 80
Table 5. 4 : CC of operational properties with cumulative gas production 81
Table 5. 5 : Overall dataset characteristics 84
Table 5. 6 : Summary of multiple linear regression model on the full dataset. 85
Table 5. 7 : Multiple linear regression with 13 variables 92
Table 5. 8 : Overall model results, SPE 2021 shale wells data 101

LIST OF FIGURES

Page

Figure 1. 1 : Cumulative gas and oil production from unconventional reservoirs in the US . Data from US Energy Information Agency (2017) as cited in (Zoback & Kohli, 2019) 2
Figure 1. 2 : Annual natural gas production in the US and CO_2 emissions from the electrical power sector. Data from the Energy Information Agency (2017) as cited in (Zoback & Kohli, 2019) 2
Figure 1. 3: Schematic of horizontally drilled well with a multi-stage hydraulic fracturing technique (Michael, 2021) 6
Figure 2. 1 : Matrix of scatterplot (Schuetter et al., 2018) 11
Figure 2. 2 : Histograms of predictors (Zhong et al., 2015) 12
Figure 2. 3 : Scatterplot matrix of predictors (Zhong et al., 2015). 12
Figure 2. 4: Model evaluation using Scatterplot (Schuetter et al., 2018) 17
Figure 2. 5: Feature importance by (a) GAM, (b) RF (Artun, 2020). 18
Figure 3. 1 : General workflow of a machine learning-based forecasting model. 23
Figure 4. 1 : Data analysis cycle (Mishra & Datta-Gupta, 2018) 26
Figure 4. 2 : The sequence of different data analysis approaches (Holdaway, 2014) 28
Figure 4. 3 : Location of mean, mode, and median for varying distributions (Mishra & Datta-Gupta, 2018) 30
Figure 4. 4 : Boxplot example (Wilcox, 2021). 33
Figure 4. 5 : Box-Whiskers plot with outliers (Holdaway, 2014). 33
Figure 4. 6 : Histogram example (Martinez et al., 2017) 34
Figure 4. 7 : Sensitivity of histograms to bins number. (Mishra & Datta-Gupta, 2018) 35
Figure 4. 8 : Scatterplots for collinearity (left), and outliers (right) (James et al., 2021) 38
Figure 4. 9 : Example of scatterplots with corresponding Pearson CC (Mishra & Datta-Gupta, 2018) 39
Figure 4. 10 : Correlation matrix (Wei & Simko, 2021). 40
Figure 4. 11 : Scatterplot matrix along with CC (Beck & Latif, 2022). 40
Figure 4. 12 : Simple regression fit (James et al., 2021) 42
Figure 4. 13 : Multiple linear regression model fit (James et al., 2021). 46
Figure 4. 14 : A schematic of the validation set approach (James et al., 2021). 50
Figure 4. 15 : A schematic of LOOCV (James et al., 2021) 51
Figure 4. 16 : k-Fold cross-validation representation (Schuetter et al., 2018) 52
Figure 4. 17 : Residuals vs fitted values (Kassambara, 2018) 53
Figure 4. 18 : Scale-location plot for homoscedasticity (Kassambara, 2018). 54
Figure 4. 19 : Normal Q-Q plot of residuals (Kassambara, 2018). 55
Figure 4. 20 : Residuals versus leverage plot (Kassambara, 2018). 56
Figure 4. 21: Decision tree example (Belyadi & Haghighat, 2021) 57
Figure 4. 22: Regression tree using rectangular regions (James et al., 2021). 57
Figure 4. 23 : RF modeling process (Mishra & Datta-Gupta, 2018) 60

Figure 4. 24 : XGBoost final prediction estimation (Chen & Guestrin, 2016). 61
Figure 5. 1 : Reservoir parameters boxplots..... 68
Figure 5. 2 : Operational parameters boxplots 70
Figure 5. 3 : Cumulative gas production boxplot..... 71
Figure 5. 4 : Cumulative gas production histogram 73
Figure 5. 5 : Histograms of reservoir parameters. 74
Figure 5. 6 : Operational parameters histograms 75
Figure 5. 7 : Reservoir parameters scatterplots 77
Figure 5. 8 : Operational parameters scatterplots..... 79
Figure 5. 9 : Correlation matrix..... 82
Figure 5. 10 : R-squared plot..... 86
Figure 5. 11 : RSS Plot..... 86
Figure 5. 12 : Adjusted R-squared plot 87
Figure 5. 13 : Cp plot 88
Figure 5. 14 : BIC plot 88
Figure 5. 15 : Validation set approach plot 90
Figure 5. 16 : Leave-one-out cross-validation plot 90
Figure 5. 17 : k-fold cross-validation plot..... 91
Figure 5. 18 : Observed vs predicted cumulative gas production for MLR model. ... 93
Figure 5. 19 : MLR diagnostic plots. 94
Figure 5. 20 : Regression tree for the full dataset. 95
Figure 5. 21 : Regression tree for training data. 96
Figure 5. 22 : Cross-validation for pruning the regression tree. 97
Figure 5. 23 : Pruned regression tree..... 97
Figure 5. 24 : Observed vs predicted cumulative gas production for regression tree. 98
Figure 5. 25 : Observed vs predicted cumulative gas production for bagging. 99
Figure 5. 26 : Observed vs predicted cumulative gas production for RF. 100
Figure 5. 27 : Observed vs predicted cumulative gas production for XGBoost. 100
Figure 5. 28 : Importance of variables in the random forest model 102
Figure 5. 29 : Relative influence for XGBoost model 103
Figure B. 1 : Regression tree model, 3 years-production..... 125
Figure B. 2 : RF variable importance, 3 years-production..... 125
Figure B. 3 : XGBoost relative influence, 3 years-production..... 126

FORECASTING THE PERFORMANCE OF SHALE GAS WELLS USING MACHINE LEARNING

SUMMARY

Utilization of real data to develop data-driven models in the petroleum industry has gained momentum in the past decade. The challenges related to modeling unconventional reservoirs has been recognized as the driving force behind this change in approach. Data-driven models help to enhance operations, increase efficiency and save time. In the meantime, several numerical reservoir simulators are used for modeling and forecasting the performance of shale gas wells. However, these models are computationally expensive and the simulators could indirectly face with difficulties in forecasting performance for the unconventional shale reservoirs comparing to conventional ones. This study employs a data analytics approach to investigate and gain understanding into the main driver parameters that influence the gas production performance in unconventional reservoirs (i.e. cumulative gas production after one year). The dataset utilized in this study is acquired from SPE Data Repository and consist of 53 wells (SPE, 2021). The study essentially utilized two primary methods, namely exploratory data analysis (EDA) and predictive data analytics modeling. Through the utilization of exploratory data analysis (EDA), the correlation between each reservoir and operational parameter with the cumulative gas production (Gp) is clearly identified. A number of reservoir and operational parameters display a strictly monotonic relationship with the gas production. Out of all variables, gas saturation was the variable, which demonstrated the strongest correlation. Furthermore, predictive data-analytics models based on statistical and machine learning algorithms were developed to forecast the cumulative gas production after 1 year. Among the five conducted models, extreme gradient boosting machine (XGBoost) proved to be the optimal technique for forecasting gas production, as it yielded the highest Coefficient of Determination (R^2) and the lowest Root Mean Square Error (RMSE). Finally, an analysis of variable importance was conducted to determine the key variables, which have the highest predictive power in forecasting gas production performance in unconventional shale reservoirs. The operational parameters such as the number of stages, lateral length and bottom perforation along with reservoir properties such as gas saturation, porosity and thickness are more dominant than the other reservoir and operational parameters. Gas saturation is the most critical parameter, which is considered as the key driver of forecasting the gas production. The findings of this study will be beneficial in the design and development similar forecasting modelling projects.

MAKİNE ÖĞRENMESİ KULLANARAK ŞEYL GAZ KUYULARININ PERFORMANSININ TAHMİN EDİLMESİ

ÖZET

Enerji talebinin artışı, son iki on yılda sıradışı şeyl rezervuarlarından üretimi teşvik eden itici güç olmuştur. Bu, gelişen yatay sondaj teknolojisi ve çok aşamalı hidrolik kırılma teknikleri ile birlikte gelmiştir (Kulga et al., 2017). Bu rezervuarlar, nanodarsilerle ölçülen ultra düşük geçirgenlikli kayalar olarak karakterize edilmektedir. Bu nedenle, sıradışı şeyl gibi karmaşık sistemlerde sayısal rezervuar simülatörleri gibi geleneksel rezervuar mühendisliği yaklaşımları, rezervuar analizi birçok farklı yönüyle zorluklar yaşamıştır. Bu, yavaşlamasına ve hatta ilerlemeyi zorlaştırmasına neden olmuştur. Ayrıca, hesaplama açısından verimsiz olabilir ve hatta daha pahalı olabilir, bu da sayısal modellere güvenmenin zorlaştığını gösterir (Ertekin ve Sun, 2019). Bu zorluklar, son on yılda ivme kazanan petrol endüstrisinde gerçek verileri kullanma yaklaşımındaki değişimin ardındaki itici güç olarak kabul edilmektedir. Veri tabanlı makine öğrenimi algoritmaları modelleri, rezervuar simülatörleri ile karşılaşılan zorluklardan kaçınmaya yardımcı olurken, hesaplama yükünü azaltma, verimliliği artırma, zaman tasarrufu sağlama ve hızlı tepki verme yeteneklerine sahiptirler (Schuetter et al., 2018). Bu, mühendislerin modelleme hedeflerine ulaşmasına ve hızlı kararlar almasına yardımcı olur (Kulga et al., 2017). Bu çalışma, yakın tarihli bir veri kümesi kullandı (SPE, 2021). Veri kümesi, Eagle Ford, Marcellus ve Haynesville şeyl oluşumlarından 53 şeyl kuyusunu içeriyor ve rezervuar ve operasyonel özellikler arasında dağıtılan 20 parametreyi içermektedir. Bu çalışmada, sıradışı rezervuarlardaki gaz üretim performansını etkileyen ana parametreleri incelemek ve anlamak için bir veri analitiği yaklaşımı kullanılmıştır (örneğin, 1 yıl sonrası gözlenen toplam gaz üretimi). Çalışma temelde iki ana tekniği kullanmıştır: keşifsel veri analizi (EDA) ve tahmin edici veri analitiği modellemesi. Keşifsel veri analizi, veri kümesinin yapısını incelemek ve veri hakkında daha derinlemesine bir anlayış elde etmek için kullanılmıştır. Ayrıca, değişkenler arasındaki ilişkileri daha iyi anlamaya katkı sağlamıştır. keşifsel veri analizi, tek değişkenli, çift değişkenli ve çok değişkenli analizi içeren üç aşamalı bir analizdir. Tek değişkenli analiz, veri kümesindeki her parametreyi, kutu grafikleri kullanarak bir boyutta (birikmiş gaz üretimi dahil) görsel olarak temsil etmek için kullanılır. Bu, değişkenleri tek tek, dağılım simetrisini ve dağılım simetrisini yanı sıra veri kümesindeki aykırı değerleri algılama konusunda yardımcı olur. Çift değişkenli analiz, birikmiş gaz üretimi ile her bir rezervuar ve operasyonel parametre arasındaki ilişkiyi araştırmak amacıyla uygulanmıştır. Bu bağlamda, Pearson ve Spearman korelasyon katsayıları, giriş ve yanıt değişkenleri arasındaki lineer ilişkinin gücünü değerlendirmek için kullanılmıştır. Ayrıca, çok değişkenli analiz, çok sayıda rezervuar ve operasyonel parametre arasındaki ilişkiyi ve parametreler arasındaki ilişkiyi belirlemeye yardımcı olan korelasyon matrisini kullanarak birçok parametre arasındaki korelasyonu belirlemeye yardımcı olmaktadır. Bu çalışmada kullanılan EDA, birikmiş gaz üretimi ile ilişkili olan parametreleri doğrulamanın temel adımıdır ve değişkenler arasındaki

ilişkileri ve her bir parametre ile birikmiş gaz üretimi arasındaki ilişkiyi anlamak için kullanılır. Model geliştirmek için kullanılan iki temel yöntem vardı: çoklu doğrusal regresyon (MLR) ve ağaç tabanlı yöntemler. Ağaç tabanlı yöntemler, regresyon ağacı, bagging, random forest (RF) ve ekstrem gradyan artırma makinesi (XGBoost) içeriyordu. Çoklu doğrusal regresyon, tüm tahmin edici değişkenlerin eğim katsayısına dayalı olarak hedef değişkenin tahminini yapar bir proxy modeldir. Sonuçlanan proxy denklem daha sonra görünmeyen veriler için birikmiş gaz üretimini tahmin etmek için kullanılabilir. Regresyon ağacı, basit ve oldukça yorumlanabilir bir yöntemdir. Veri kümesini, öncelikle en etkili tahmin ediciyi birincil bölünme (kök düğüm) olarak kullanarak böler. Daha sonra, veri ağaç üzerinde iki tarafın her iki tarafında daha az etkiye sahip karar düğümleri oluşturarak dallanır. Karar düğümleri, kök düğümünden hedef değişken üzerinde daha az etkiye sahiptir. Bu dallanma, daha fazla bölünme mümkün olduğu sürece devam eder; aksi takdirde, süreç bir terminal düğüme ulaşır. Bu terminal düğümler sonunda yanıt değişkeninin ortalama tahmin değerlerini temsil eder, bu bağlamda 1 yıl sonra birikmiş gaz üretimini temsil eder. XGBoost, algoritmaların optimizasyonunun mevcut olması nedeniyle ölçeklenebilir bir makine öğrenme yöntemidir. Bu nedenle, XGBoost modeli, kullanılan özel veri kümesine uygun olacak şekilde ayarlanabilir ve bu, aşırı uyum riskini önlemeye yardımcı olur. En iyi ayar parametrelerini tahmin etmek için çapraz doğrulama tekniği kullanılabilir. RF'den farklı olarak, regresyon ağacı paralel olarak inşa edilen ağaçlar yerine bağımlı bir modeldir, bu da her bir son tahmin edilen ağacın çıktısına bağlı olduğu anlamına gelir. Bu modelin son tahminleri, tüm bireysel ağaçların tahminlerini toplamak suretiyle hesaplanır. Kısacası, çoklu doğrusal regresyon (MLR), regresyon ağacı, bagging, RF ve XGBoost dahil olmak üzere beş istatistiksel ve makine öğrenimi yöntemi kullanılarak şeyl gaz kuyularının performansını doğru bir şekilde tahmin edebilen bir model geliştirmek için kullanılmıştır. Her bir model için veri kümesi %80 ve %20 olarak ikiye bölünmüş ve eğitim ve test için kullanılmıştır. Eğitim verileri modeli eğitmek için kullanılırken, test verileri modeli doğrulamak için kullanılmıştır. Bu çalışmada her bir modelin performansını değerlendirmek için üç yaygın metrik kullanılmıştır. Bu metrikler ortalama mutlak hata (AAE), kök ortalama kare hatası (RMSE) ve belirlilik katsayısı (R^2) şeklinde sıralanabilir. En düşük tahmin hatası ve en yüksek R2 değeri olan model seçilecektir. Bu çalışmada kullanılan son yöntem, şeyl gaz kuyularımızın performansı üzerinde en fazla etkisi olan parametreleri ve geliştirilen modeldeki en fazla tahmin yeteneğine sahip olan parametreleri belirlemek için kullanılan değişken önemliliği yaklaşımıdır. Bu çalışmanın sonuçları, birikmiş gaz üretimi ile simetrik bir dağılıma ve kesin bir monoton ilişkiye sahip bir dizi rezervuar ve operasyonel parametreyi göstermektedir, bunlar arasında gaz doygunluğu, su doygunluğu, gözeneklilik, rezervuar kalınlığı, lateral uzunluk, hidrolik kırılmaya ilişkin aşama sayısı ve alt perforasyon bulunmaktadır. Tüm değişkenler arasında gaz doygunluğu, en güçlü korelasyona sahip olan değişkendir. Ayrıca, 1 yıl sonra birikmiş gaz üretimini tahmin etmek için kullanılan beş model arasında, ekstrem gradyan artırma makinesi (XGBoost) en iyi sonuçları göstererek gaz üretimini tahmin etmek için en iyi teknik olduğunu kanıtlamıştır, çünkü en düşük Kök Ortalama Kare Hatası (RMSE) ve en yüksek Belirlilik Katsayısı (R^2) değerlerini vermiştir. XGBoost modeli, çapraz doğrulama yoluyla türetilen en iyi ayar parametrelerini kullandı, bu da tüm gözlemleri ve yalnızca kullanılabilir tahmin edicilerin %80'ini kullanmayı ve ağaçların maksimum üç düzeyine sahip olmayı içerir. Bu, kök düğüm ile terminal düğüm arasındaki düğüm sayısının en fazla üç bölünme olacak şekilde olduğu anlamına gelir. Ayrıca, sadece 150 adet artırma turu (iterasyon) kullanmayı önermiştir, bu da modelin daha fazla iterasiyon yapmayarak işlemi durdurur ve aşırı uyum riskini durdurmayı

önerir. Son olarak, sıradışı şeyl rezervuarlarda gaz üretim performansını tahmin etme konusunda en yüksek tahmin yeteneğine sahip olan anahtar değişkenler belirlenmiştir. XGBoost ve rastgele orman yöntemleri, şeyl kuyularının performansını etkileyen en üstteki 6 etkili parametre konusunda uyumludur. Bu parametreler, rezervuar ve operasyonel parametreler arasında dağılmıştır. Aşama sayısı, lateral uzunluk ve alt perforasyon gibi operasyonel parametrelerle gaz doygunluğu, gözeneklilik ve kalınlık gibi rezervuar parametreleri diğer rezervuar ve operasyonel parametrelerden daha baskındır. XGBoost ve RF tarafından yapılan genel değişken önemliliği yaklaşımı, ayrıca regresyon ağacının lateral uzunluğu ve aşama başına küme sayısını ana parametreler olarak kullanarak ağacı bölmek için ana parametreleri sıralama eğilimindedir. Bu sonuçlar, benzer çalışmaların bulgularıyla tutarlıdır (Schuetter et al., 2018). Gaz doygunluğu, birikmiş gaz üretimini tahmin etmek için en kritik olarak değerlendirilen temel bir parametredir. Bu çalışmanın bulguları, benzer tahmin modellemesi projelerinin geliştirilmesine katkı sağlayacaktır.

1. INTRODUCTION

The increase of energy demand has been the driving force behind the production from unconventional shale reservoirs during the past two decades. This came with the developing techniques such as horizontal drilling technology and the application of hydraulic fracturing in multiple stages (Kulga et al., 2017). These reservoirs are characterized as ultra-low permeability rocks - measured in nanodarcies, not millidarcies - that include several distinctive features, including high clay minerals, high organic content, fine clastic grains, and very low porosity. Unlike conventional gas reservoirs, which contain all free gas in the pore space, the unconventional shale reservoirs store the gas through a combination of compression (as free gas) and adsorption (either organic matter or minerals) onto the surface of the solid minerals (Esmaili & Mohaghegh, 2016; Guo et al., 2012). Thanks to the advanced technologies that makes the production of oil and gas from these systems economically feasible. Owing to the advancement technology in hydraulic fracturing and horizontal drilling, it is incontrovertible that unconventional reservoirs have produced an undeniable and significant amount of both natural gas and oil in USA. Figure 1.1 shows a five times increase in unconventional oil production and three times increase in unconventional gas production from approximately 100,000 horizontal wells drilled in the unconventional shale formations in the United States in just over a decade (Zoback & Kohli, 2019). Thus, unconventional reservoirs have revolutionized the oil and gas industry, enabling the country to achieve energy independence (Sahai, 2022). As well as the dramatic increase of natural gas production from unconventional resources has great impact on the reduction of greenhouse gas emissions since it is cleaner burning than other fossil fuel cousins like coal or petroleum (Speight, 2007). Moreover the dramatic increase of natural gas production in the United States from unconventional resources began from 2005 to 2006 where the CO_2 emissions began to drop (Figure 1.2) at about that time because of the fuel switching process from coal to natural gas to produce electrical power. Thus, the use of coal for electrical power generation dropped from 48.2% in the US to 33.4% while the use of gas increased from providing

21.4% of electricity to 32.5% in about a decade. As a result, the emissions of CO_2 fell about 15% to levels not observed for over 25 years (Zoback & Kohli, 2019).

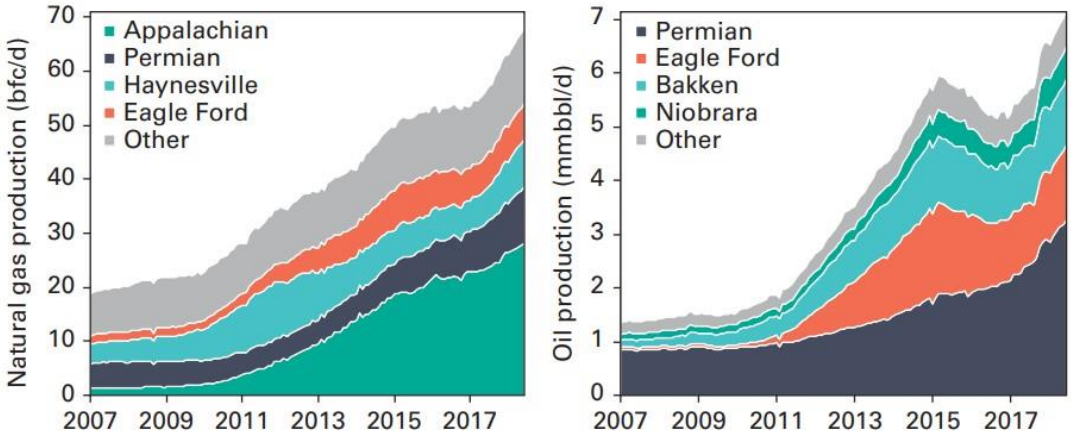


Figure 1. 1 : Cumulative gas and oil production from unconventional reservoirs in the US . Data from US Energy Information Agency (2017) as cited in (Zoback & Kohli, 2019)

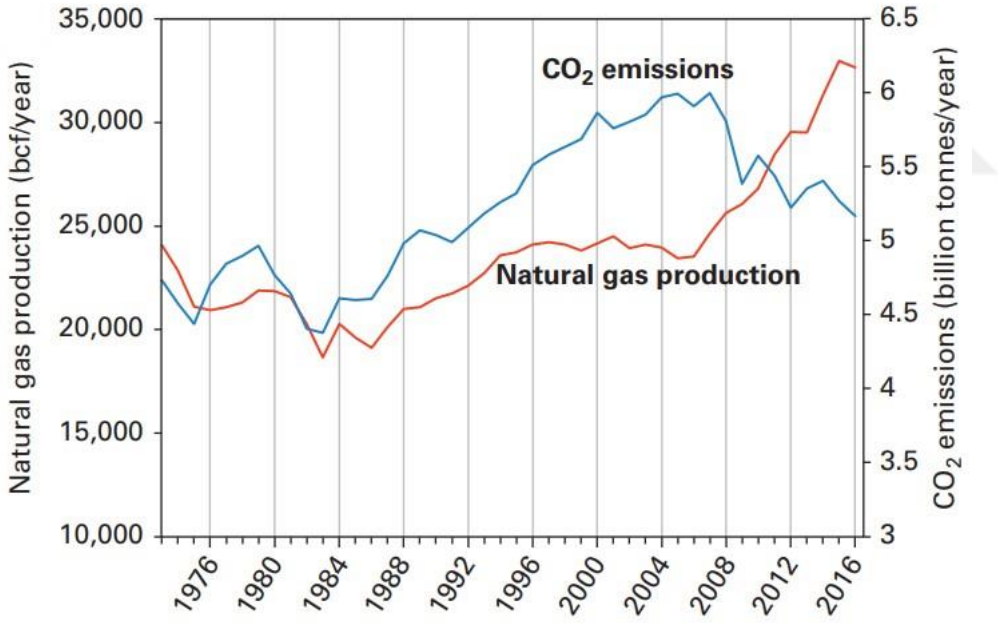


Figure 1. 2 : Annual natural gas production in the US and CO_2 emissions from the electrical power sector. Data from the Energy Information Agency (2017) as cited in (Zoback & Kohli, 2019)

The emerging of unconventional shale resources have transformed the oil and gas industry due to the abundance of natural gas production from it. However the real success that boom the global economy began in North Texas when Mitchell Energy Company between 1997 and 2002 combined horizontal drilling with the slick water fracturing in Barnett shale and this yielded increase of 250% in production from the

Barnett shale play (Bowker, 2007; NTNG, 2016). Since then energy companies started developing and pursuing other shale formations such as Marcellus, Haynesville, Eagle Ford, Woodford and others (U.S. Energy Information administration, 2011).

In this study, a forecasting model was developed using data collected from Marcellus, Haynesville, and Eagle Ford shale formations, some details about these shale formation are listed below.

Marcellus shale : The Marcellus shale gas play spans across the Eastern Region of the US and is considered to be one of the world's biggest and most prolific shale formations, covering a potential area of around 44,000 square miles (Zagorski et al., 2017) as cited in (Suriamin & Ko, 2022). The Marcellus shale has a depth varying from 4,000 to 8,500 feet. While it has thickness varying from 50 to 600 feet and the undeveloped technically recoverable resources (TRR) amount to 410.34 trillion cubic feet (Tcf) (U.S. Energy Information administration, 2011). Initial production rates of Marcellus shale have been reported to be between 0.5- 4 million cubic feet per day. Furthermore, it is classified as dry gas that contains relatively low levels of carbon dioxide and nitrogen thus the transportation through pipeline does not require natural gas liquid removal (Speight, 2020).

Haynesville shale : The Haynesville shale is a unique shale gas paly located in northwest Louisiana and east Texas, which possess numerous petrophysical, engineering, and mechanical properties that are close to optimal (Hammes & Gale, 2014). The entire area of Haynesville shale covers 9,000 square miles divided into 3,574 square miles active shale paly and 3,574 square miles undeveloped shale play. The Haynesville is known to be deep as the shale play depths ranges between 10,500 to 13,500 feet with average thickness of 250 feet (U.S. Energy Information administration, 2011). The Haynesville formation is considered one of the most productive shale gas play as the daily production through the use of developed horizontal drilling and multistage fracturing reached higher than 7 billion cubic feet per day (Bcf/d) in 2011 (Hammes & Gale, 2014). Thus, the Haynesville gas shale has a huge potential to serve as a considerable resource for the US, featuring an original gas in place (OGIP) place amounting to 717 Tcf and an estimated TRR of 251 Tcf (Speight, 2020).

Eagle Ford shale: Geographically, the Eagle Ford shale play covers an area of 20,000 square mile in south Texas and extends across the Western Gulf Coast Basin. Geologically, it is Cretaceous sediment in age which contains significant proportion of carbonate shale, typically as high as 70% (Railroad Commission of Texas, 2013). The Eagle ford shale production depth varying from 5,000 to 14,000 feet and it has thickness between 50 to 400 feet. Comparing to other shale plays, Eagle Ford is easy to be stimulated through hydraulically fracturing as it has low clay content and high carbonate content which represented in percentage as 10 to 20% clays and 65 to 75% carbonate. Furthermore, the formation demonstrated greater success comparing to other traditional shale plays due to its abundant production of not only dry gas, but also gas condensate, volatile oil, and black oil (Pope et al., 2012).

In this study, a real data from Marcellus, Eagle Ford and Haynesville shale is utilized to develop predictive data-analytics models using machine learning algorithms to predict the cumulative gas production after 1 year. The conventional reservoir engineering approaches such as numerical reservoir simulators in such complex systems like unconventional shale have experienced challenges in many different aspects of reservoir analysis. As it become slow and even making it difficult to move forward. Furthermore, it could be computationally ineffective and even more expensive making it difficult to rely in numerical models (Ertekin & Sun, 2019). Therefore, it was preferred to utilize machine learning algorithms for their fast response speeds as well as their capability to reduce the computation load (Schuetter et al., 2018) and developing a model that can be generalized well to new data of similar projects. Thus, engineers will be able to achieve the modeling goals and make decisions in a rapid way (Kulga et al., 2017).

Finally, Figure 1.3 can depict the horizontally drilling and multi-stage fracturing technique used in unconventional shale reservoirs. This figure would help in understanding the hydraulic fracturing regarding parameters used in this study such as number of clusters, stage spacing, number of clusters per stage, and number of stages. For better understanding as well, the full parameters used in this study will be briefly defined as follows:

Initial reservoir pressure: The pressure of the reservoir at the discovery stage (before production) and is typically measured in psi.

Reservoir temperature: The thermal state of the reservoir rock at a specific depth, typically measured in degrees Fahrenheit (°F).

Net Pay Thickness: The vertical extent of the reservoir rock that holds the hydrocarbons and thus play a role in determining the reserve capacity, typically measured in feet.

Porosity: The pore spaces in the reservoir rock that store the gas and fluids. It's usually expressed as a percentage.

Water and gas saturations: Is a measure to define how much of the rock pore space is filled with water or gas, typically expressed as a percentage.

Gas Specific Gravity: The ratio of gas density to the density of air at standard conditions to identify how heavy or light the gas is when compared to air.

CO_2 and NO_2: The concentrations of carbon dioxide and nitrogen in the reservoir.

True Vertical Depth: The distance from the surface to a specific point in the well or reservoir, usually measured in feet.

Stage Spacing: The distance between each stage in a hydraulically fractured horizontal well. It affects the reservoir productivity.

Stages: A stage refers to a section along the length of the horizontal wellbore that undergoes fracturing treatment.

Clusters: The individual perforation points that are created in the well casing to allow for the injection of fracking fluids into the reservoir.

Clusters per Stage: The number of perforation clusters created in a single hydraulic fracturing stage.

Proppant: Refers to solid material (sand or ceramic beads), mixed with fracturing fluids, and injected into a well to keep fractures open after the fracturing process.

Lateral Length: The horizontal distance of the wellbore through the reservoir rock, typically measured in feet.

Top Perforation and Bottom Perforation: Refers to the openings nearer to the start and end of the horizontal wellbore, respectively, and are created to allow the passage of fluids during hydraulic fracturing.

Sandface Temperature: The temperature of the reservoir rock at the point where it interfaces with the wellbore, usually measured in degrees Fahrenheit (°F).

Static Wellhead Temperature: The temperature of gas or fluids at the surface wellhead, typically measured in degrees Fahrenheit (°F).

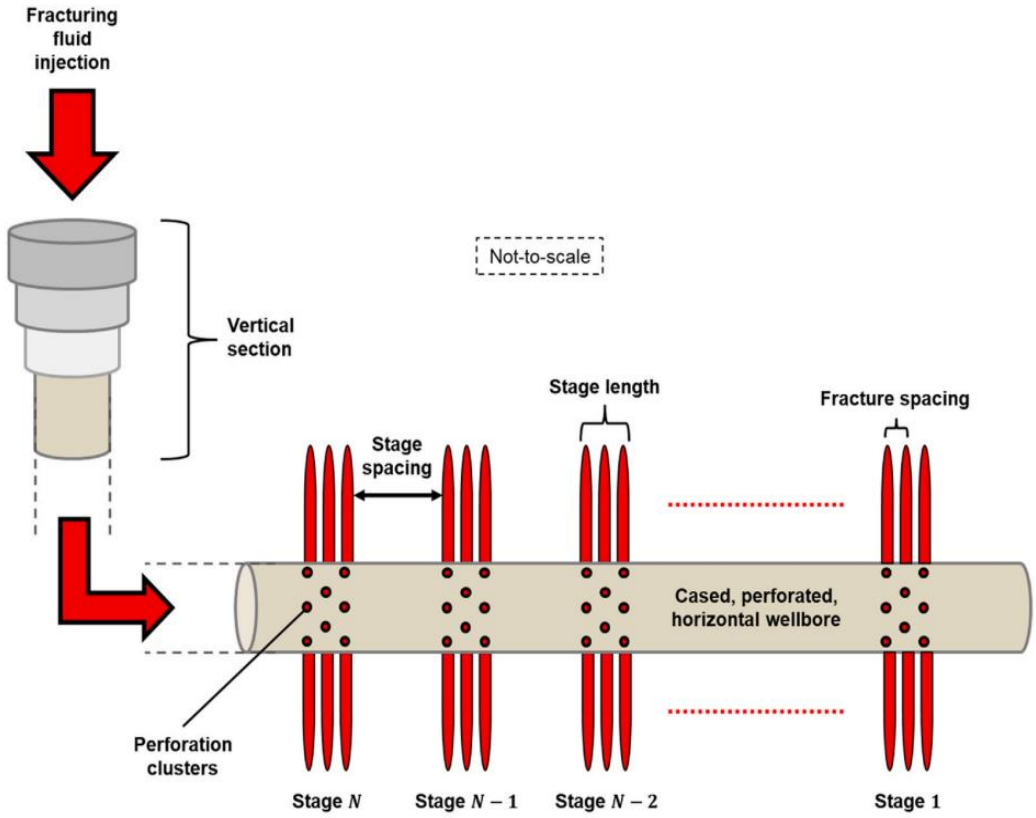


Figure 1. 3: Schematic of horizontally drilled well with a multi-stage hydraulic fracturing technique (Michael, 2021).

1.1 Thesis Structure

Chapter 1: The current chapter provides an introduction to the thesis, encompassing the research context, including the importance of shale gas and performance forecasting using machine learning.

Chapter 2: This chapter discusses the literature review, data-analytics, and machine learning applications that is used for forecasting the performance of shale gas production from unconventional shale reservoirs.

Chapter 3: This chapter will address the research problem clearly and precisely, along with outlining the research primary goals and the objectives of the study.

Chapter 4: This chapter will discuss the methodology used in this study, evaluating the way the research was carried out utilizing two primary methods namely, exploratory data analysis (EDA) and predictive modeling based on machine learning algorithms.

Chapter 5: This chapter will provide a concise presentation of the main study findings and a logical discussion with comments on the results.

Chapter 6: This chapter will highlight the notable conclusions drawn from the application of data-analytics based on ML for forecasting shale gas production as discussed in the preceding chapters. Additionally, this chapter will include recommendations for future studies.

2. LITERATURE REVIEW

Shale gas resources have attracted significant global interest because of its high capacity to provide the world with huge quantity of clean energy (Nguyen-Le et al., 2020). Differing from the conventional resources, unconventional shale reservoirs are organic-rich gas resources and share a common characteristic of having extremely low matrix permeability (Heller et al., 2014). The matrix permeability of a gas shale is known as ultralow-permeability, typically ranging from 1 to 100 nD, and possessing less than 10% porosity (Kim & Lee, 2015). Thus, long transient flow periods is occurred which make production forecasting and reserves estimation more challenging. Furthermore, hydraulic-fracture properties and horizontally drilled wells are also associated with the challenges of forecasting production (Gong et al., 2013). Therefore, numerical reservoir simulators in these systems is a highly complicated task and finding more practical alternatives is necessary (Schuetter et al., 2018). Recently, the emergence of the 4th Industrial Revolution (IR 4.0) is powered by a range of technological advancements including data analytics, artificial intelligence (AI), and internet-of-things (IoT). Data analytics has been addressed as one of the major component of IR 4.0, given its potential to provide valuable insights and patterns (Kaliraj & Devi, 2021). For a familiar literature, this section briefly expresses exploratory data analysis (EDA), machine learning applications and variable importance and model evaluation.

2.1 Exploratory Data Analysis

EDA is a data analysis philosophy that utilizes a range of techniques, mostly in a graphical representation, to achieve many objectives. The major goal of EDA is to gain insights into the dataset by uncovering underlying patterns, detecting the outliers and anomalies hence one is able to summarize the main characteristics of the dataset. Furthermore, it is used to examine the distribution of variables as well as study the relationship among various variables to generate hypotheses for further investigation (Holdaway, 2014). EDA is employed by numerous studies for reservoir Characterization, production optimization in unconventional reservoirs as well as for Performance assessment and forecasting. In their research, Holdaway (2009) utilized EDA in a reservoir characterization project by conducting histograms, box plots and scatter plots in order to investigate the structure of each individual reservoir variable, then the relationship between each pair of reservoir variables and finally between groups of variables to attain a deeper knowledge of the structure of the reservoir data.

Likewise, Schuetter et al. (2018) emphasizes the importance of conducting a preliminary exploratory data analysis prior building a model to identify any possible issues before moving forward. In this study, they utilized a scatterplots matrix as illustrated in Figure 2.1 to examine the relationship between each of the input variables (predictor) and output variable (response), along with histogram for each individual variable along the diagonal. The plot also indicates a robust relationship between pairs of predictors.

Artun (2020) employed histograms to identify the design parameters that were more frequently observed, which could potentially lead to a better performance of cyclic gas injection into a hydraulically fracked well. Another study by Lolon et al., (2016) utilized histograms and bivariate analysis to visualize the data graphically, and therefore identify potential outliers, deviation and the range of the data between mean, maximum or minimum value of each predictor. These analysis then was utilized to initiate a model to forecast the oil production.

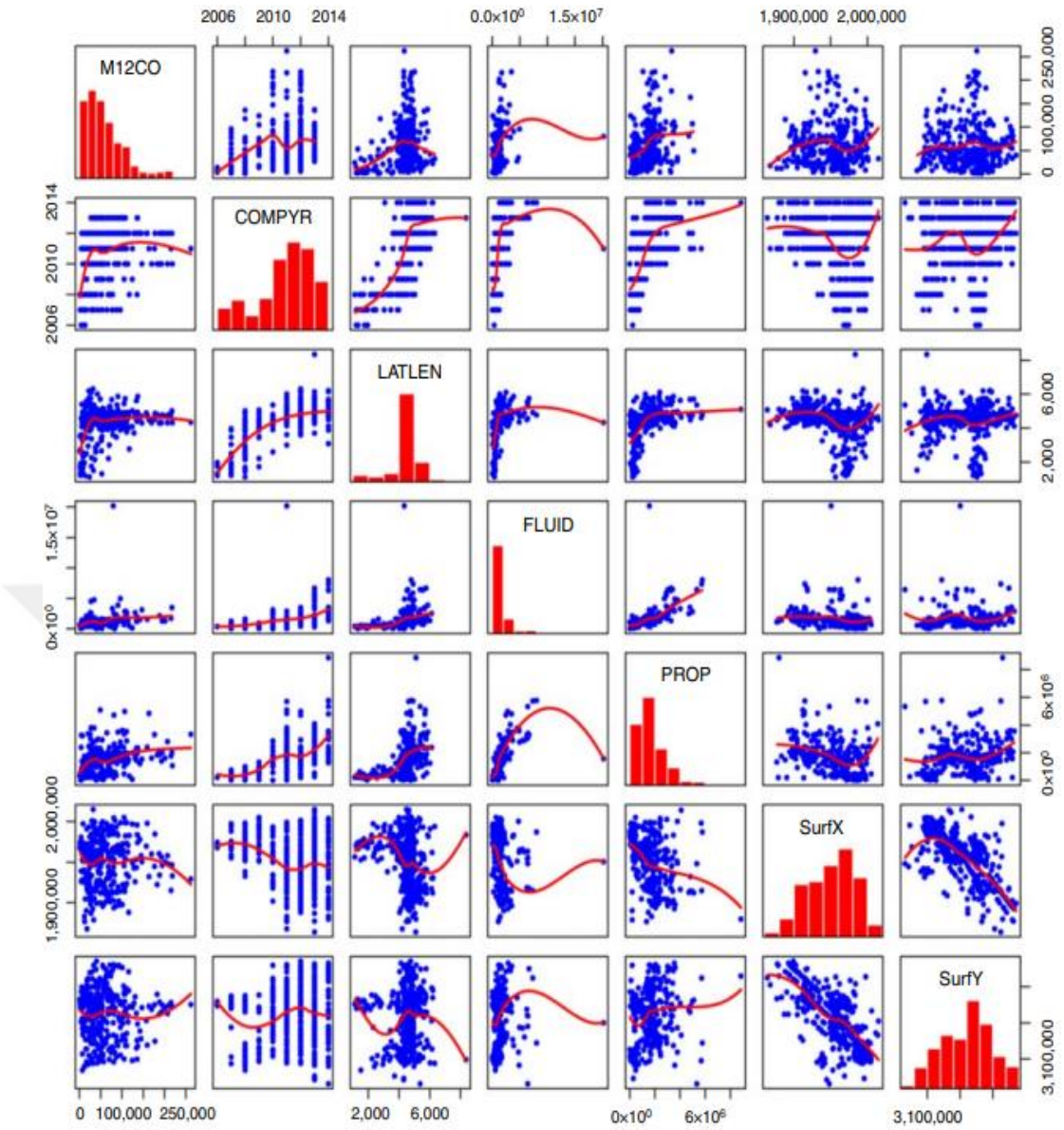


Figure 2. 1 : Matrix of scatterplot (Schuetter et al., 2018)

A Wolfcamp "Shale" Case Study by Zhong et al. (2015) conducted exploratory data analysis through histograms, the most commonly utilized univariate analysis, to analyze each variable characteristic and distribution (Figure 2.2). Then, scatterplot matrix is utilized to visualize the relationship between each pair of variables as shown in Figure 2.3 to determine any general trends or patterns as well as detecting unusual data points, including outliers and leverage points.

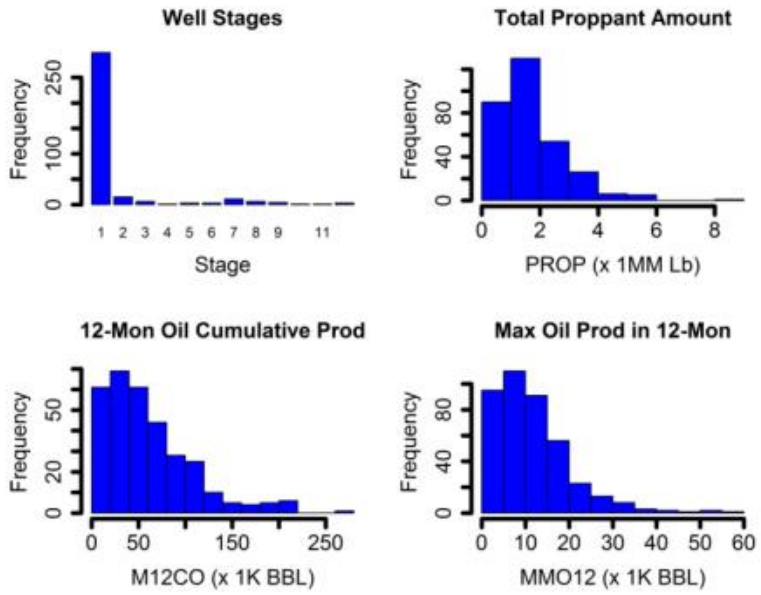


Figure 2. 2 : Histograms of predictors (Zhong et al., 2015).

Another study by LaFollette et al. (2013) performed EDA using histograms and matrix of scatterplots to graphically present the frequency distributions of each variable and the relationship between each pairwise of predictors respectively. All researches mentioned above, imply the importance of performing EDA before jumping into modeling stage.

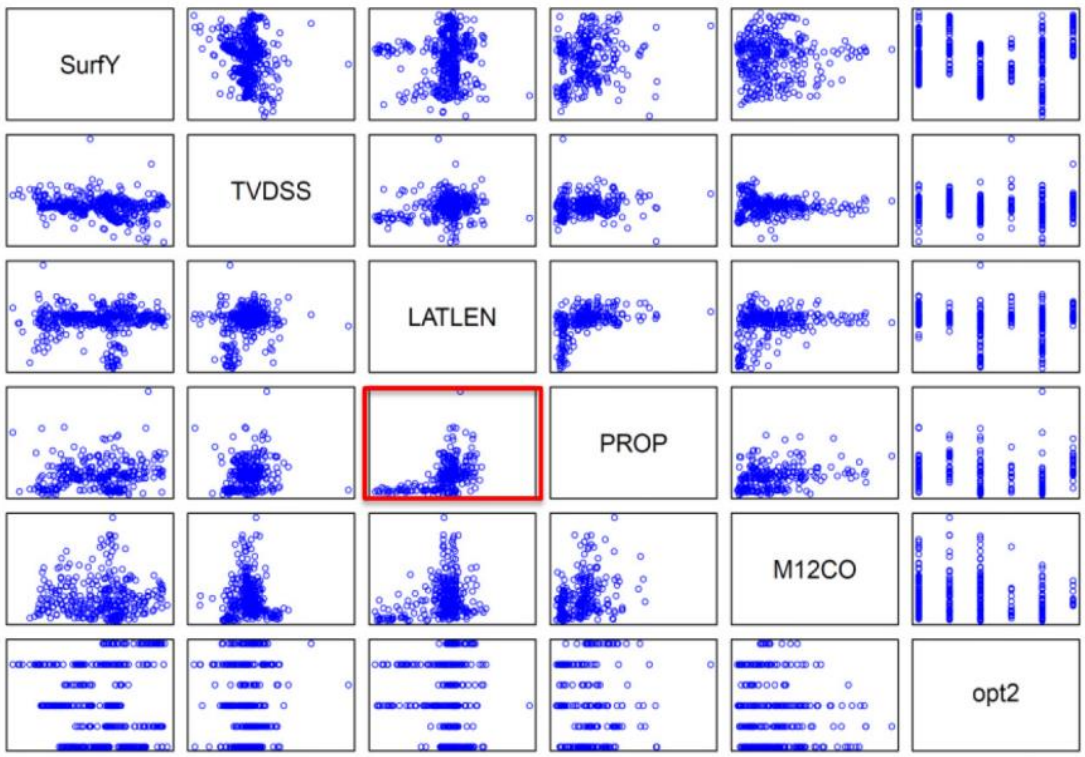


Figure 2. 3 : Scatterplot matrix of predictors (Zhong et al., 2015).

2.2 Machine Learning Applications

Over the last decade, many studies in oil and gas industry and many other engineering sectors show a considerable attention to the applications of machine learning (ML) as it proven to be the answer to many real-world challenges, which develop accurate models that allow engineers to make informed decisions ranging from exploration to field development (Arora & Doshi, 2021; Hanga & Kovalchuk, 2019; Kalantari-Dahaghi, 2022; Lee et al., 2018; Leem et al., 2022; S D Mohaghegh et al., 2017; Mohammed et al., 2016; Saldaña et al., 2023).

Predictive modeling using machine learning algorithms is defined by Kuhn & Johnson (2013) as the technique of developing or creating a mathematical tool, which uses an equation-based model, that produce an accurate prediction. Although predictive models have shown widely their ability to produce a desired or intended results fast, there are some situations which could limit their applications such as, the model have to be trained before problem-solving, unjustified extrapolation and the need of certain volume of data since the model could be questionable if the dataset is small (Ertekin & Sun, 2019). The following literature will discuss predictive models utilized in the oil and gas industry, specifically in the unconventional shale reservoirs.

A research by Lolon et al. (2016), utilized machine learning-based predictive models, such as gradient boosting machine (GBM), random forests (RF), and multiple linear regression to forecast cumulative oil production from 3,600 wells in the Three Forks formation and 6,800 wells drilled horizontally in the Middle Bakken shale play based on available production and completion data.

Kong et al. (2021) performed eXtreme Gradient Boosting machine (XGBoost) as a base model to estimate the cumulative production from 519 wells in the liquid-rich Duvernay shale paly using a database consisting of 46 variables, including drilling, operational, production and reservoir properties. Al-Alwani et al. (2019) also utilized a machine learning technique, partial least square regression, to assess the effect of different well completion parameters and stimulation parameters. These parameters were used as predictor variables to predict the gas production from 2700 wells in Marcellus shale.

Another study by Shahkarami et al. (2018) utilized 820 gas wells from 2000 wells available after running intensive pre-processing and hence build predictive model to

forecast shale gas wells estimated ultimate recovery (EUR) from Marcellus shale play, along with 25 input parameters including, location, reservoir , drilling, completion properties and initial well performance using several machine learning algorithms such as artificial neural networks (ANN), simple linear regression (SLR) and support vector machine (SVM). Bhattacharya et al. (2019) also effectively performed RF, ANN and SVM models to predict daily gas production in their case study from Marcellus Shale. Stating that these machine learning algorithms exhibit a remarkable speed over conventional physic-based reservoir simulations.

Decision trees based Machine learning algorithms were also successfully implemented in modeling of cumulative production in the Wolfcamp Shale play. In each model, the input parameters (predictors) from several aspects including, well location, production and hydraulic fracture properties had varying degrees of influence on predicting the target variable (Schuetter et al., 2018; Zhong et al., 2015).

Many other studies used predictive models in oil and gas industry to evaluate different aspects such as modeling of Eagle Ford shale based on well completion and frac strategies (Nejad et al., 2015), historical production data based forecasting the performance of shale gas reservoir (Gupta et al., 2014) and predicting EUR from 161 production gas wells in Sichuan Basin, China based on early data including geological and engineering properties (Niu et al., 2022).

Regarding the analysis of production behavior in unconventional shale reservoirs, several predictive modelling techniques utilized for classification and regression problems. The following literature will briefly clarify some of them, mainly the models used in this study, while the methodology chapter will provide a more comprehensive analysis.

Multiple Linear Regression: also called Ordinary-Least-Squares (OLS) Regression, a simple yet extremely effective technique that describe the response as a linear functions of the predictors (Schuetter et al., 2018). It aims to minimizes the Residuals Sum of Squares (RSS) between the actual values and forecasted response values (Lesmeister, 2019).

Classification and Regression Trees (CART): are methods based on machine-learning utilized to establish predictive models by partitioning the data recursively. The partitions is illustrated graphically as a tree hence is easily interpretable. The

classification trees is used when the predicted is categorical while the regression trees is utilized when the predicted outcome is numerical. Both calculate the prediction error by subtraction the squared of actual and predicted values (Loh, 2011).

Bagging: bagging is decision tree based predictive model, also known as bootstrap aggregation, used in general-purpose to reduce the variance within a noisy dataset by using multiple prediction trees on different subset of a training data. These prediction trees mostly use the strong predictors on the top splits. The predictions generated from these splits are then averaged to obtain the final prediction (James et al., 2021).

Random Forest (RF): Random forest is an improvement over bagging method. Instead of building the decision trees based on the strong predictors, Random forest choose subset of the predictors randomly hence not only the strong predictors but all other predictors will contribute to the final prediction (James et al., 2021). Concisely, RF was defined by Breiman et al. (1984) as an ensemble regression tree method with each tree being trained using random subset of predictors (as cited in Mishra & Lin, 2017).

Extreme gradient boosting (XGBoost): is machine learning system that can be scalable and optimized. It overcome the other existing methods in term of fast respond and accuracy due to the availability of algorithms optimizations (Chen & Guestrin, 2016). XGBoost aims to build several classification and regression trees and keep adding them sequentially (Tang et al., 2021). The final prediction can be estimated by summing up the predictions of all individual trees (Chen & Guestrin, 2016).

XGBoost have been employed in different aspects in oil and gas industry, such as permeability prediction (Zhao et al., 2022), fracturing parameters optimization in shale reservoir (Zhou & Ran, 2023), detecting the sweet spot in shale reservoir (Tang et al., 2021), production prediction and operation parameters optimization in shale gas well (Tan et al., 2021), forecasting reservoir porosity (Pan et al., 2022), forecasting crude oil price (Gumus & Kiran, 2017), well-placement optimization (Mousavi et al., 2020) and prediction of water absorption in sublayer (Liu et al., 2019). In addition to that, XGBoost is utilized over many other fields including medicine. For instance, Ma et al. (2020) performed XGBoot for diagnostic classification of cancers as well as Ogunleye & Wang (2020) conducted it for chronic kidney disease diagnosis. Nevertheless, XGBoost have shown a better performance than other machine learning

models including linear regression, (Pan et al., 2022; Zhao et al., 2022), multiple linear regression(Tan et al., 2021), and random forest , support vector machine (Ma et al., 2020; Nguyen-Le et al., 2020; Pan et al., 2022; Tan et al., 2021; Zhao et al., 2022).

The machine learning-based predictive models discussed earlier exhibit varying levels of predictive ability depending on the datasets used. For instance, Zhong et al. (2015) conducted research on the Wolfcamp dataset and found that tree-based methods, such as RF and GBM need less time for pre-processing the raw data. Additionally, these models were less affected by data quality issues and produced superior predictions compared to other models. In contrast, a research conducted by Lolon et al. (2016) revealed that despite the GBM model had the lowest root mean squared error (RMSE) and highest determination coefficient (R^2) when trained on a specific dataset, the model exhibited poor prediction ability when applied to the same dataset. Indicating that GBM is overfitting the training data and not suitable as prediction model for this dataset since the test RMSE significantly increases compared to the training RMSE.

2.3 Model Evaluation

Evaluating the performance of any statistical and machine learning-based model on a specific dataset is a must. The quality of fit is a way to measure how accurately the predicted values actually align with the actual data (James et al., 2021). For instance, Schuetter et al. (2018) proposed a commonly used method for model evaluation by constructing a scatterplot of predicted versus actual data. Hence if all data points fall along the 45˚ line, the model is fitted well to the training data. However, that does not imply the model generalizability and ability to work on new and unseen data. Figure 2.4 illustrates a case of a poor predictive model which is overfitting. However, it seem to represent a good fit when assessed solely on the training dataset. Thus, it is important to held out a subset of the training data to test and evaluate the fit on the test observations (Schuetter et al., 2018).

For model performance evaluation, quality of fit metrics such as R-squared (R^2), mean squared error (MSE) and mean absolute error (MAE) are performed (Chahar et al., 2022).

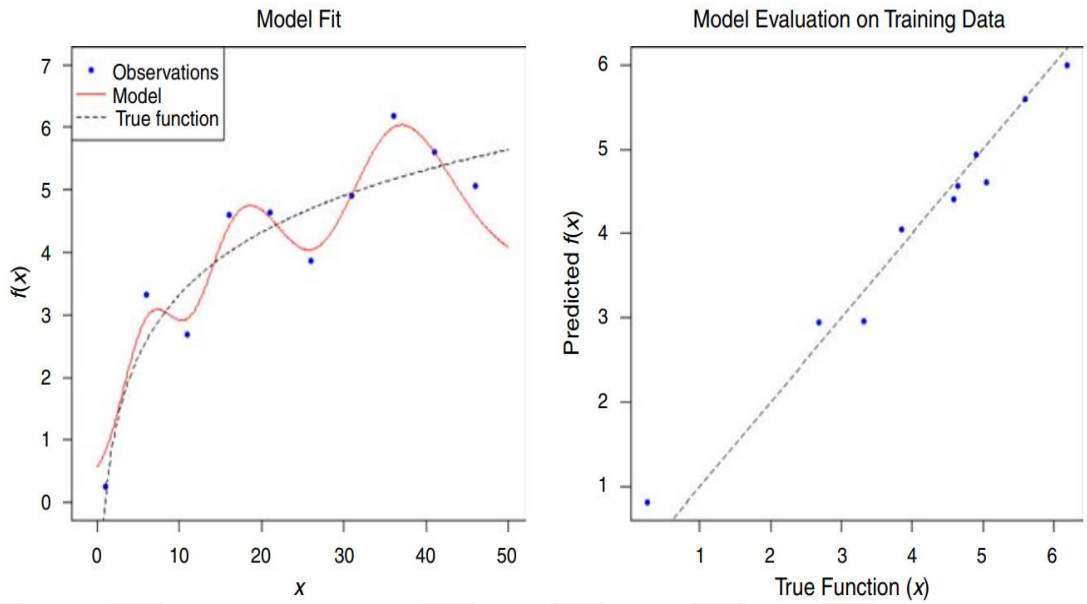


Figure 2. 4: Model evaluation using Scatterplot (Schuetter et al., 2018)

Furthermore, several researchers have utilized various methods to evaluate the model quality of fit. For example, Kong et al. (2021) in their study used the MSE and R^2 as evaluation metrics. Lolon et al. (2016) as well as utilized determination coefficient (R^2) with training root mean squared error (RMSE) and test RMSE to compare the performance of different models.

Further studies proposed average absolute error (AAE), MSE and pseudo-R^2 as a model goodness-of-fit metrics (Mishra & Lin, 2017; Schuetter et al., 2018). While Zhong et al. (2015) adopted AAE and MSE in their study.

2.4 Variable Importance

In certain scenarios, the objective may not be to establish a predictive model for a specific response. Instead, the objective may be to identify the main drivers that influence that response among a wide range of predictors(Schuetter et al., 2018). For instance, to determine the predictive power of each variable in a RF model, the relative importance should be computed by evaluating the increase in MSE when that specific variable is permutated while all other variables remain unchanged (Breiman, 2001) (as cited in Mishra & Lin, 2017).

Several studies used RF, GBM and many other methods to implement the variable importance algorithms to present the most influential variables in their research. As

seen in Figure 2.5 Artun (2020) performed the generalized additive model (GAM) and RF variable importance and both resulted with the cyclic injection volume as the most influential variable among the others.

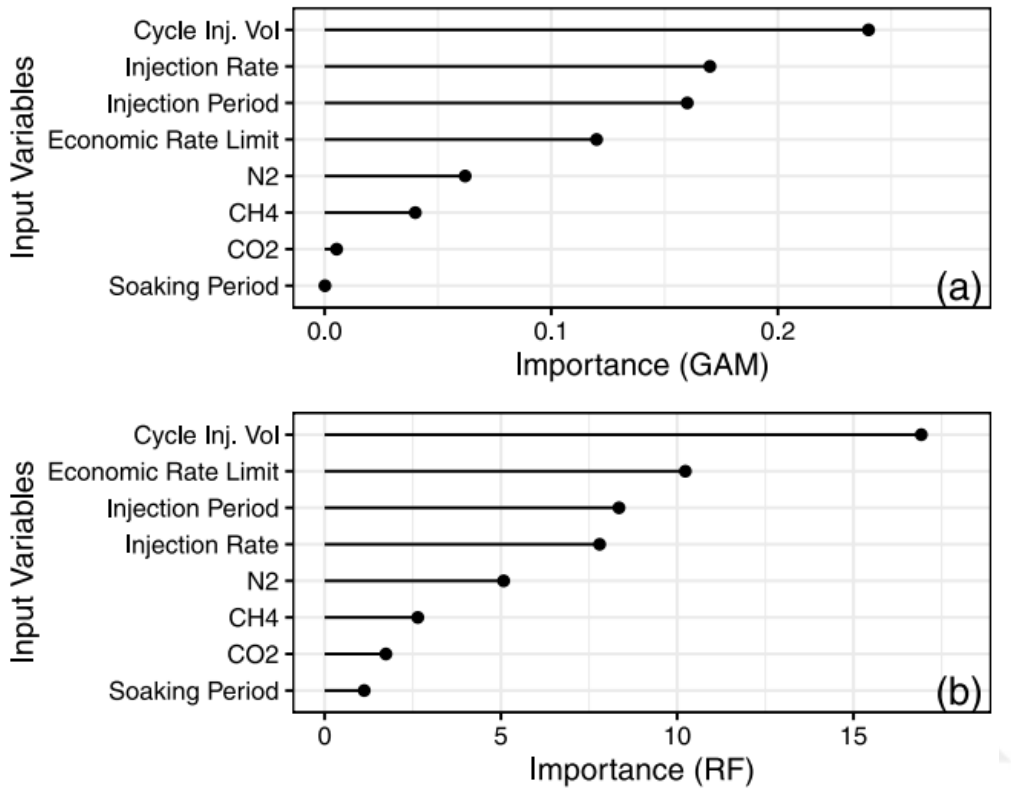


Figure 2. 5: Feature importance by (a) GAM, (b) RF (Artun, 2020).

Multiple studies also utilized different methods including RF, GBM and XGboost. For example, Schuetter et al. (2018) and Zhong et al. (2015) employed RF and GBM to compare and determine the main drivers and worst predictors. Another study by Bhattacharya et al. (2019) used RF relative influence method. Finally, to observe the difference in ranking of influential predictors, different methods like RF, GBM and XGBoost could be implemented at once as a measures of variable importance analysis (Han & Kwon, 2021). In random forest (RF), the relative importance is calculated by the increase of mean square error (%IncMSE) of the prediction for every tree (Han & Kwon, 2021; Schuetter et al., 2018). While the GBM and XGBoost calculating the relative importance upon to the partial gain of each predictor to the total gain resulting from variable splits as well as measuring the relative frequency for each predictor utilized within the tree splits (Han & Kwon, 2021).

2.5 Summary

The literature review discussed in this chapter intended to explore the challenges associated with using numerical reservoir simulation for providing insights and evaluated the applications of data analytics and machine learning-based predictive models in unconventional reservoirs. The findings consistently indicate that numerical reservoir simulators in such complicated systems like unconventional shale reservoirs where the permeability is measured in nanoDarcy is a complicated task and economically infeasible. Thus, data-driven methods are gaining growing significance in production optimization in oil and gas industry. Furthermore, the decrease of greenhouse gases emissions in the past decade came in conjunction with the abundance production of gas from shale resources. It is crucial to focus on understanding and optimizing the production of shale gas for fuel switching process from coal to natural gas. Thus, this would effectively limit CO_2 emissions. To do this economically, data-driven models based on machine learning for forecasting and optimizing the gas production as well as understanding the key driver parameters of shale gas production should be given utmost priority in our journey.

3. STATEMENT OF THE PROBLEM

Although shale gas production is becoming a significant source of energy supply, forecasting the performance of shale gas wells remains a challenging task due to the complex and heterogeneous nature of shale formations.

Existing reservoir simulation models in such unconventional shale reservoirs, which are characterized by more complex fluid flow dynamics compared to conventional reservoirs, behave to assume uniform relationship between input and response variables. This approach overlooks the dynamic and nonlinear nature of shale gas reservoirs. Consequently, these models are less reliable, computationally intensive and costly, and unable to capture a comprehensive full-field analysis such as reservoir properties and operational parameters. As a result, the accuracy of these predictions is limited.

To address this problem, a machine learning techniques are being employed to develop robust predictive models that can effectively with fast computational speed correlate various field data, extract valuable information, identify insightful patterns and relationships between variables. These machine learning models have the potential to accurately make predictions of the performance of shale gas wells. Moreover, they have the capability to identify the primary parameters that have the most significant impact on the performance of shale gas wells.

In this study, the successfully developed machine learning model will contribute in the development of sustainable designs for similar forecasting and modelling projects.

3.1 Objectives of the study

The main aim of this study is to build a machine learning-based model that can effectively and accurately forecast the performance of shale gas wells as well as identify whether reservoir or operational properties have the most impact on the performance of these wells. Furthermore, depicting the way of implementing machine-learning predictive models to effectively forecast the cumulative gas production after one year. Additionally, the study outlines the process of selecting the best performing

model that adequately represents the dataset used in this study. This selection will be based on the use of statistical evaluation metrics.

It is typical to present a list of research objectives in the following manner:

- Performing initial investigations using exploratory data analysis to uncover any underlying patterns on data.
- Calculate the correlation between cumulative gas production (response) and each input variable.
- Forecasting the cumulative gas production using machine-learning predictive models.
- Identify the key driver parameters of gas production among the whole predictors using variable importance algorithms.

A set of questions to be answered later are listed in the following manner:

- What are the main characteristics of dataset (number of observations and variables)?
- What kind of patterns and trend exist in the data, Is their outliers?
- Is there any correlation between predictors and response as well as among predictor variables?
- Is the predictive model capable to accurately make predictions on unseen data?
- What parameters influence the performance of shale gas wells the most?

3.2 General Workflow

The general workflow adopted in this study to develop a machine learning-based predictive model for forecasting the performance of shale gas wells is illustrated in Figure 3.1.

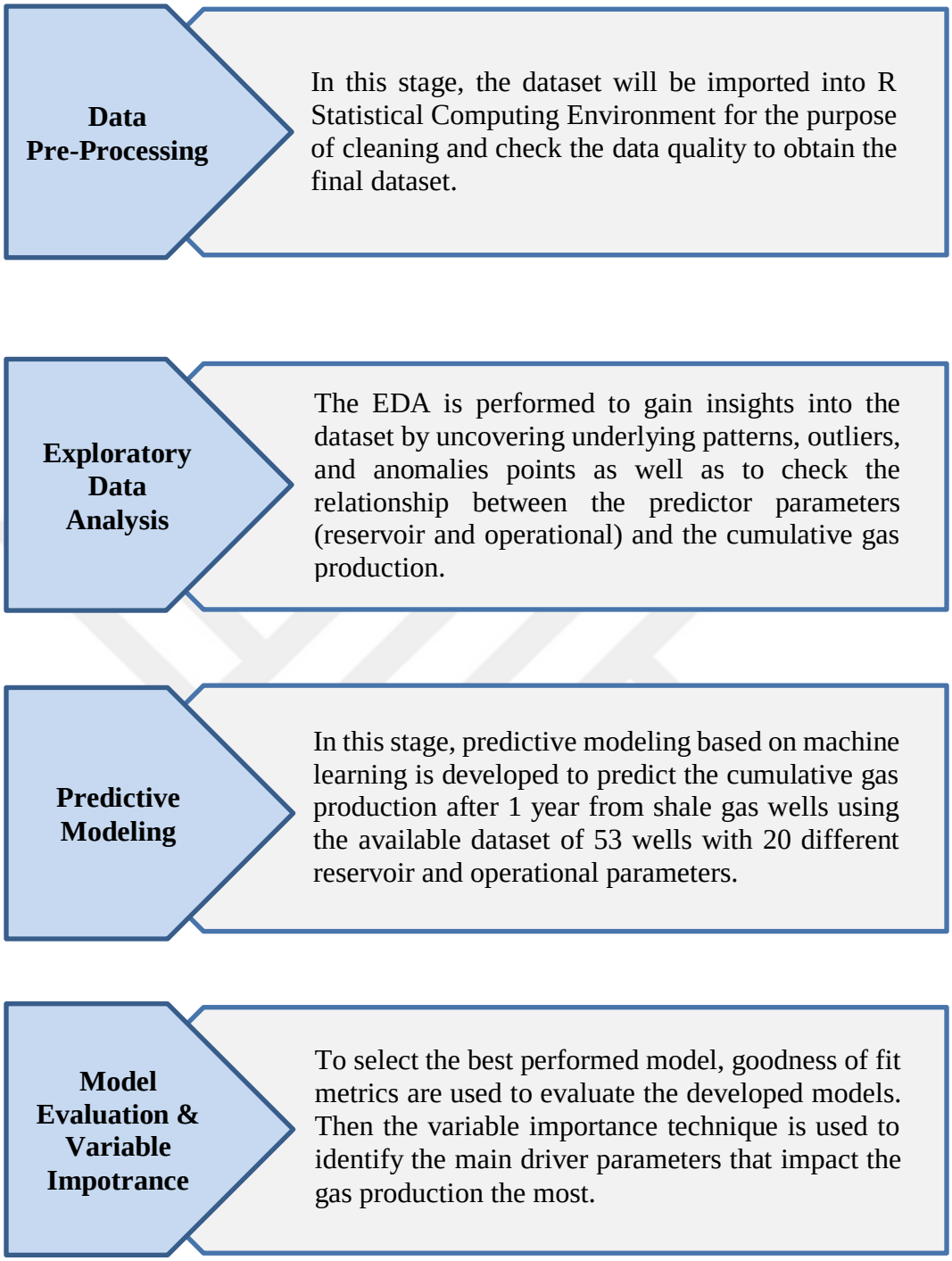


Figure 3. 1 : General workflow of a machine learning-based forecasting model.

4. METHODOLOGY

In Chapter 3, Figure 3.1 depicts a flowchart that represents the essential components and procedures undertaken in this research. The initial step following data pre-processing is to conduct exploratory data analysis, which aims to ensure the quality and reliability of the data being used. A forecasting models using machine learning methods including multiple linear regression and tree-based approaches such as regression tree, bagging, RF and XGBoost are used to forecast the cumulative gas production after one year from shale wells. After training and testing the model, evaluation of the models is done using goodness of fit metrics. Finally the variable importance is conducted using random forest and XGBoost algorithms. Consequently, the key parameters, whether reservoir or operational, that influence the performance of shale wells is determined.

The software employed to develop the models and execute the analysis procedures and generate meaningful insights from the collected data in this research is R Statistical Computing Environment (R Core Team, 2021).

This methodology chapter provides a detailed explaining of all methods employed throughout the research process, beginning with the exploratory data analysis stage to the predictive modeling. Each method used to successfully complete the study will be explained comprehensively. According to (Mishra & Datta-Gupta, 2018), for petroleum geoscience applications, it is advisable to pay attention to data analysis and statistical modeling cycle. This cycle is shown in Figure 4.1.

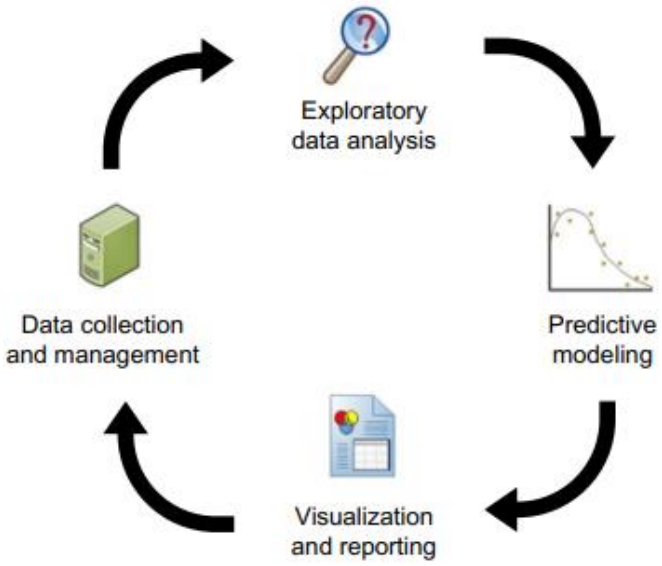


Figure 4. 1 : Data analysis cycle (Mishra & Datta-Gupta, 2018)

4.1 Dataset Collection

The dataset utilized in this study is acquired from SPE Data Repository (SPE, 2021). The dataset is consisting of 53 shale wells from Eagle Ford, Marcellus and Haynesville shale formations. After completing the data pre-processing stage, a resulting dataset comprising of 53 wells with 20 distinct parameters is obtained. The parameters are distributed among reservoir and operational properties, with Table 4.1 providing an overview of the parameters contained within the dataset utilized in this research.

Table 4. 1 : Dataset parameters

Parameter Description	Variable	Unit	Type
Cumulative Gas Production	Gp	MMscf	Response
Initial Pressure	Pi	psi	
Reservoir Temperature	Tr	deg F	
Net Pay Thickness	h	ft	
Porosity	Porosity	%	
Water Saturation	Sw	%	
Gas Saturation	Sg	%	
Gas Specific Gravity	Yg	%	
CO_2	CO2	%	
N_2	N2	%	
True Vertical Depth	TVD	ft	Predictor
Stage Spacing	Spacing	ft	
Stages	Stages	#	
Clusters	clusters	#	
Clusters per Stage	Clusters_Stage	#	
Proppant	Proppant	lbs	
Lateral Length	Lateral_Length	ft	
Top Perforation	Top_Perf	ft	
Bottom Perforation	Bottom_Perf	ft	
Sandface Temperature	Tsf	deg F	
Static Wellhead Temperature	Twhs	deg F	

4.2 Exploratory Data Analysis

Exploratory data analysis (EDA) is defined by Tukey as a numerical detective work trying to collect evidence from the data and assessing the strength of these evidence. It can serve as a foundation stone in every field involving data, providing a solid foundation for analysis and interpretation. For instances, a crime detective to successfully accomplish his works needs some clues and tools in order to understand where the criminal most likely put his fingers as well as a tool such as fingerprint powder to find fingerprints in surfaces. Equally, a similar scenario is useful in the data analysis (Tukey, 1977). In data analysis basis the evidence is the predictor variables,

and assessing the correlation between predictors can determine the strength of them (evidence).

EDA is employing a several tools, most of them visualized graphically, to achieve many objectives. The primary objective is to gain the greatest possible insight from a dataset (Holdaway, 2014). Furthermore, a list of objectives are listed below:

- Reveal any underlying structure.
- Pinpoint any anomalies or outliers in the data.
- Identify the key variables.
- Identify the relationship between variables.

EDA approach is built to disclose "what does data telling us" by generating hypothesis that guide us for improving the process performance. EDA, Classical, and Bayesian are three well-known data analysis approaches that exhibit general similarities in their principles and objectives. However, EDA is differing from them in the sequence as shown in Figure 4.2. For EDA sequence, after collecting the data the model is not developed as in Classical and Bayesian analysis. Instead, analysis is conducted immediately after the data collection in order to find out what model would be appropriate to those data (Holdaway, 2014).

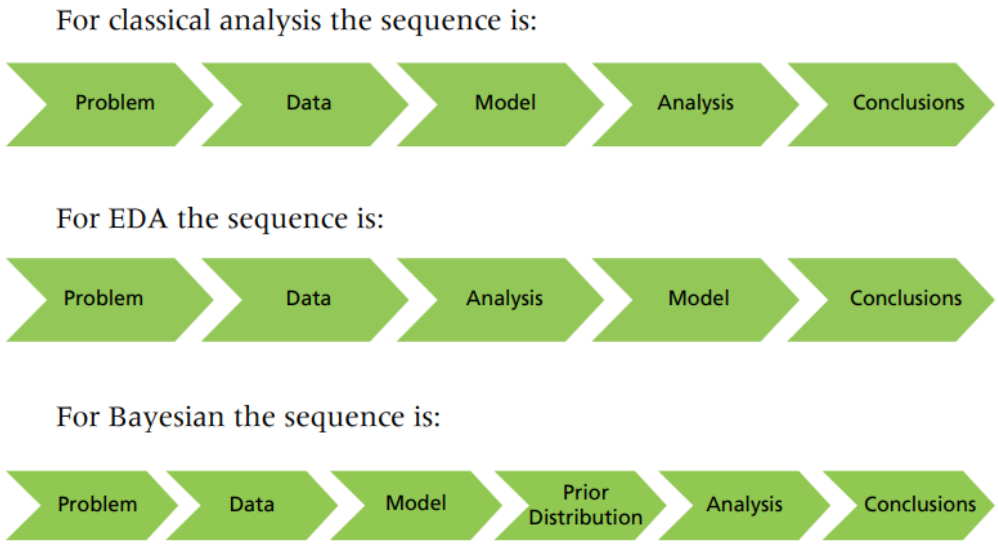


Figure 4. 2 : The sequence of different data analysis approaches (Holdaway, 2014)

In this research the exploratory data analysis sequence is followed in order to firstly obtain as much as possible insights of the dataset then being able to choose the best appropriate model for our data set. To achieve this, EDA was partitioned into three main components in this study, each of which plays a significant role in comprehensively exploring and understanding the data. These three main components are listed as the following:

- Univariate analysis
- Bivariate analysis
- Multivariate analysis

4.2.1 Univariate analysis

Univariate data analysis is a crucial initial step of EDA, involves outlining the characteristics of individual variables. This includes measuring the standard descriptions, such as like, the degree of clustering or uniformity in the data, as well as measuring the largest and smallest values, the spread, and distributional symmetry. Univariate analysis also illustrate the descriptors graphically such as standard deviation, mean, median, and mode (Holdaway, 2014). Below, the above mentioned univariate measures, along with commonly used graphical methods for visually analyzing and summarizing the data are explained.

4.2.1.1 Measures of center

Measuring the center of data is executed by identifying the mean which is the most frequently utilized measure of central tendency. For any variable X, to calculate the central tendency, equation 4.1 is used. x_i representing the single outcome (Mishra & Datta-Gupta, 2018).

$$E[X] = \bar{X} = \sum_{i=1}^{N} f_i x_i = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{4.1}$$

Where,

f_i : Relative frequency

x_i : Outcome

X : Random variable

N : Number of observations

$\bar{X}$: Arithmetic mean

Besides the mean, two more useful measures of central tendency are utilized, namely median and mode. The median represents the center of the distribution, while the mode corresponds to the most commonly occurring value. Whereas, the mean is the arithmetic average of all observations in the range of a variable (Mishra & Datta-Gupta, 2018).

The statistical location for mean, median, and mode may look similar when the distribution of data is symmetrical or nearly symmetrical. Nevertheless, for skewed distributions (asymmetrical), they can differ significantly. This because the mean is highly affected by the outlier values. In contrast the median is less influenced by the presence of outliers (Mishra & Datta-Gupta, 2018). As shown in Figure 4.3, the mode and mean is changing location up to the data distribution, whether it is left-skewed or right-skewed. However, the two cases show that median is located between the mode an mean (Mishra & Datta-Gupta, 2018).

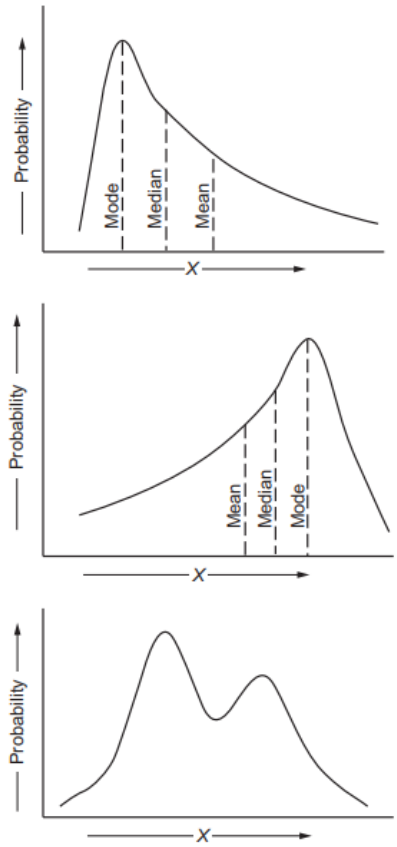


Figure 4. 3 : Location of mean, mode, and median for varying distributions (Mishra & Datta-Gupta, 2018)

4.2.1.2 Measures of dispersion

Thera are two main measures to evaluate the dispersion namely variance and standard deviation. Both trying to define whereas the data are tightly clustered or widely scattered around the mean. If the data exhibit a wide spread around the mean, the calculated variance, and standard deviation tend to be relatively low. Conversely, they would be relatively high if the data points are tightly clustered (Nisbet et al., 2018).

Mishra & Datta-Gupta (2018) stated that the fundamental measure of dispersion around the mean is the variance. It is defined by the following equation (4.2).

$$V[X] = \sigma_x^2 = \sum_{i=1}^{N} f_i (x_i - E[X])^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - E[X])^2$$
$$V[X] = \sigma_x^2 = \frac{\sum x_i^2}{N} - (E[X])^2 = E[X^2] - (E[X])^2 \quad (4.2)$$

Simply, the variance can be described as the difference between the mean of squared values and the square of the mean itself (Mishra & Datta-Gupta, 2018). The variance is equal in value with the mean squared error (MSE). While the standard deviation is obtained by taking the square root of variance (Bruce et al., 2020). Consequently, the standard deviation is also equal in value with the RMSE. The coefficient of variance, often denoted as CV, is a statistical metric used as a normalized measure of data spread. It is typically presented as a percentage and is defined as in equation 4.3 (Mishra & Datta-Gupta, 2018).

$$CV[X] = \frac{\sigma_x}{E[X]} \times 100\% \quad (4.3)$$

Comparing CV with variance or standard deviation when dealing with the heterogeneity of reservoir parameters like permeability, the coefficient of variance (CV) is more reliable and consistent since it give priority to the variable data spread, regardless of the size of observed values (Mishra & Datta-Gupta, 2018).

4.2.1.3 Univariate data graphing

In this study, to visualize univariate data graphically, two main tools were employed namely boxplots and histograms.

Boxplots: A boxplot, commonly known as a box and whisker diagram, is utilized to depict the dispersion and center of the data for a single variable as well as to detect outliers and anomalies. To illustrate the center and dispersion of data using boxplots five statistics are used: the minimum observation, maximum observation, median, first quartile and third quartile (Smith, 2011).

In Figure 4.4, the rectangle box indicating the interquartile range (IQR), the line inside the box is representing the median of the data. The upper and down whisker is depicting the data values that are higher and lower from the data in the IQR (Holdaway, 2014).

To summarize the above, in a numerical order, the boxplot is dividing the data into four groups, where the median is halving them by 50 percent. The lower whisker illustrates the minimum value to 25th percentile of the data, the data between lower edge of the box and the median illustrate 25th percentile to 50th percentile, data between median and top edge of the box is the 50th percentile to 75th percentile of the data, and the upper whisker represents 75th percentile to maximum (Smith, 2011).

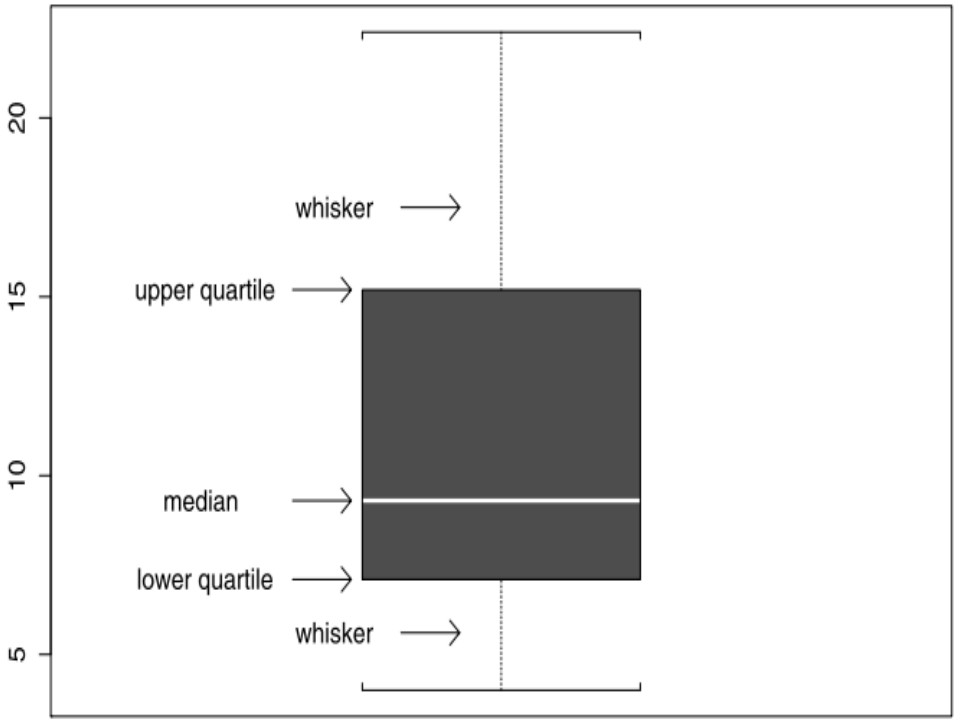


Figure 4. 4 : Boxplot example (Wilcox, 2021).

For the extreme values or so called outliers, which fall outside the box below or above the whiskers (below minimum or above maximum values). These outlier data points are 1.5 times farther than the ones inside the interquartile range (IQR). Outliers are illustrated in Figure 4.5 as separate points (Holdaway, 2014).

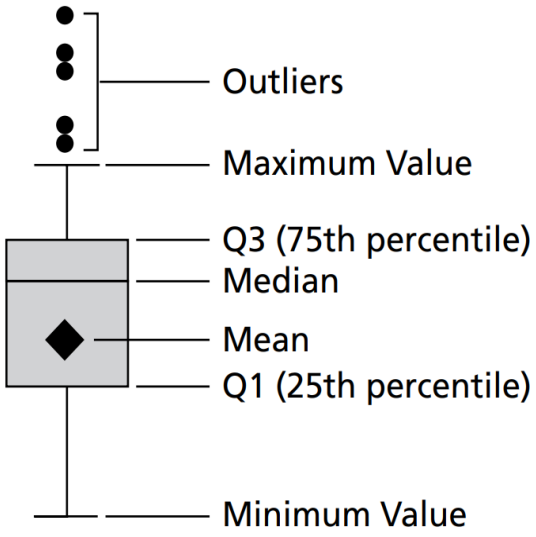


Figure 4. 5 : Box-Whiskers plot with outliers (Holdaway, 2014).

Histograms:

Histograms aim to represent the data graphically in order to summarize a dataset by depicting the distribution with the relative frequency or occurrences using vertical bins. In order to determine the frequency of each variable, one must count the number of observations that fall into each respective bin as shown in Figure 4.6 (Martinez et al., 2017). Histograms are widely used method of univariate data which is useful to extract meaningful information such as dispersion, the central tendency, and the overall shape of the distribution (Holdaway, 2009).

Histograms are capable to be applied to a large-scale datasets. Moreover, they are simple and computationally feasible (Martinez et al., 2017). They are implemented by first splitting the actual values into set of intervals or bins. Then, these bins is corresponded to their frequency of occurrence. For determining the bins number in histograms, common rules are typically employed. This process may involve using trial and error as listed below (Mishra & Datta-Gupta, 2018).

- The bins number or intervals (k) for an observation size of N can be obtained by selecting the smallest integer that satisfies the condition $\mathbf{2^k \geq N}$, as suggested by Iman and Conover in 1983 (as cited in Mishra & Datta-Gupta, 2018).
- A default value of the bins number can be set as {3.3log (N) + 1}. This was suggested by Venables and Ripley (1997) (as cited in Mishra & Datta-Gupta, 2018).

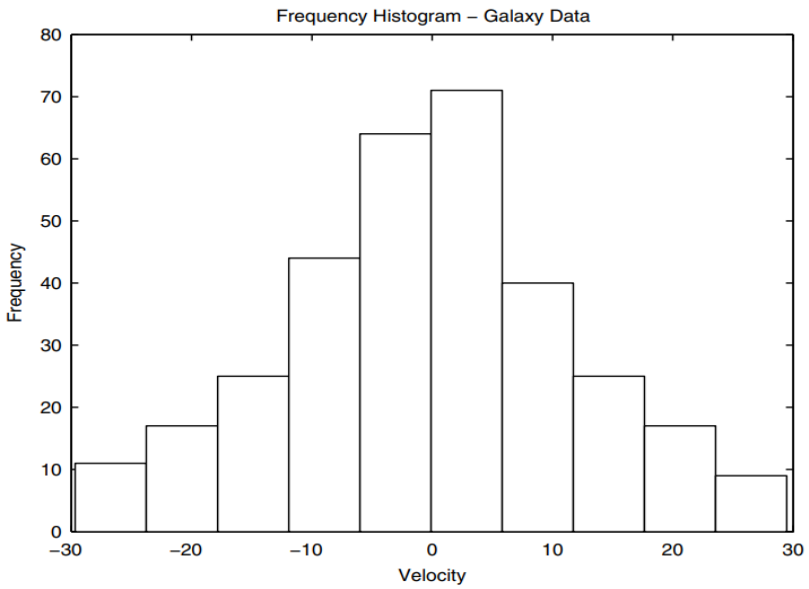


Figure 4. 6 : Histogram example (Martinez et al., 2017)

Since the shape of histograms are heavily reliable on the number of bins, they may not be considered as a highly dependable graphical tool. For instance, Figure 4.7 shows the histograms of the same variable, which depict different results, using different number of bins. Thus, it is recommended to use multiple bin numbers until a strong shape indication is obtained (Mishra & Datta-Gupta, 2018).

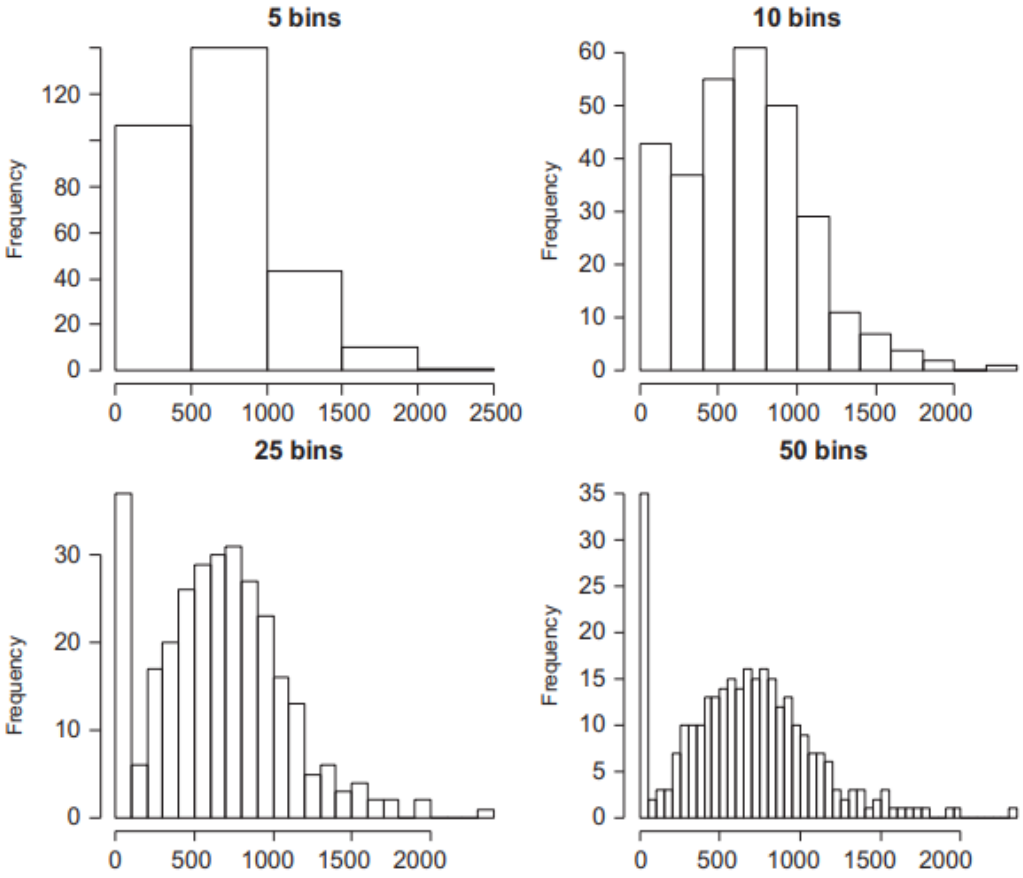


Figure 4. 7 : Sensitivity of histograms to bins number. (Mishra & Datta-Gupta, 2018)

4.2.2 Bivariate analysis

Bivariate analysis is a statistical method used to examine and define the relationship between multiple variables, often involving two or more variables. The main tool used in bivariate analysis is scatterplot (Holdaway, 2014). For instance, the behavior of a variable can be observed over the other variable using a line plot. These plots also can visualize the relationship of a response variable versus predictor variables (Holdaway, 2014). One of the measures used in this study to identify the strength and direction of correlation between variables is explained below.

4.2.2.1 Correlation and rank correlation

The correlation coefficient (CC) is measure metric to examine how strong the linear relationship among predictor variables, and between predictors and response (Bruce et al., 2020). The CC, also referred to as the Pearson correlation coefficient, which spans from -1 to +1. A value of +1 and -1 denotes a strong positive and a strong negative correlation respectively. Thus, the negative and positive sign are determining the directions of correlated variables, whereas the absolute value is referring to the strength of the correlation (Mishra & Datta-Gupta, 2018). The CC can be calculated using equation (4.4).

$$CC = \rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{1}{N-1} \sum_{i=1}^{N} \left(\frac{x_i - \bar{X}}{\sigma_x} \right) \left(\frac{y_i - \bar{Y}}{\sigma_y} \right) \tag{4.4}$$

Where,

σ_{xy} : Covariance

σ_x : Standard deviation of X

σ_y : Standard deviation of Y

$\bar{X}$: mean of X

$\bar{Y}$: mean of Y

$N - 1$: degree of freedom

x_i : Single output for x

y_i : Single output for y

The rank correlation coefficient (RCC) is utilized as robust metric to examine the correlation between variables if they are correlated in a non-linear manner. The RCC, also referred to as the Spearman correlation coefficient and is defined by equation (4.5) (Mishra & Datta-Gupta, 2018).

$$RCC = \rho_{xy(rank)} = \frac{\sigma_{xy(rank)}}{\sigma_{x(rank)} \sigma_{y(rank)}} = \frac{1}{N-1} \sum_{i=1}^{N} \left(\frac{R_{x,i} - \bar{R}_x}{\sigma_{Rx}} \right) \left(\frac{R_{y,i} - \bar{R}_y}{\sigma_{Ry}} \right) \tag{4.5}$$

Where,

$\sigma_{xy(rank)}$: Rank variables covariance

$\sigma_{x(rank)}$: Rank variable standard deviation of X

$\sigma_{y(rank)}$: Rank variable standard deviation of Y

Furthermore, RCC can be simply calculated using equation 4.6 based on the ranks difference.

$$RCC = 1 - \frac{6\sum d^2}{N(N^2 - 1)} \tag{4.6}$$

Where,

d : Difference of two ranks at each observation

N : Number of observations

4.2.2.2 Bivariate data graphing

In this study, scatterplots were the primary bivariate visualization technique used in to analyze the relationship among variables, as well as between variables and response (target variable).

Scatterplots: Scatterplots is useful way to represent the data graphically. The primary objective of a scatterplot is to display the relationship or correlation between two variables. This is accomplished by plotting dots on the graph, with each dot represents a single individual, unless two or more individuals share the same data point (Utts, 2014). The independent variable (predictor) always plotted on the x axis, while the response is on the y axis (Moore et al., 2017).

In addition to illustrating the relationship between variables, scatterplots provide valuable insights into various aspects of the data such as outliers, variance, (Moore et al., 2017) and the presence of clusters or collinearity as shown in Figure 4.8 (James et al., 2021).

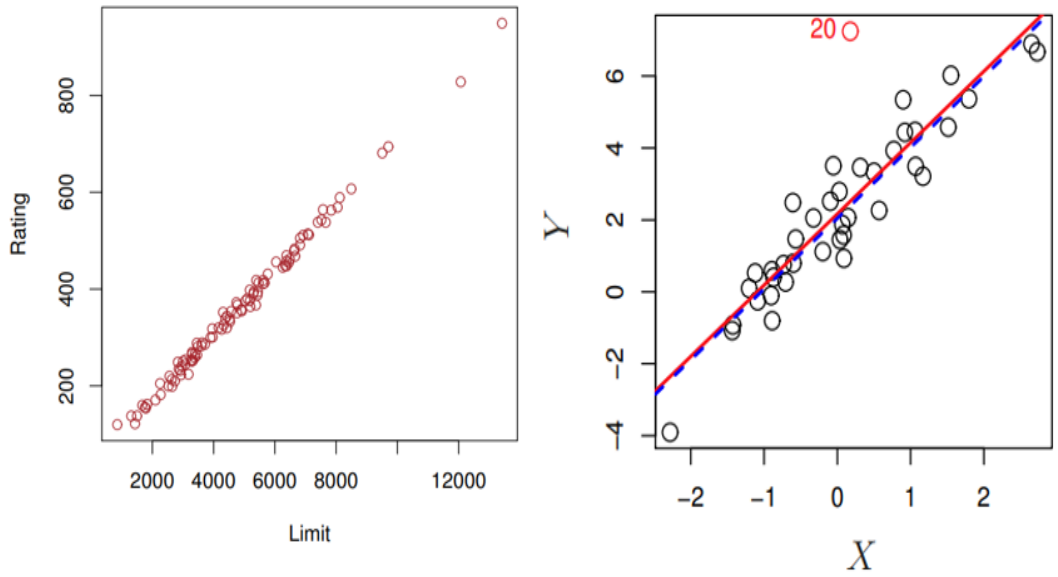


Figure 4. 8 : Scatterplots for collinearity (left), and outliers (right) (James et al., 2021)

Furthermore, the overall pattern of scatterplot can be interpreted by three main aspects namely, direction, form, and strength of the correlation. The direction can determine the relationship trend whether from lower left to the upper right or vice versa. Form identifies the form of the line whether straight, curved or oscillate. The strength indicates how closely the data points fall along the line or the form (Moore et al., 2017).

Multiple example of scatterplots are shown in Figure 4.8 by Mishra & Datta-Gupta (2018), illustrating the behavior between two variables with their corresponding Pearson correlation coefficient. As it observed in the top left scatterplot (A), a trend from lower left to the upper right, showing a strong relationship between the two variables and the direction of them is positive since the average values of one variable increases, the other variable also tend to increase. The top right scatterplot depict a strong linear relationship with a negative direction. The bottom left and right scatterplots illustrating a weak negative and modest positive correlation respectively between the two variables.

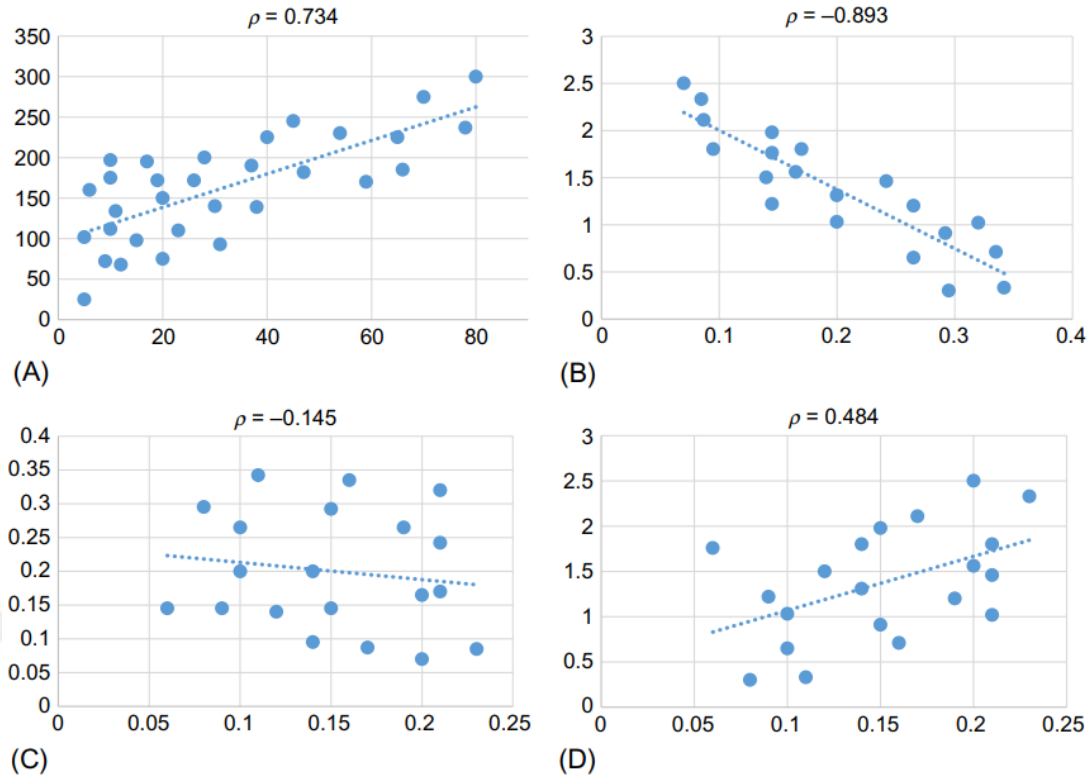


Figure 4. 9 : Example of scatterplots with corresponding Pearson CC (Mishra & Datta-Gupta, 2018)

4.2.3 Multivariate analysis

Multivariate analysis is an extension of bivariate analysis. Instead of calculating the Pearson or Spearman correlation coefficient for only two variables, multivariate analysis involves the calculation of correlation coefficient (CC) for all variables at once and represent them graphically using a correlation matrix as shown in Figure 4.10 (Mishra & Datta-Gupta, 2018). Correlation matrix is an easy to use exploratory tool that help in the detection of hidden patterns among variables. It supports reordering the variables automatically to give insights of the relationship between variables. The correlation Matrix also assists in identifying the strength and direction of the correlation between two generic variables (Wei & Simko, 2021).

Similarly, to make comparison of variables, a scatterplot matrix or pairs plot is used to display a combination of scatterplots for a selected variables to examine the relationship between all variables simultaneously. In addition, the values of Pearson CC can be as well as combined with scatter plots as shown in Figure 4.11. Thus many valuable insights can be reached easily and visually for all variables in the dataset (Beck & Latif, 2022).

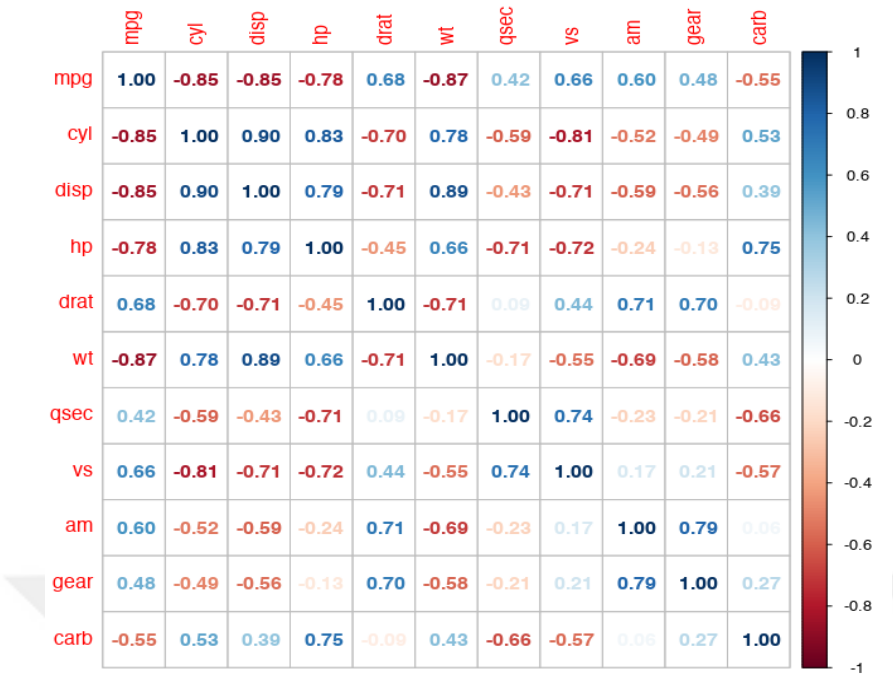


Figure 4. 10 : Correlation matrix (Wei & Simko, 2021).

In Figure 4.11, the left of the diagonal representing the scatterplots for each pair of variables, the diagonal depicting the distribution for each single variable (univariate) and the right of the diagonal showing the Pearson CC along with the asterisks (*) that outline the statistical significance (Beck & Latif, 2022).

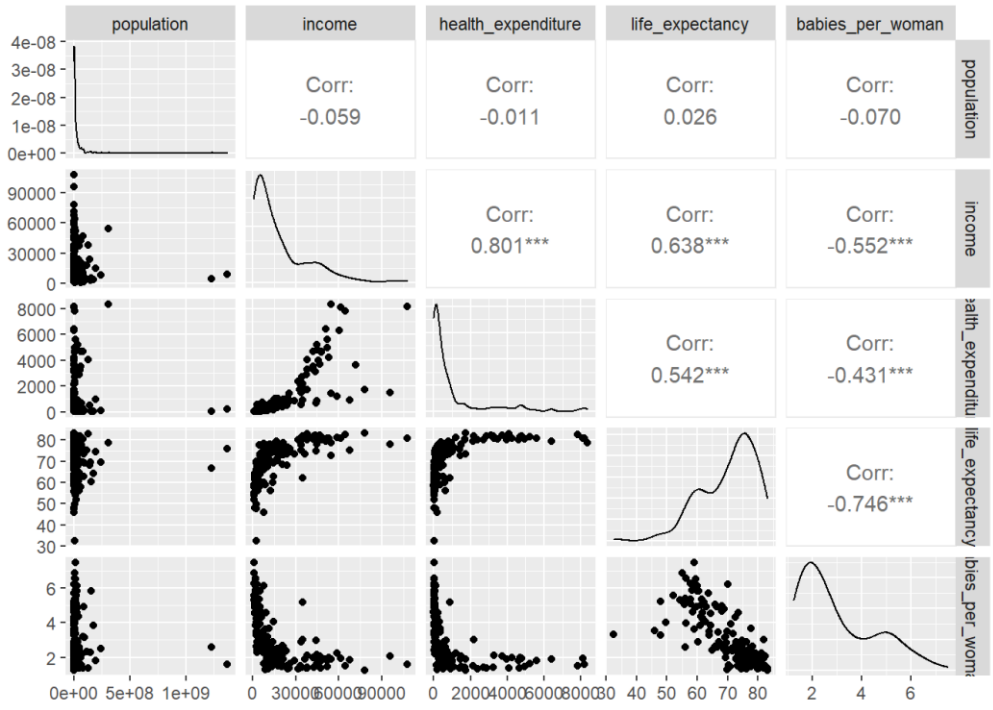


Figure 4. 11 : Scatterplot matrix along with CC (Beck & Latif, 2022).

4.3 Machine Learning-based Predictive Modeling

The primary purpose of this study is to develop a machine learning-based predictive model that is capable to accurately forecast the performance of gas wells from different shale formations. Through exploratory data analysis (EDA) discussed before, which is a critical step in terms of pre-processing and investigation the data along with understanding the distribution of each variable (Operational and reservoir parameters), and relationship between variables and the response (cumulative gas production). These EDA approach help us to develop an appropriate predictive model for our data. This section will discuss the methods used in this study to develop the needed models that will be able to forecast the output variable (cumulative gas production after 1 year).

Mohaghegh (2013) stated that data-driven predictive modeling seem to be the most optimal way for predicting production from unconventional shale formations when comparing to analytical and numerical techniques which tend to make a lot of assumptions to get the model workable. Within reservoir engineering, the majority of machine-learning algorithms belong to the category of supervised learning. (Ertekin & Sun, 2019), where the value of a target variable is predicted based on a number of predictors (input variables) (Mishra & Datta-Gupta, 2018) .

This study utilizes supervised machine learning algorithms, including multiple linear regression and various tree-based methods such as regression tree, bagging, random forest, and extreme gradient boosting machine.

4.3.1 Simple linear regression

Simple linear regression is the simplest method to predict a target variable based on only one single predictor. This can be achieved by assuming a linear relationship between the output Y and the input X as seen in equation 4.7 (James et al., 2021).

$$\boldsymbol{Y \approx \beta_0 + \beta_1 X} \tag{4.7}$$

Where,

$\boldsymbol{Y}$: Target variable (response)
$\boldsymbol{\beta_0}$: Intercept
$\boldsymbol{\beta_1}$: Slop
$\boldsymbol{X}$: Predictor variable

The intercept $\boldsymbol{\beta_0}$ and the slop $\boldsymbol{\beta_1}$ are the model constant coefficient. These coefficients can be produced by using the training data set. Then a future target variable Y can be predicted on a particular value of X using equation 4.8 (James et al., 2021).

$$\hat{y} \approx \hat{\beta}_0 + \hat{\beta}_1 x \tag{4.8}$$

Where,

$\hat{y}$: Prediction of Y

$\hat{\beta}_0$ & $\hat{\beta}_1$: Coefficient estimates

In order to estimate $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ coefficients, before using equation 4.6, a data must be used to be able to obtain the coefficients. For example let the following

$$(x_1, y_1), (x_2, y_{21}), \dots, (x_n, y_n)$$

show $\boldsymbol{n}$ observation pairs of X and Y. For instance, let X denote TV advertising and Y denote sales in $\boldsymbol{n = 200}$ different market. The main goal is to calculate the coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ to let the linear model try to fit the available data as seen in Figure 4.12. In other words, the intercept and slop ($\hat{\beta}_0$ and $\hat{\beta}_1$) should be calculated in order to find a line that is close as possible to the $\boldsymbol{n = 200}$ data points (James et al., 2021).

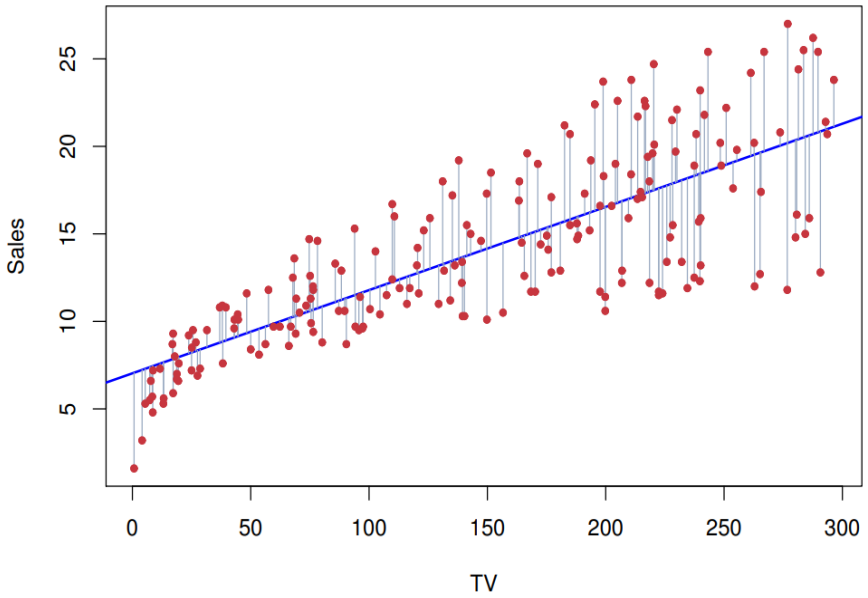


Figure 4. 12 : Simple regression fit (James et al., 2021).

To measure the data fit (closeness), many ways can be utilized. However, the most known approach is minimizing the least squares to obtain the ptimal fit as shown in

the figure 4.12 where the grey lines depict the residuals. The residual sum of squares is calculated as shown in equation 4.9 (James et al., 2021).

$$\boldsymbol{RSS = \left(y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1\right)^2 + \left(y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2\right)^2 + \cdots + \left(y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n\right)^2} \quad (4.9)$$

In order to minimize RSS, the least square approach selects the coefficient of estimates ($\boldsymbol{\hat{\beta}_0}$ and $\boldsymbol{\hat{\beta}_1}$) by doing some calculations listed in equation 4.10 (James et al., 2021).

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (4.10)$$

Where,

$\boldsymbol{\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i}$ & $\boldsymbol{\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i}$ are the sample means (James et al., 2021).

In term of nonlinear relationship between X and Y, there may be hidden variables that could cause variation in Y. Hence equation 4.6 will not be a proper way to make predictions on Y since a measurement error is occurring. Thus equation 4.11 will be utilized along with a random error (James et al., 2021).

$$\boldsymbol{Y \approx \beta_0 + \beta_1 X + \epsilon} \quad (4.11)$$

Where,

$\boldsymbol{\epsilon}$: Mean-zero random error

After building the linear regression model, the model accuracy evaluation is a mandatory process to quantify to which degree the model is fitting the data. Thus two main assessing metrics for linear regression is used namely, the R^2 statistic and the residual standard error (RSE) (James et al., 2021).

Residual standard error:

Residual standard error (RSE) is defined as the difference between the responses and the true regression line. Recalling equation 4.6 which is depicting each observation with an error term ($\boldsymbol{\epsilon}$) . These errors prevent us from perfectly predicting Y from X. Even if the coefficient of estimates ($\boldsymbol{\hat{\beta}_0}$ and $\boldsymbol{\hat{\beta}_1}$) that represent the regression line are known. Thus, RSE serves as an estimate for the standard deviation of ϵ. To calculate it equation 4.12 is used (James et al., 2021).

$$RSE = \sqrt{\frac{1}{n-2}RSS} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (4.12)$$

Where, the residual sum of squares $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

R^2 statistic:

While dealing with *RSE* which provide an absolute measure for the deviation of Y from the regression, R^2 statistic define an alternative method for measuring the fit of data by taking a scale value ranges between from 0 to 1. Hence, it is independent of the unit of Y, and it is computed using equation 4.13 (James et al., 2021).

$$R^2 = \frac{TSS-RSS}{TSS} = 1 - \frac{RSS}{TSS} \quad (4.13)$$

Where, $TSS = (y_i - \hat{y}_i)^2$ is defined the total variance in Y.

Statistical significance and P-value:

Statistical significance is methods used by researchers to determine whether the observed relationship between two variables is statistically significant. First, a null hypothesis that suggest that there is no relationship between the variables is established. Then an alternative hypothesis is trying to challenge the null hypothesis by measuring how unlikely the test statistic would result if the null hypothesis were true. The resulting number from this situation is called the p-value. Where the p-value is an evidence against null hypothesis. A small value of p-value indicating a strong evidence against the null hypothesis. The usual criterion used by researchers for small p-value to be considered statistically significance is when it is less than 0.05 or 0.01 (Utts, 2014).

4.3.2 Multiple linear regression

Predicting a target variable based on one single predictor is accomplished by using simple linear regression which we explained above. However, in practice, usually more than one predictor is existing in the model. Thus, instead of applying many separate linear regression models which is not satisfactory, a better approach is to conduct one linear regression that capable of accommodating multiple predictors. This approach called multiple linear regression which is able to assign each predictor a distinct slop coefficient resulting with the following form represented by equation 4.14 (James et al., 2021).

$$\boldsymbol{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon} \quad (4.14)$$

Where $\boldsymbol{X_j}$ is the $\boldsymbol{j}$**th** predictor and $\boldsymbol{\beta_j}$ is the coefficient that determine the relationship between that variable and the target one. The regression coefficients $\boldsymbol{\beta_j}$ is the average impact on Y when $\boldsymbol{X_j}$ is increased by one unit while keeping other variables fixed. Since the regression coefficients in equation 4.8 are unknown and must be estimated, a similar scenario of conducting simple linear regression case will be repeated along with equation 4.9 to generate prediction using the $\boldsymbol{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p}$ estimates (James et al., 2021).

$$\boldsymbol{\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p} \quad (4.15)$$

After estimating the parameters using the same approach as in simple regression, $\boldsymbol{\beta_1, \beta_2, \ldots, \beta_p}$ are selected to make the residual sum of squares minimized as possible.

$$\begin{aligned} \boldsymbol{RSS} &= \boldsymbol{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \\ &= \boldsymbol{\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip}\right)^2} \end{aligned} \quad (4.16)$$

The least square regression is transformed from line into plane (Figure 4.13) wen there is more than one predictor and one response. The following figure illustrate three predictor and one response in three dimensional setting to reduce the sum of squared between each observation represented by a set of red points and the plane (James et al., 2021).

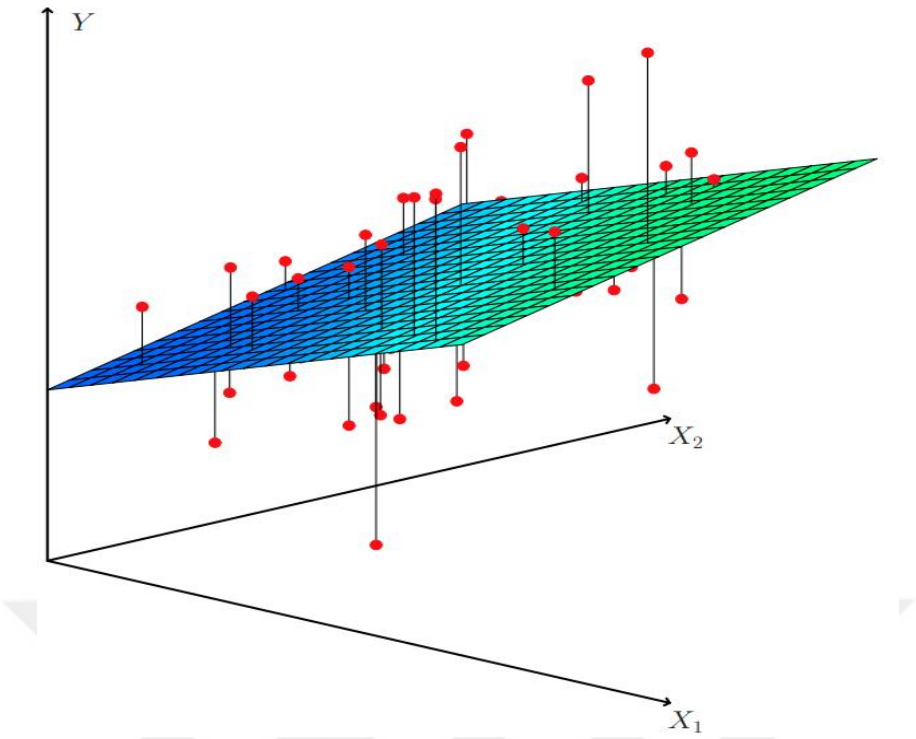


Figure 4. 13 : Multiple linear regression model fit (James et al., 2021).

When applying multiple linear regression, penalty of important questions that require answers are arising, such as (1) can any of the predictors be considered useful to forecast the response? (2) Do all the predictors describe the response or just subset of them? (3) To what extent the model represent and fit the data? (4) What is the level of prediction accuracy (James et al., 2021)?

F-statistic

Evaluation of the multiple linear regression model would be instrumental in addressing these questions. To achieve that, it involves examining the relationship between predictors and the response as well as validating the linear model using an analogous statistic, namely the F-statistic (Mishra & Datta-Gupta, 2018). F-statistic is given by equation (4.17).

$$F = \frac{(TSS-RSS)/p}{RSS(n-p-1)} \tag{4.17}$$

F-statistic is utilized to test the hypothesis whether the regression coefficient are zero, $\boldsymbol{\beta_1 = \beta_2 = \cdots = \beta_n = 0}$. This null hypothesis should be rejected using an alternative

hypothesis that prove at least one of the coefficient is not zero based on evidence from the data. This could be achieve by obtaining a large value of F-statistic or low value of p-value (Mishra & Datta-Gupta, 2018). In other words, value of F-statistic close to 1 would reveal that predictors are not explaining the response. Hence, there is no evident relationship between the response and predictors. In contrast, if the F-statistic value is far larger than 1, this would be a strong evidence against null hypothesis. Thus, an association between response and predictors is existed. Moreover, a value of F-statistic that is even a little larger than 1 could also be a good indicator to reject the null hypothesis (James et al., 2021).

Variable selection:

Assessing multiple linear regression model requires initially to identify the F-statistic and p-value to detriment if any of the predictors is correlated with the response. If there is so, then it is important to identify which is the guilty ones that does not relate. In practice, the target variable (response) can be often associated with only a portion of the predictors. In some cases, all the predictors could be related to the response. Therefore, multiple models can be created based on the selected variables as we can start with no variables model, gradually adding variables until we reach a satisfactory model that best fit the data. For instance, having 30 $\boldsymbol{p}$ variables could generate $\mathbf{2^{30}}$= 1,073,741,824 models and this is not practical (James et al., 2021). Thus, an automated approaches are required to choose best a few models for evaluation such as forward selection approach and backward selection approach. After determining the variables that make up the model, several statistic metrics can be used to evaluate the model quality, such as Akaike information criterion (AIC), Bayesian information criterion (BIC), Mallow's Cp, and adjusted R^2 (James et al., 2021). These statistics will be explained in details below.

Model Fit:

Measuring the model fit is essential to determine the quality of the model. Performing this requiring two commonly used numerical measuring metrics, namely RSE (Residual Standard Error) and R^2, which represent the proportion of the variance explained. These metrics are computed similarly to how they are in simple linear regression. However, in multiple linear regression, R^2 signifies the square of $\boldsymbol{Cor}\left(\boldsymbol{Y}, \widehat{\boldsymbol{Y}}\right)^{\mathbf{2}}$ between the response and the predicted value of the response (fitted

linear model). When R^2 is close to 1, it implies a better fit. Hence explaining a large portion of the variance on the target variable (response) Y (James et al., 2021).

4.3.2.1 Best subset selection

In this section, we need to identify the best subset of predictors that perfectly represent our model. Starting by fitting separate regression models for a combination of the p predictors. This process including models with one predictor, two predictor and so on until we reach up to models with all predictors. This can be achieved by performing separate p models with exactly one predictor, then fit $(\boldsymbol{2^p}) = \boldsymbol{p(p-1)/2}$ models that include every possible pairs of predictors and fit a separate regression model for each pair of variables and continue this process with three, four predictors, and so on, up to the model with all p predictors. Finally, we check the results from the performed models to select the best model (James et al., 2021).

As stated by James et al. (2021), it is essential to note that selecting the best model from $\boldsymbol{2^p}$ model can be computationally expensive. Therefore, some algorithms can help in selecting a single model that has the highest predictive accuracy or provides the most meaningful insights. These algorithms including the use of cross-validation, Cp, BIC, AIC, or adjusted R^2. These metrics are explained in the next section of deciding on the best model.

4.3.2.2 Best Model Selection

After we obtain a preliminary knowledge on selecting different subset of predictors for multiple models, we need to identify which of these models will best fit the data. As explained before RSS and R^2 can be used to assess a regression model. However including all predictors in a model usually results in a small RSS and large R^2 model. This is because these metrics are associated with the training error. Therefore, the model that have a low test error is required since RSS and R^2 are not the proper measures for choosing the best model among multiple models with varying numbers of predictors (James et al., 2021).

Estimating the test error for best model selection can be done using two main common approaches. (a) Indirectly, by adjusting training error to estimate test error. (b) directly, by using cross-validation or validation set approach to estimate test error (James et al., 2021). These approaches are explained below.

Mallow's Cp

This technique is used for adjusting the train error and estimate the test error for a least squares model fitted with d predictor variables. To calculate the Cp estimate of test mean squared error (MSE), equation 4.18 is used.

$$\boldsymbol{C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)} \tag{4.18}$$

Where $\boldsymbol{\hat{\sigma}^2}$ is representing the error variance ϵ and is typically obtained using all predictor variables in the model. To adjust the train error (RSS), Cp adds $\boldsymbol{2d\hat{\sigma}^2}$ term to the training RSS. Among a set of models, Cp statistic tends to choose the model that has the low test error. Hence, the best model is the one with the lowest Cp value (James et al., 2021).

Akaike information criterion (AIC)

The AIC is an estimator of test error that is applicable for a wide range of models using maximum likelihood estimation. Maximum likelihood and least squares are similar when using the model in equation 4.8 with Gaussian errors. The AIC and Cp criteria exhibit a direct proportionality to each other in the case of least squares models. (James et al., 2021).

Bayesian information criterion (BIC)

The initial derivation of the BIC originates from a Bayesian approach. Despite this, it exhibits similarities to both Cp and AIC. Hence to choose a model, it should correspond to the lowest BIC. Equation 4.19 is used to calculate BIC (James et al., 2021).

$$\boldsymbol{BIC = \frac{1}{n}(RSS + \log(n)2d\hat{\sigma}^2)} \tag{4.19}$$

Adjusted R^2

The adjusted R^2 is a widely common statistic measure for choosing the best model among various models with distinct number of predictor variables. As discussed before, R^2 is always increasing by adding more variables to the model. In contrast, the

RSS is decreasing. However, for least square models with d predictors equation 4.20 can be utilized to calculate the adjusted R^2 (James et al., 2021).

$$\boldsymbol{Adjusted\ R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}} \quad (4.20)$$

Where a model with adjusted R^2 value close to 1 is a good indicator of a model with low test error, which is the opposite of Cp, BIC, and AIC, where a smaller value suggests a model with low test error

The validation set approach

Validation set approach is a statistical technique to measure the test error linked with a specific set of observations by splitting the data into two portions, namely train set and validation set. The train set is employed to fit or train the model, whereas the validation set is utilized to predict the response (James et al., 2021). The resulted test error from the validation set method is usually evaluated by mean square error (MSE). Figure 4.14 illustrated the approach of validation set, where the data consist of n observations is splitted into training and validation sets shown in blue and beige colors respectively.

Figure 4. 14 : A schematic of the validation set approach (James et al., 2021).

The validation set approach is relatively simple and uncomplicated while implementation. However, two possible drawbacks of this method should be considered (James et al., 2021).

1- The estimated test error can vary significantly based on the observations chose for the training and validation sets.

2- As the model only use small portion of the entire data for training, the model could perform worse and the validation set could overestimate the test error.

Therefore, to refine the set validation approach and fixing its drawbacks, cross-validation techniques will be explained below.

Leave-one-out cross-validation (LOOCV)

LOOCV shares similarities with the validation set method. However, instead of splitting the data into two random portions for training and validation, only one observation is leaved for validation and the all other observations is used for training. Therefore, the prediction is made only on the excluded observation while the model is fitted to n-1 training observations (James et al., 2021). Figure 4.15 is illustrating the LOOCV, where the data is splitted repeatedly into n observations for training shown in blue color and leave one observation for validation shown in beige color. This procedure is repeatedly done to estimate the test error by averaging the errors from each procedure.

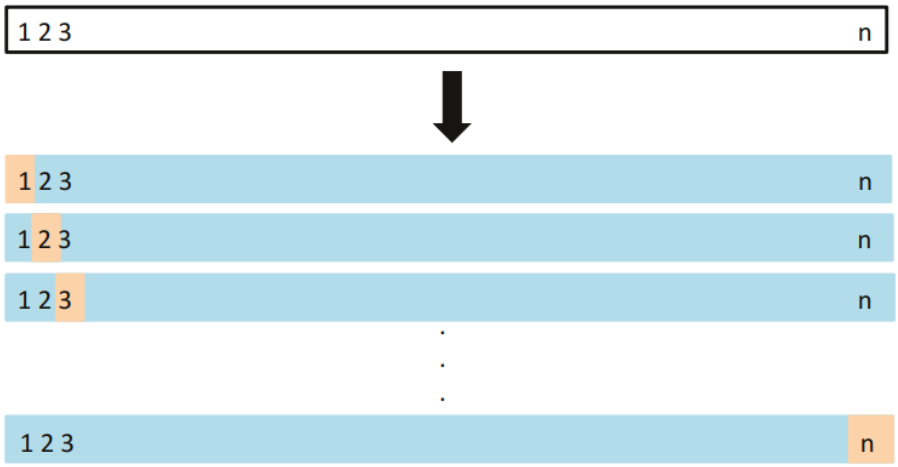


Figure 4. 15 : A schematic of LOOCV (James et al., 2021)

k-Fold cross-validation

Cross-validation approach started by splitting the training data into k random groups, also known as folds. Then, the model is trained using k-1 folds while the other fold is left out for validation and prediction purposes. The whole k folds is cycled in order to obtain validation prediction from each observation. The predictions is obtained from the observations that were not used in the training dataset. It is possible to extend the

cross-validation by repeated the entire process with different folds selected randomly and thus will generate distinct predictions on each observation (Schuetter et al., 2018).

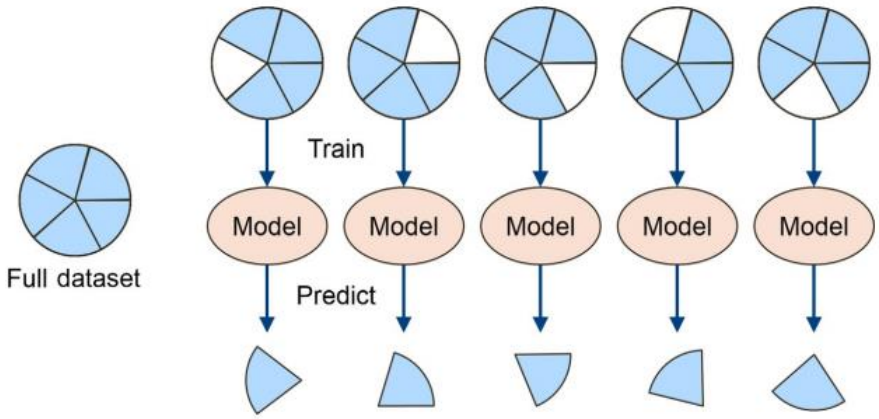


Figure 4. 16 : k-Fold cross-validation representation (Schuetter et al., 2018)

By repeatedly training and validating the model using different subsets of the data, cross-validation can be used just for evaluation purpose and it grants a more reliable indication of how well and predictive the model is expected to work on unseen data for future applications (Schuetter et al., 2018).

4.3.2.3 Goodness of fit

Numerous common metrics exist for measuring model goodness of fit. In this context, we will discuss two of them that will be used in this study. (1) Average absolute error (AAE), and (2) mean squared error (MSE).

Average absolute error: the calculation of average absolute difference between the observed values and the predicted ones is referred to as the average absolute error, also known as the average size of the residuals (Mishra & Datta-Gupta, 2018). Equation 4.21 is used to calculate it.

$$\boldsymbol{AAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|} \qquad (4.21)$$

Mean squared error: The MSE metric is similar to the AAE. However, MSE takes into account the average squared difference between actual and prediction values as shown in equation 4.22 (Mishra & Datta-Gupta, 2018).

$$\boldsymbol{AAE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (4.\ 22)$$

In term of units, AAE uses the same unit of the target variable, while MSE uses the squared unit. It is familiar to mention that RMSE is s a popular variation of mean squared error (MSE), estimated by taking the square root of MSE. A small value of RMSE (i.e., values closer to zero) indicating better model performance as the deviation between actual and prediction values is small (Mishra & Datta-Gupta, 2018). In general, Navidi, (2008) expressed that MSE or its square root RMSE is usually prioritized over the AAE because of its well-established distribution (as cited in Mishra & Datta-Gupta, 2018).

4.3.2.4 Regression assumptions and diagnostics

After we build the regression model, now it is essential to check if the model is suitable for our data or not. To do this, the regression model should validate the regression assumptions since the model is considered as good as it validates its assumptions (Lesmeister, 2019). These assumptions are listed below.

Linearity:

This assumption assumes that there is a linear relationship between the target variable and predictors. If not, transformations techniques on response or predictor variables will help in solving nonlinearity (Lesmeister, 2019). Linearity can be checked using residuals vs fitted diagnostic plot as depicted in Figure 4.17

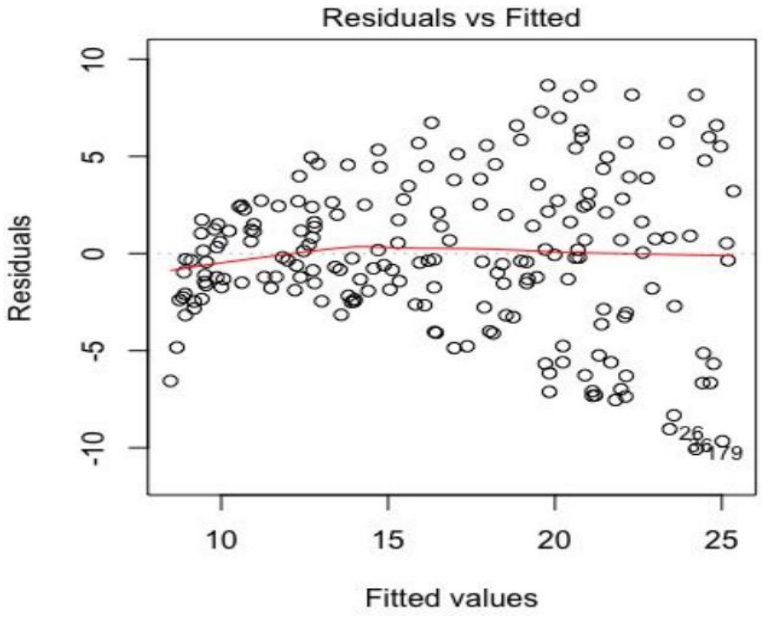


Figure 4. 17 : Residuals vs fitted values (Kassambara, 2018)

When the residual vs fitted plot present the red dashed line approximately horizontally at zero, that is mean no pattern is exist. If a pattern exist (may be, u-shape), it would be an indicator of non-linearity. If non-linear relationship is detected, some transformations (i.e., log(x), $\sqrt{\mathbf{x}}$ or x^2) of the predictor variables is needed (Kassambara, 2018).

Homoscedasticity:

This assumption assumes constant variance of errors (residuals) around the input values. When this assumption is violated, it is referred to as heteroscedasticity (Lesmeister, 2019). In the diagnostic plots, the scale-location plot as shown in Figure 4.17 is used to check the homoscedasticity (Kassambara, 2018).

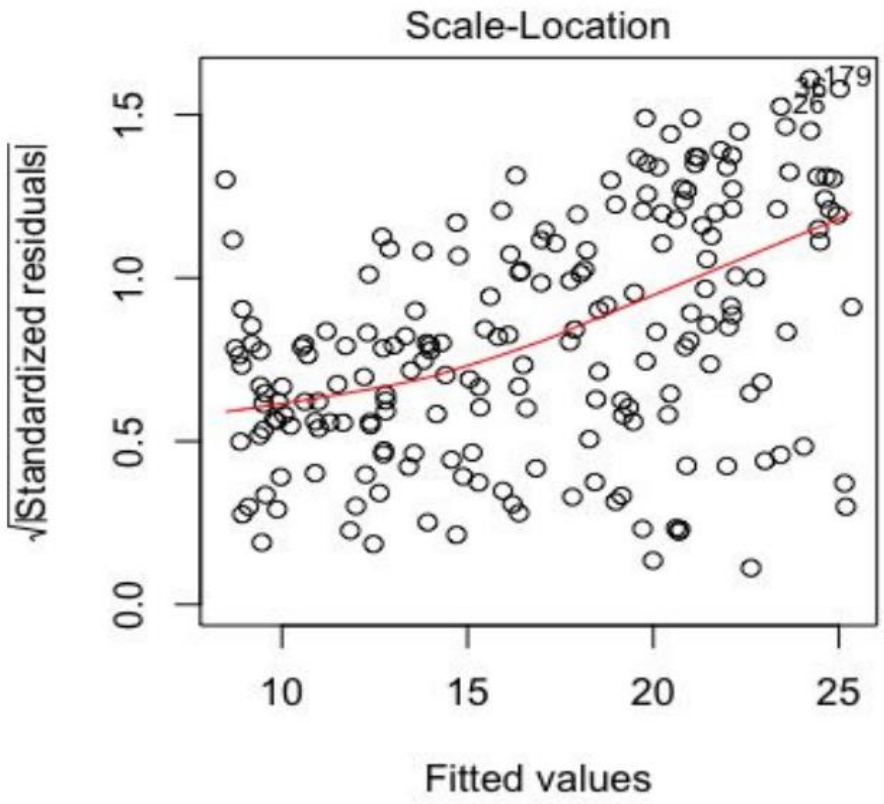


Figure 4. 18 : Scale-location plot for homoscedasticity (Kassambara, 2018).

If the data points (residuals) around the red horizontal line, which is not the case of figure 4.17, are normally (equally) distributed, this would valid the assumption of homoscedasticity. Log transformation of the response (y) would be a good solution for the violation homoscedasticity (Kassambara, 2018).

Normality of residuals:

This assumption supposes that the residual errors are distributed normally. The Normal Q-Q plot is used to represent graphically the quantile values of two variables against each other. This assumption could be violated by the outlier values (Lesmeister, 2019). Figure 4.19 shows approximately normality of the residuals by following the straight dashed line.

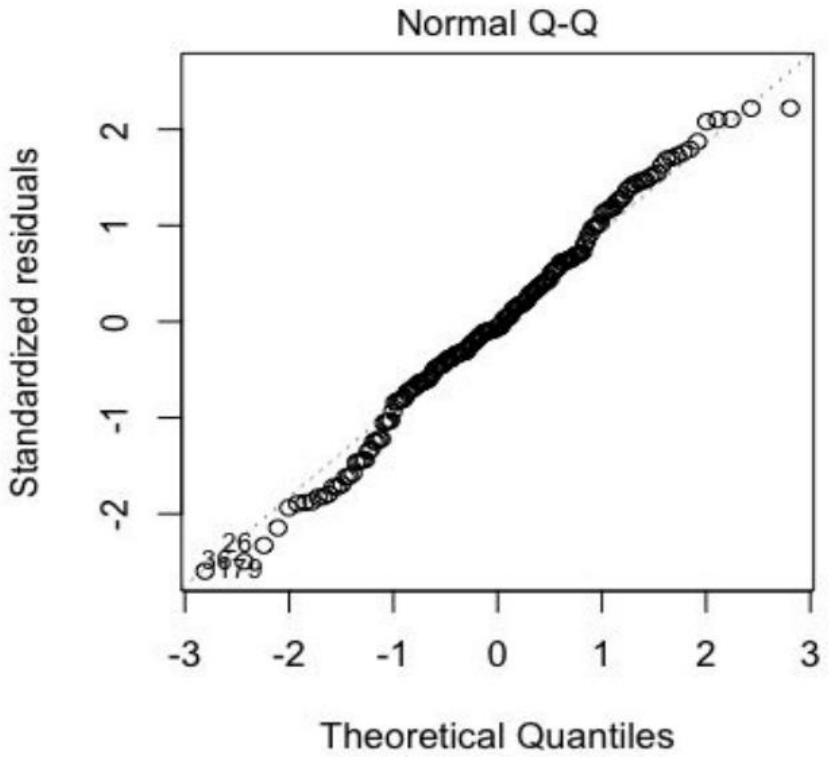


Figure 4. 19 : Normal Q-Q plot of residuals (Kassambara, 2018).

Presence of influential values:

This assumption is used to determine if our dataset contain influential values such as outliers and high-leverage points. While the outlier values are influential in the response variable (y), the leverage points are extreme in the predictors (x). These values can significantly impact our model hence it should be detected (Kassambara, 2018).

Figure 4.20 depict the residual vs leverage plot in order to check the potential outliers or high leverage points. It depicts that there is no outliers since the observations with the top extreme points (#179, #26 and #36) are under the absolute value 3 of standardized residuals (James et al., 2021).

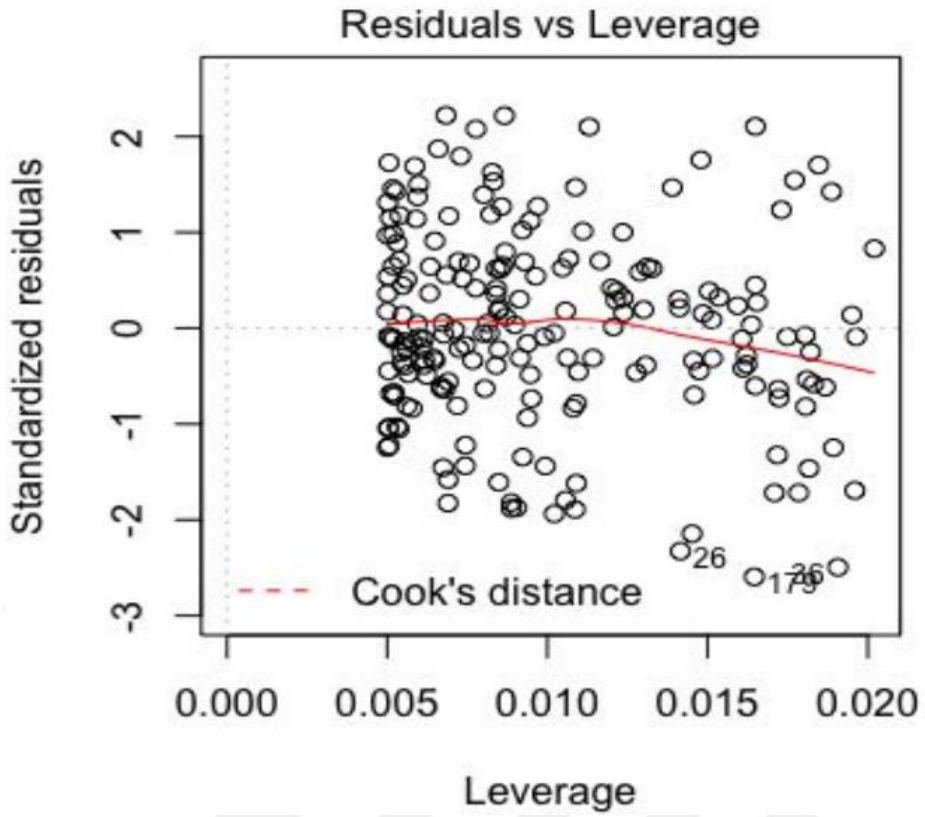


Figure 4. 20 : Residuals versus leverage plot (Kassambara, 2018).

Furthermore, the high leverage points are determined by a common metric called hat-value or leverage statistic which assume values above 2(p + 1)/n as points with substantial leverage (Bruce et al., 2020). Where p and n represent the size of predictors and the number of observations respectively.

4.3.3 Tree-based methods

In this study, tree-based methods is utilized including regression tree, bagging, random forest (RF), and extreme gradient boosting machine (XGBoost) to build data-driven based model that is able to predict the cumulative gas production after one year from shale formations. These methods are explained below.

4.3.3.1 Classification and Regression Trees (CART)

CART can be defined as a tree based technique that is relatively simple and highly interpretable. CART main objective is to identify how the predictors affecting the respond (Breiman et al., 1984) (as cited in Mishra & Datta-Gupta, 2018). The fundamental principle of a decision tree is partitioning the predictor variables by starting with the variable that reduce the RSS the most. The first split is called the root

node (Lesmeister, 2019). The root node is defined as the most important predictor variable. As shown in Figure 4.21, the binary splits continue leading to a new node called decision nodes, which are the variables that have less degree of influence than the root node. When the decision node has no more splits, it will lead to the final terminal node (leaf node), where the predicted mean value of the response is reached (Favero et al., 2023).

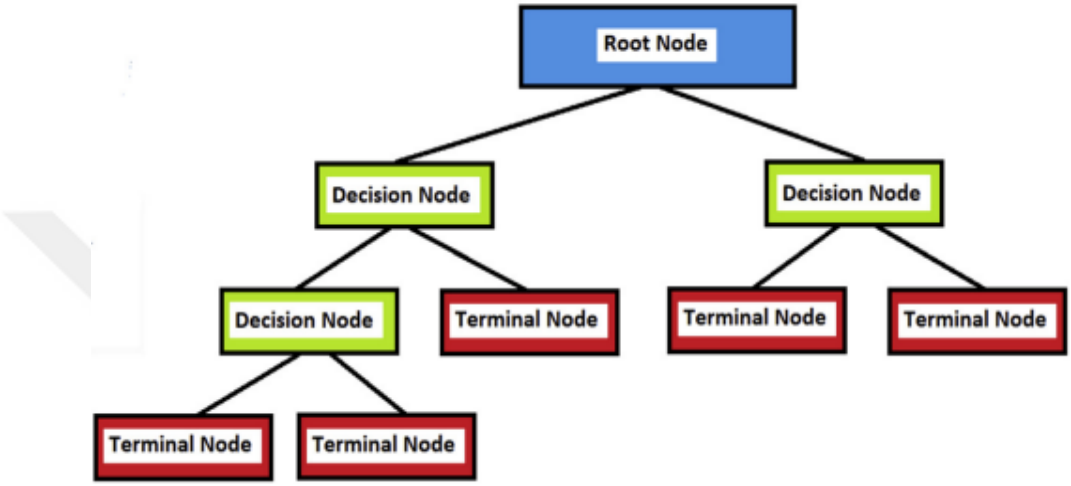


Figure 4. 21: Decision tree example (Belyadi & Haghighat, 2021)

This top-down approach can be illustrated using rectangular regions as shown in Figure 4.22 where each region representing the mean predicted value of each observation that fall in that region (James et al., 2021).

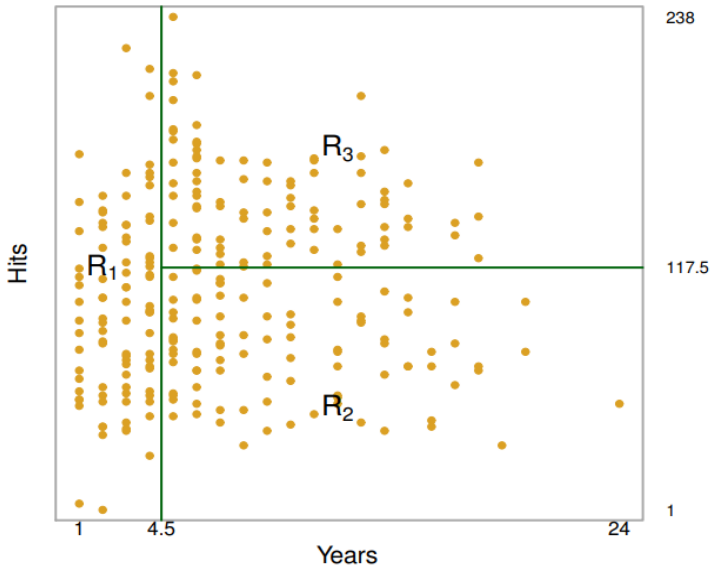


Figure 4. 22: Regression tree using rectangular regions (James et al., 2021).

Since the dataset in our study does not consist of categorical variables, we will explain only the fundamentals of building regression tree. James et al. (2021) introduced two main steps for building regression tree.

1- The space of predictors that consist of all available values of $X_1, X_2,...,X_P$ will be partitioned to J different $R_1, R_2,...,R_J$ and non-overlaping regions.

2- The observations that fall to the same region R_J will have identical prediction values, equating to the mean of response values that are within that specific region.

For simplicity and ease interpretation, the space of predictors was partitioned into rectangle regions in order to spot the regions $R_1,...,R_J$ that reduce the RSS the most (James et al., 2021). This is expressed by equation (4.23).

$$\sum_{j=1}^{J}\sum_{i\in R_j}(y_i - \hat{y}_{R_j})^2 \qquad (4.23)$$

Where,

$\hat{y}_{R_j}$: Mean response value of a training observation

However, it is unfeasible to evaluate all potential partitions of the predictor space into J regions. Therefore, a greedy approach namely top-down is used, starting from the root node of the tree and continue splitting the feature space until reaching the destination so-called terminal node. It called greedy since at each time of splitting the tree, the best split is selected without regard to the later splits that may lead to a better tree. Conducting a top-down method or so-called recursive binary splitting initially require the selection of predictor variable X_j as well as the cutpoint s, and so dividing the predictor variable space to $\{X|X_j < s\}$ and $\{X|X_j \geq s\}$ regions that help in generating the best possible decrease in RSS. With this splitting method, each predictor cutpoint s value and all predictors $X_1, X_2,...,X_P$ are considered in order to obtain the ccutpoint and predictor that produce the tree with the minimum RSS (James et al., 2021). Furthermore, for any j and s, half-planes that divide the predictor space into distinct regions are defined using equation (4.24).

$$R_1(j,s) = \{X|X_j < s\} \text{ and } R_2(j,s) = \{X|X_j \geq s\} \qquad (4.24)$$

Then, we find specific values for j and s that make equation 4.25 minimized.

$$\sum_{i:x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2 \quad (4.25)$$

Where,

$\hat{y}_{R_1}$: Mean response value of a training observation in region one.

$\hat{y}_{R_2}$: Mean response value of a training observation in region two.

The process of finding the values of j and s in equation 4.18 is repeated in order to find the best further splits that reduce the RSS. This process can be achieved fast if the number of predictor is not large (James et al., 2021). However, with large number of predictors, you may have a complete tree of unwanted branches with a high variance and low bias causing overfitting. To avoid this, the tree must be properly pruned to obtain the best size of the tree (Lesmeister, 2019).

The pruned tree can be achieved by conducting cross-validation or validation set approaches to estimate the minimum test error that correspond the best subtree. Cost complexity pruning metric also can be more effectively since it is not looking at each subtree, instead it examine a series of trees organized with a nonnegative tuning parameter as an index (James et al., 2021).

4.3.3.2 Bagging

Bagging is decision tree based predictive model, also known as bootstrap aggregation, used in general-purpose to reduce the variance within a noisy dataset by using multiple prediction trees rather than one single tree on different subset of a training data. These prediction trees mostly use the strong predictors on the top splits. The predictions generated from these splits are then averaged or taking the most frequently occurring prediction by majority voting to obtain the final prediction (James et al., 2021).

The exception of bagging when compared to basic regression trees is instead of fitting several models to the same training data, bagging applying a new model for each bootstrap resample (Bruce et al., 2020). The bootstrap resample technique is iteratively choosing a random sample of n observations from the training data and assessing the model on each sample. Then the average error from these samples is computed to check the total variance of the model performance (Kassambara, 2018).

4.3.3.3 Random Forest (RF)

Random forest is an improvement over bagging method. Instead of building the decision trees based on the strong predictors, Random forest choose subset of the predictors randomly hence not only the strong predictors but all other predictors will contribute to the final prediction (James et al., 2021). Concisely, RF was defined by Breiman et al. (1984) as an ensemble regression tree method with each tree being trained using random subset of predictors (as cited in Mishra & Lin, 2017). The main concept of building random forest model is initially by generating a group of regression trees, as shown in Figure 4.23, which is developed using a random subset of observations and predictor variables.

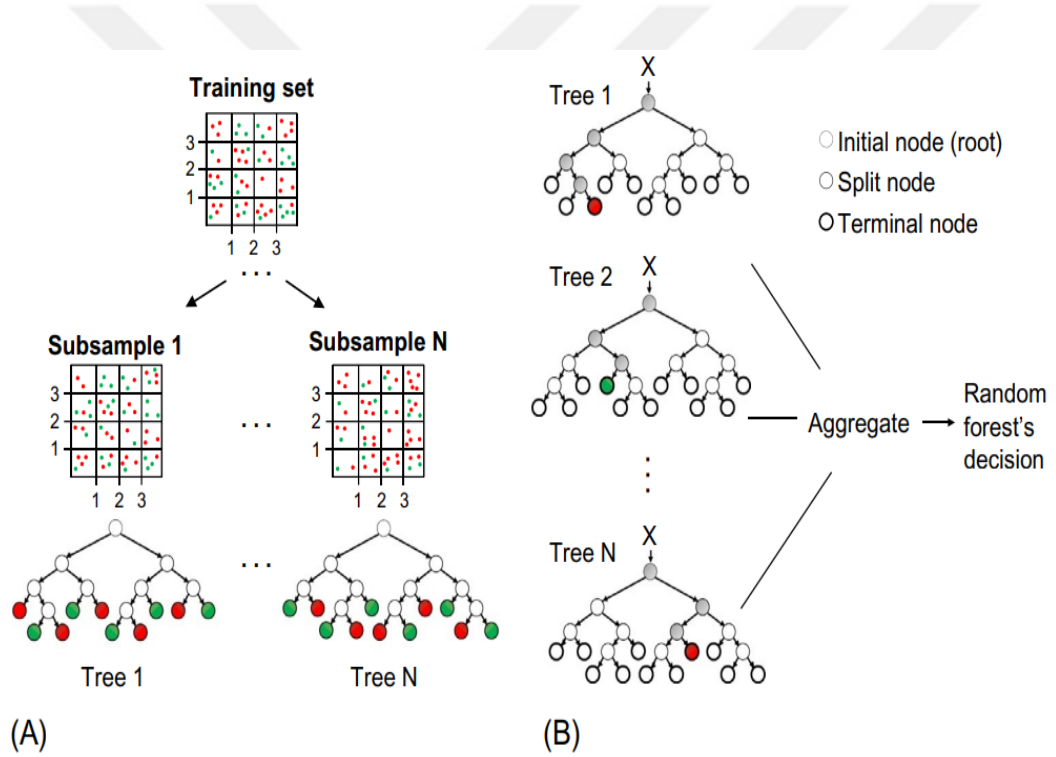


Figure 4. 23 : RF modeling process (Mishra & Datta-Gupta, 2018)

For prediction in random forest, in order to obtain a unique regression estimate from each tree, each observation should be involving through all the trees in the ensemble. Consequently, the average of all these regression estimates is the final prediction of the model (Mishra & Datta-Gupta, 2018). The RF algorithm provide an appropriate built-in cross-validation feature that simplifies the process of validating the prediction model. This feature allows to easily assess the model performance and its accuracy (Mishra & Datta-Gupta, 2018).

4.3.3.4 Extreme gradient boosting (XGBoost)

XGBoost is machine learning system that is scalable and was introduced by Chen and Guestrin in 2016, that overcome the other existing methods in term of fast respond and accuracy due to the availability of algorithms optimizations (Chen & Guestrin, 2016). XGBoost aims to build several classification and regression trees and keep adding them sequentially resulting with a unique prediction (Tang et al., 2021). The success of XGBoost over the other tree ensemble methods is owed to its exceptional scalability that make it an efficient solution in all scenarios along with its speed efficiency in tree construction, ten times faster than the alternative existing models. The scalability of the XGBoost is coming from availability of several important algorithmic optimizations (Chen & Guestrin, 2016). The final prediction can be estimated by summing up the predictions of all individual trees, see Figure 4.24 (Chen & Guestrin, 2016).

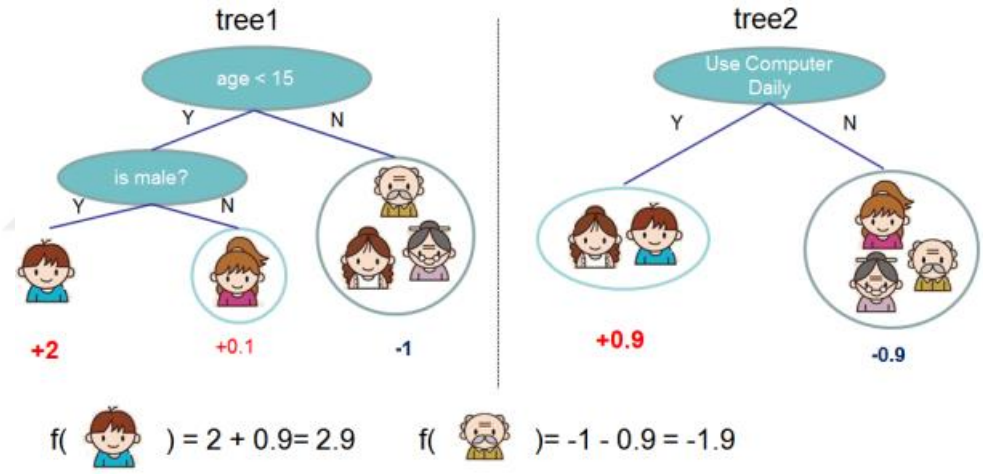


Figure 4. 24 : XGBoost final prediction estimation (Chen & Guestrin, 2016).

Considering a dataset $D = \{(x_i, y_i)\}$ that consist of *n* observations and *m* predictors, in order to predict the response XGBoost (a tree ensemble model) employs K additive functions given by equation 4.26 (Chen & Guestrin, 2016).

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \ , \quad f_k \in \mathcal{F} \tag{4.26}$$

Where,

$\mathcal{F}$: Regression tree space also known as CART

f_k : Corresponds an independent tree with leaf scores

In XGBoost, equation 4.27 is utilized to minimize the objective function to acheive the optimal tree structure (Chen & Guestrin, 2016).

$$\mathcal{L}(\emptyset) = \sum_i \boldsymbol{l}(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_i) + \sum_i \boldsymbol{\Omega}(\boldsymbol{f}_k) \quad (4.27)$$

Where $\boldsymbol{l}$ is the difference between the estimated variable $\hat{\boldsymbol{y}}_i$ and the observed response $\boldsymbol{y}_i$, also known as loss function and $\boldsymbol{\Omega}$ is a regularization term used to prevent overfitting by penalizing the model's complexity (Chen & Guestrin, 2016). Where $\boldsymbol{\Omega}$ can be written as in equation (4.28).

$$\boldsymbol{\Omega}(\boldsymbol{f}) = \boldsymbol{\gamma}\mathbf{T} = \frac{1}{2}\boldsymbol{\lambda}\|\boldsymbol{w}\|^2 \quad (4.28)$$

Where,

T : Tree leaves number

$\boldsymbol{w}$: Score of prediction on every leaf

Finally, equation 4.23 can be derived and written below.

$$\mathcal{L}^t = \sum_{j=1}^{T}\left[(\sum_{i\in I_j} \boldsymbol{g}_i)\boldsymbol{w}_j + \frac{1}{2}\left(\sum_{i\in I_j} \boldsymbol{h}_i + \boldsymbol{\lambda}\right)\boldsymbol{w}_j^2\right] + \boldsymbol{\gamma}\mathbf{T} \quad (4.29)$$

Where, g_i and h_i represent the loss function of first order and second order gradient statistics respectively. The parameters λ and γ are coefficients that play a crucial role in regulating the model. They are utilized to prevent overfitting, ensuring that the model generalizes well to unseen data (Chen & Guestrin, 2016).

5. RESULTS AND DISCUSSION

In this chapter, we will illustrate and analyze the results achieved through data analytics, statistical and machine-learning based modeling techniques. Furthermore, this chapter aims to identify the primary controlling parameters (both reservoir and operational parameters) governing the performance of shale gas wells in unconventional reservoirs. The analysis of this study was run in the following manner:

- Descriptive statistics was employed to gain an overview and insight into the data.
- Histograms and boxplots was employed to represent the data graphically.
- Employ scatterplots to visually understand the data as well as the relationship among the parameters.
- Employ correlation analysis to describe the relationship between the gas production and each input parameter.
    - Employ supervised machine-learning modeling like, MLR and tree-based methods to forecast the performance of shale gas wells.
- Identify the key driver parameters of the gas production in unconventional shale reservoirs.

5.1 Descriptive Statistics

It is a good way to start our discussion through descriptive statistics to get a quick overview of our dataset. This will help us to gain insightful information about the dataset variables along with their statistics such as the range of values of each variable (minimum and maximum), the mean, standard deviation, variance, and as well as the skewness. The descriptive statistics for our dataset is conducted separately for reservoir parameters and as well as operational parameters.

5.1.1 Reservoir properties

Table 5.1 displays reservoir parameters, demonstrating the summary statistics for these parameters employed in this study. The statistics include the range values of reservoir parameters (min. and max. values), mean ($\bar{x}$), standard deviation (σ), variance (σ^2),

and the skewness. The range is an easy way to examine the variability of each reservoir parameter by depicting its minimum and maximum values. The mean value in Table 5.1 is utilized to quantify the central tendency of the data for each reservoir parameter. The data spread around the mean can be described using the variance and standard deviation. As it seen from the Table 5.1 the all parameters have some and a lot of spread around the mean. Table 5.1 as well as shows that skewness of the parameters; some parameters are nearly symmetrical such as reservoir thickness, water saturation, and porosity. Some of the parameters are slightly skewed such as initial pressure and reservoir temperature and finally we have parameters that are extremely positive or negative skewed such as gas saturation, gas specific gravity, concentration of nitrogen, and static well head temperature.

5.1.2 Operational properties

Table 5.2 displays operational properties, demonstrating the summary statistics for these parameters employed in this study. The statistics include the range values of reservoir parameters (min. and max. values), mean ($\bar{x}$), standard deviation (σ), variance (σ^2), and skewness. The general skewness of operational parameters is nearly symmetric since the descriptive values of skewness is low. A few exception is exist for the static Wellhead Temperature (Twhs), clusters and clusters per stage which show higher positive skewness values than the remaining ones. Hence, operational properties tend to depict a better distribution when compared to reservoir parameters.

Table 5. 1 : Reservoir parameters descriptive statistics

Parameter Description	Range(Min-Max)	$\bar{x}$	σ	σ^2	skewness	unit
Cumulative Gas Production (Gp)	21.97-13077.61	4132	3334	1.1E+07	0.61	MMscf
Initial Pressure (Pi)	2200-12223	6274	2881	8302365	0.77	psi
Reservoir Temperature (Tr)	115-379	211	90	8237	0.59	deg F
Net Pay Thickness (h)	56.48-268.4	157	59	3563	-0.17	ft
Porosity (Ø)	0.05-0.1	0.0699	0.013	0.00018	0.33	%
Water Saturation (Sw)	0.18-0.47	0.30	0.085	0.0073	0.27	%
Gas Saturation (Sg)	0-0.82	0.56	0.298	0.089	-1.21	%
Gas Specific Gravity (Yg)	0.57-0.99	0.63	0.12	0.015	2.11	%
CO_2	0-0.06	0.012	0.018	0.00031	1.07	%
N_2	0-0.0045	4.6E-04	0.0011	1.2E-06	3.08	%

Table 5. 2 : Operational parameters descriptive statistics

Parameter Description	Range(Min-Max)	$\bar{x}$	σ	σ^2	skewness	unit
True Vertical Depth (TVD)	5707-12668	8952	2118	4487581	0.77	ft
Stage Spacing	700-1850	1215	292	85295	0.14	ft
Number of stages	7-89	46	20	8237405	-0.17	#
Number of clusters	49-1035	317	231	356353365	1.50	#
Clusters per Stage	3-15	6.57	2.96	0.000188.77	1.67	#
Proppant	3591544-42946954	2.2E+07	9572924	0.00739.2E+13	0.23	lbs
Lateral Length	2268-13011	8019	2371	5623709	-0.062	ft
Top Perforation	5900-13153	9133	2179	4748884	0.81	ft
Bottom Perforation	10049-23203	17136	3521	1.2E+07	0.23	ft
Sandface Temperature (Tsf)	115-379	210	88	7880	0.59	deg F
Static Wellhead Temperature (Twhs)	60-236	94	48	2353	2.02	deg F

5.2 Univariate Analysis

Univariate data analysis was employed in this study to visually identify the characteristics of reservoir and operational parameters in one dimension. This means each of reservoir and operational parameter will be presented graphically and individually using boxplots and histograms. This will give us an overview of the distribution of the data including the spread, symmetry and the degree of skewness. Furthermore, it will be useful to detect if there is outlier points in the data.

5.2.1 Reservoir parameters boxplots

Figure 5.1 illustrates the reservoir parameters constructed graphically using boxplots. It can be observed that the boxplots confirm what we discussed earlier using the descriptive statistics. For instance, reservoir thickness, porosity and water saturation boxplots are depicting a nearly symmetry distribution since the median is nearly positioned at the middle of the box (interquartile range) which means that the data of these parameter are evenly distributed around the median. In addition there is no outlier points above or below the whiskers.

In contrast, for initial pressure, reservoir temperature, and carbon dioxide boxplots are depicting the median at lower position inside the box (interquartile range). This indicate that these parameters have skewness to the right (positive skewness) since the upper side whisker exceeds the length of the lower whisker. In addition, the boxplots of gas specific gravity and nitrogen are extremely skewed to the right (or positively skewed) as it seen from the median which is located at the bottom of the IQR and the boxplots for this parameters don’t have even a lower tail (lower whisker), as well as these parameters have outlier points which are 1.5 times further form the IQR . It is also observed that the gas saturation has high negative skewness with outlier values that are 1.5 times further than the data points in the IQR. The overall variability in the reservoir parameters data may be attributed to the unique nature of the unconventional shale formation.

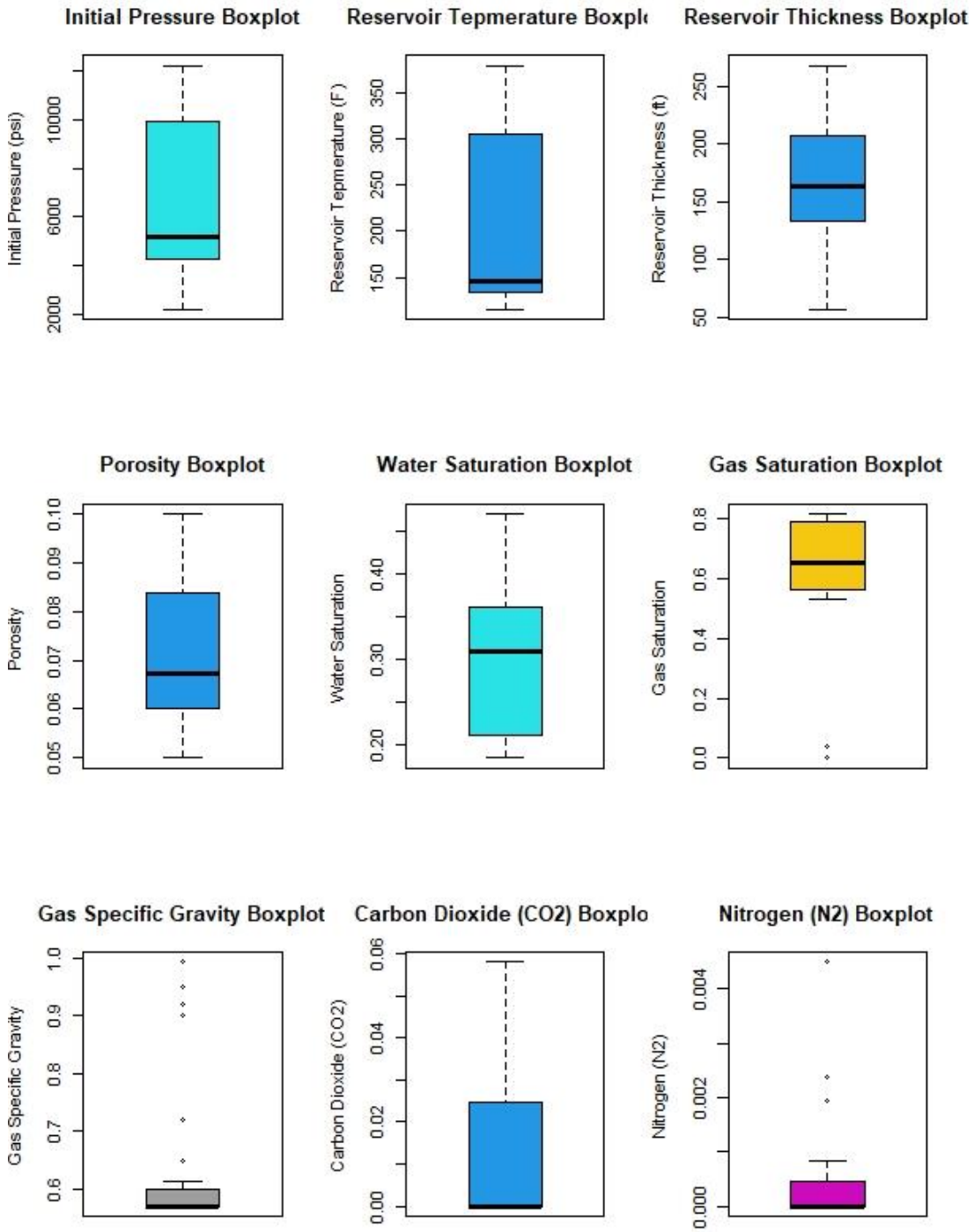


Figure 5. 1 : Reservoir parameters boxplots

5.2.2 Operational parameters boxplots

Figure 5.2 illustrates the operational parameters constructed graphically using boxplots. It is seen that the boxplots of operational parameters depict in general evenly distribution around the median except for number of clusters, clusters per stage and static well head temperature where we can see a number of data points that are out of the IQR and above the top whiskers. These values are 1.5 times further than the data distributed in the IQR. Some boxplots also show that the median is slightly near to the bottom of the box which imply that we have slightly skewness to the right. Hence the upper tail of the whiskers are longer than the lower one.

In comparison to the reservoir parameters, Figure 5.2 shows that the operational parameters exhibit less variation and fewer outliers. The operational parameters are relatively evenly distributed, with exception for the specific ones mentioned earlier. As mentioned previously, the high standard deviation of the reservoir data may be attributed to the unconventional nature of the shale reservoir.

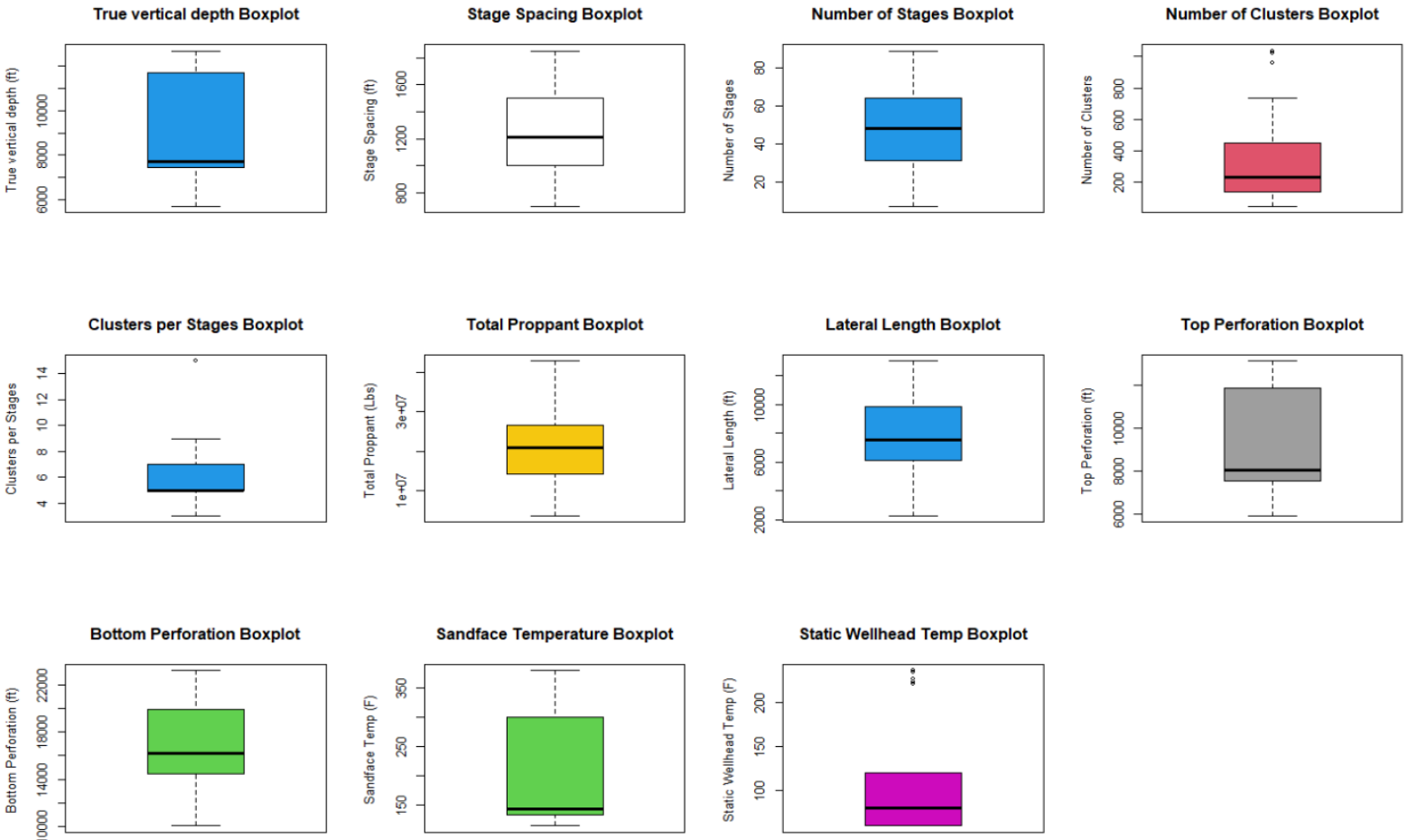


Figure 5. 2 : Operational parameters boxplots

5.2.3 Performance metric boxplot

This study focuses primary on the performance of shale gas wells. The response variable used to measure this performance is the cumulative gas production after one year, which is influenced by all the previously discussed input variables (reservoir and operational parameters).

It is shown in Figure 5.3, which illustrate the boxplot of cumulative gas production from different shale wells that the data show slightly skewed right distribution (positively skewed) since we see the elongated whisker to the upper side of the plot. This is an indicator that the majority of the data points of cumulative gas production are clustered toward the lower values shown by the IQR and a few of the data points that are extremely high (upper the IQR) trying to pull the upper whisker upward. It is also clear that the median is positioned approximately in the middle of the IQR (median = 3371 MMscf) which measures the spread of the data, indicating that most of the data, 50%, are between the first and third quartile range (between 1448 -6317 MMscf). It is also evident that there is no extreme outlier points as the all values lies in the IQR and within the range of the upper and lower whiskers.

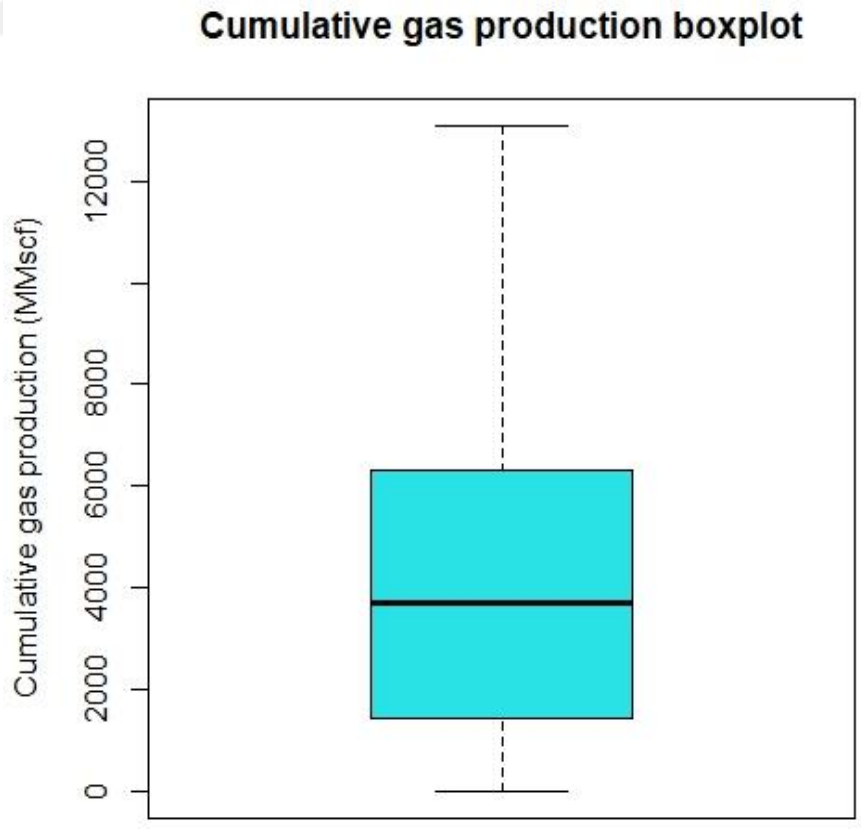


Figure 5. 3 : Cumulative gas production boxplot

5.2.4 Reservoir parameters histograms

The histograms of reservoir parameters came to visually confirm the earlier discussed descriptive statistics and boxplots. It is observed from Figure 5.5 that porosity, thickness, and water saturation are displaying nearly symmetry distribution which was also detected earlier by descriptive statistics and boxplots. The other parameters are displaying different data variability and skewness that was also observed clearly in the previous part using boxplots and descriptive statistics.

5.2.5 Performance metric histogram

Figure 5.4 illustrate the cumulative gas production after one year form different shale gas wells which is the performance metric of this study. It is observed from the histograms that gas production is distributed across different production levels. Where most of the wells are producing at low and moderate levels between 22 MMscf and 8000 MMscf. In the other hand, few exceptional wells are producing at higher rate which make the histograms bins spread toward higher values, indicating slightly positively skewed to the right.

Finally, the histograms are useful method to group the data of our different wells into group of intervals in order to estimate the wells with low, moderate, and high production rate and this will help us making accurate prediction.

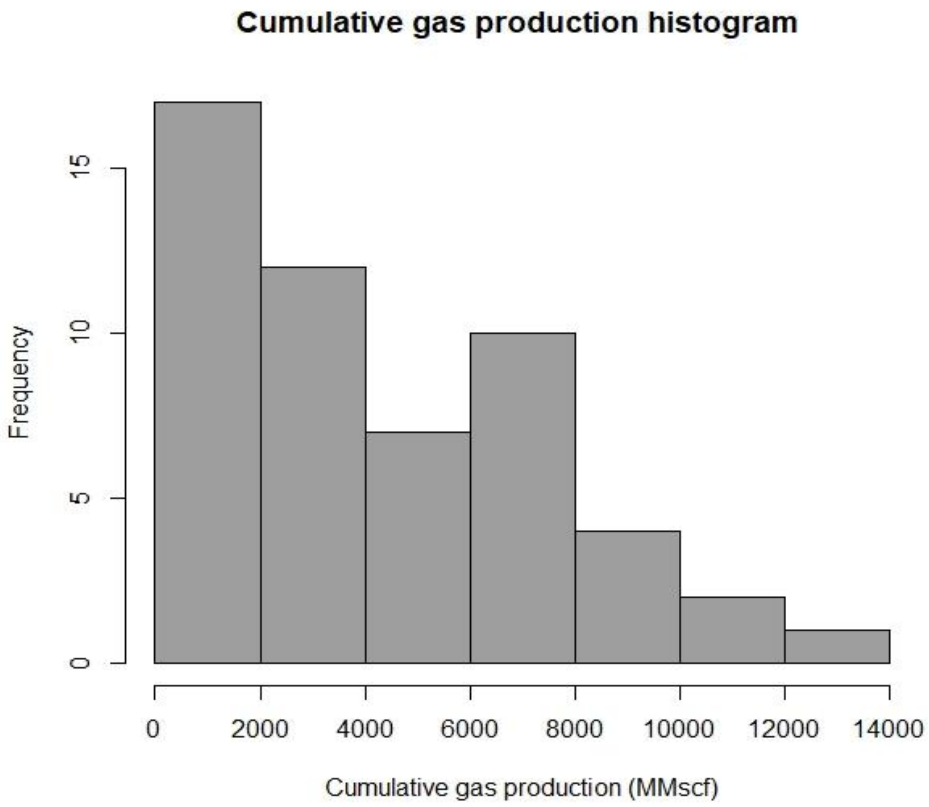


Figure 5. 4 : Cumulative gas production histogram

5.2.6 Operational parameters histograms

Figure 5.6 illustrates the histograms for operational parameters with their frequency intervals that are in general depicting a fairly symmetric distribution as in lateral length and stage spacing histograms. The degree of skewness for operational parameters is very low to be considered nearly symmetric except for some of the parameters such as static wellhead temperature, clusters, and clusters per stage where we observe from their histograms a high skew to the right due to the existence of extreme values (outliers).

When examining reservoir parameters against operational parameters through a histogram-based comparison, it becomes apparent that reservoir parameters exhibit greater variability and higher skewness in contrast to operational parameters.

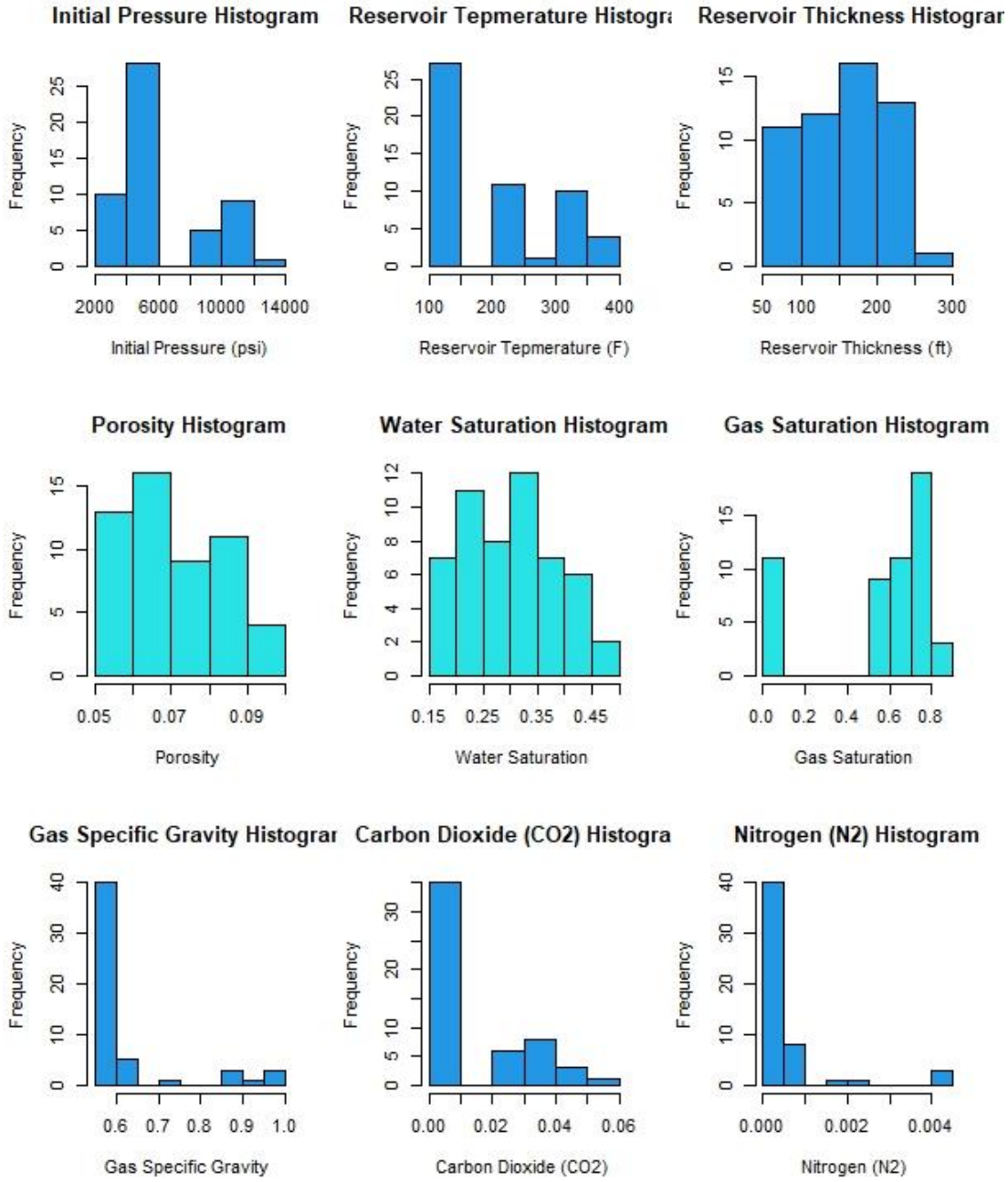


Figure 5. 5 : Histograms of reservoir parameters.

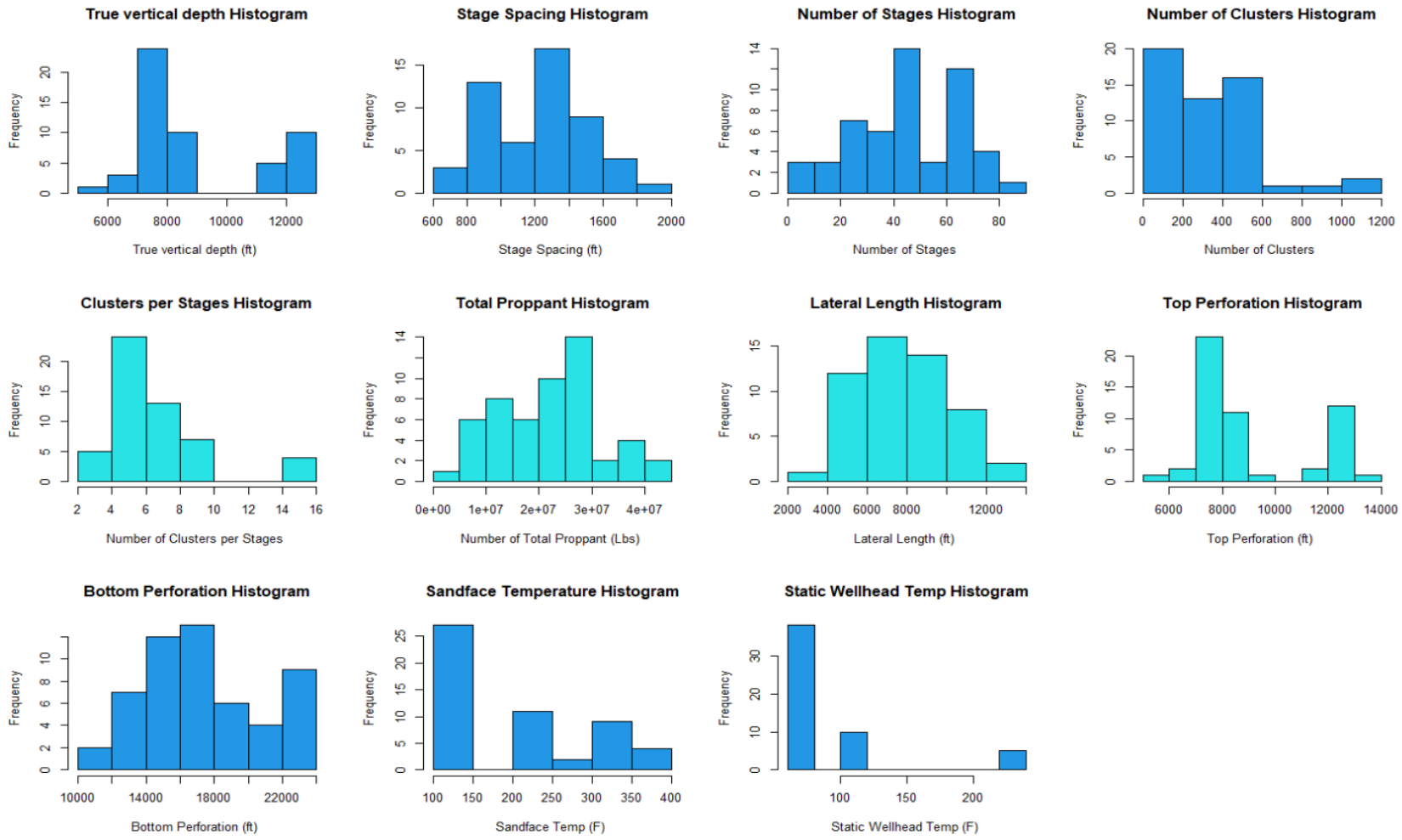


Figure 5. 6 : Operational parameters histograms

5.3 Bivariate Analysis

In this section, the bivariate data analysis was employed to examine the relationship between the performance metric which is the cumulative gas production after one year and each input reservoir and operational parameters. In order to do this, scatterplots is employed as method to examine each two variables relationship. Furthermore, the Pearson and Spearmen correlation coefficients were employed in this study to quantify the correlation between the performance metric and each input parameter.

5.3.1 Reservoir parameters scatterplots

Figure 5.7 illustrates the scatterplots for each reservoir parameter (predictor) and the cumulative gas production after one year (response variable) to examine and describe their relationship. The scatter plots of porosity and gas saturation are depicting the data points clustering around the trend line from the bottom left to the upper right which indicates a strong positive linear relationship between cumulative gas production and the gas saturation and porosity. In contrast, a modest and week relationship is presence between cumulative production and other reservoir parameters. For instance, cumulative gas production and the concentration of nitrogen depicting a weak negative linear relationship, since the data point are more spread out and are not clustered around the trend line. In addition a moderate positive linear relationship is presence in the reservoir thickness scatter plot, and a moderate negative liner relationship is observed between water saturation and cumulative gas production.

Finally, the skewness of reservoir parameter can be visualized by scatter plots as well. For instance, it is observed from Figure 5.7 that porosity and water saturation with cumulative gas production depicting nearly symmetric distribution which confirm the descriptive statistics discussed earlier. In the other hand, it is observed that the nitrogen and gas saturation have skewness to the right and left respectively. The Pearson and Spearman correlation coefficients are also shown on the scatterplots as P_CC and S_CC for Pearson and Spearman respectively. These coefficients will be explained later on this section.

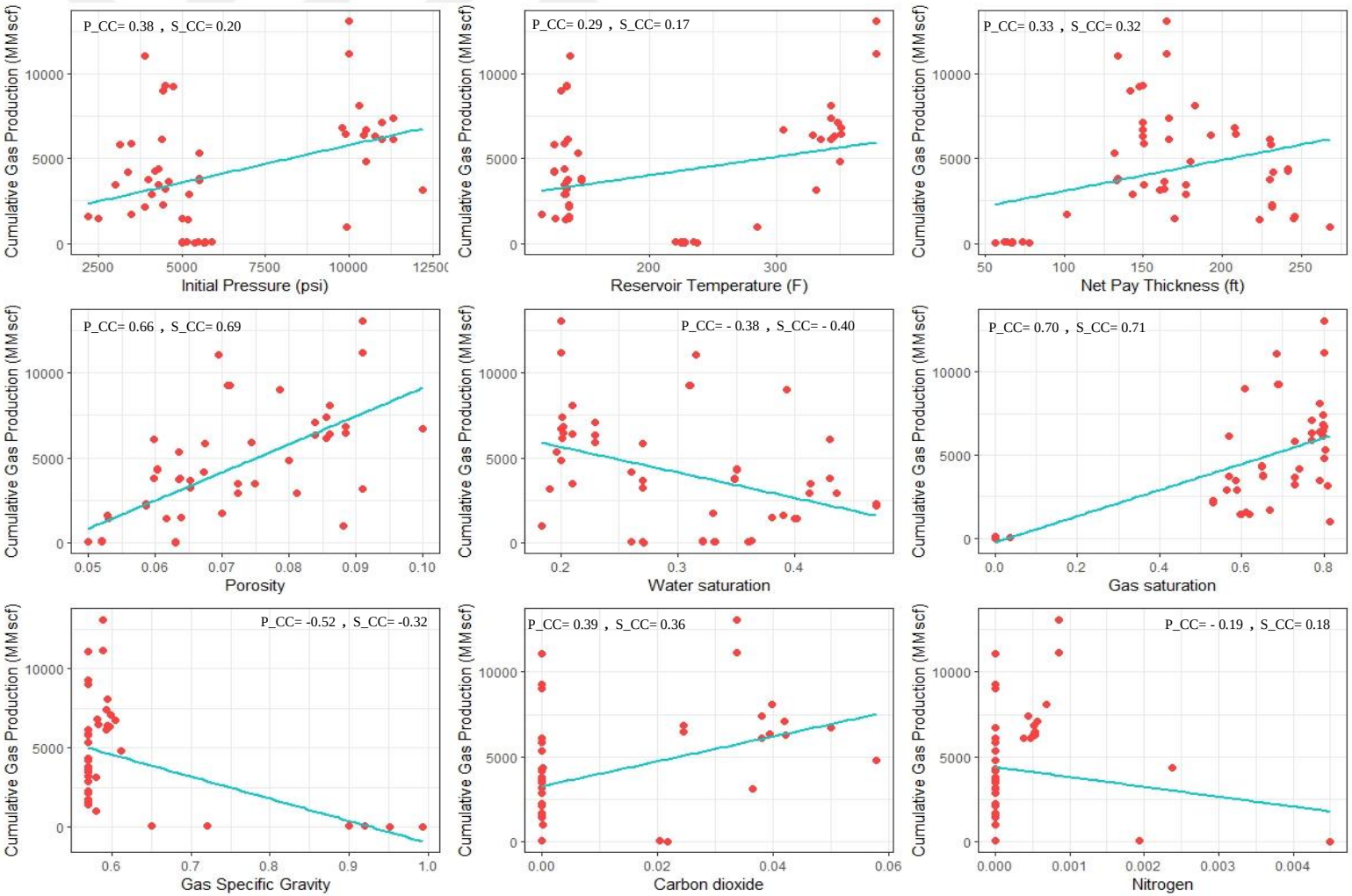


Figure 5. 7 : Reservoir parameters scatterplots

5.3.2 Operational parameters scatterplots

Figure 5.8 illustrates the scatterplots for each operational parameter (predictor) and the cumulative gas production after one year (response variable) to examine and describe their relationship. It is observed from the scatterplots in Figure 5.8 that both true vertical depth, number of stages, lateral length, top perforation, bottom perforation, sand face temperature, and static wellhead temperature are displaying a positive linear relationship with cumulative gas production. Among the above ones, there are positively strong linear relationship between the cumulative gas production with lateral length, bottom perforation, and the number of stages, since the points are clustered around the trend line. The linear relationship is positive because the trend of the data points are directed from the lower left side of the plot to the upper right. Furthermore, a strong negative linear relationship is exist between cumulative gas production and the number of clusters per stage.

A modest linear relationship is also presence between the cumulative gas production and the number of proppant, sandface temperature, and the number of clusters. This relationship can be figured out from the data points distribution around the trend line, where we can observe a greater spread out of data points. It is also observed from Figure 5.8, that the scatterplot of stage spacing displays a nonmonotonic relationship with cumulative gas production. The clusters per stage scatterplot display a negative correlation with the cumulative gas production. However this might not be the case for every shale well, since the clusters per stage is performed to increase the productivity not to decrease it as it observed in this case. In our case there are two possible reasons behind the negative effect of clusters on productivity: the different locations of our examined shale wells since every location has different perforation strategy, and the optimal number of required clusters per stage might not be considered carefully and this may lead to a negative impact on the productivity due to the interference between propagating hydraulic fractures.

Finally, the skewness of operational parameters can be visualized by scatter plots as well, and this confirm the results from descriptive statistics as explained earlier with reservoir parameters.

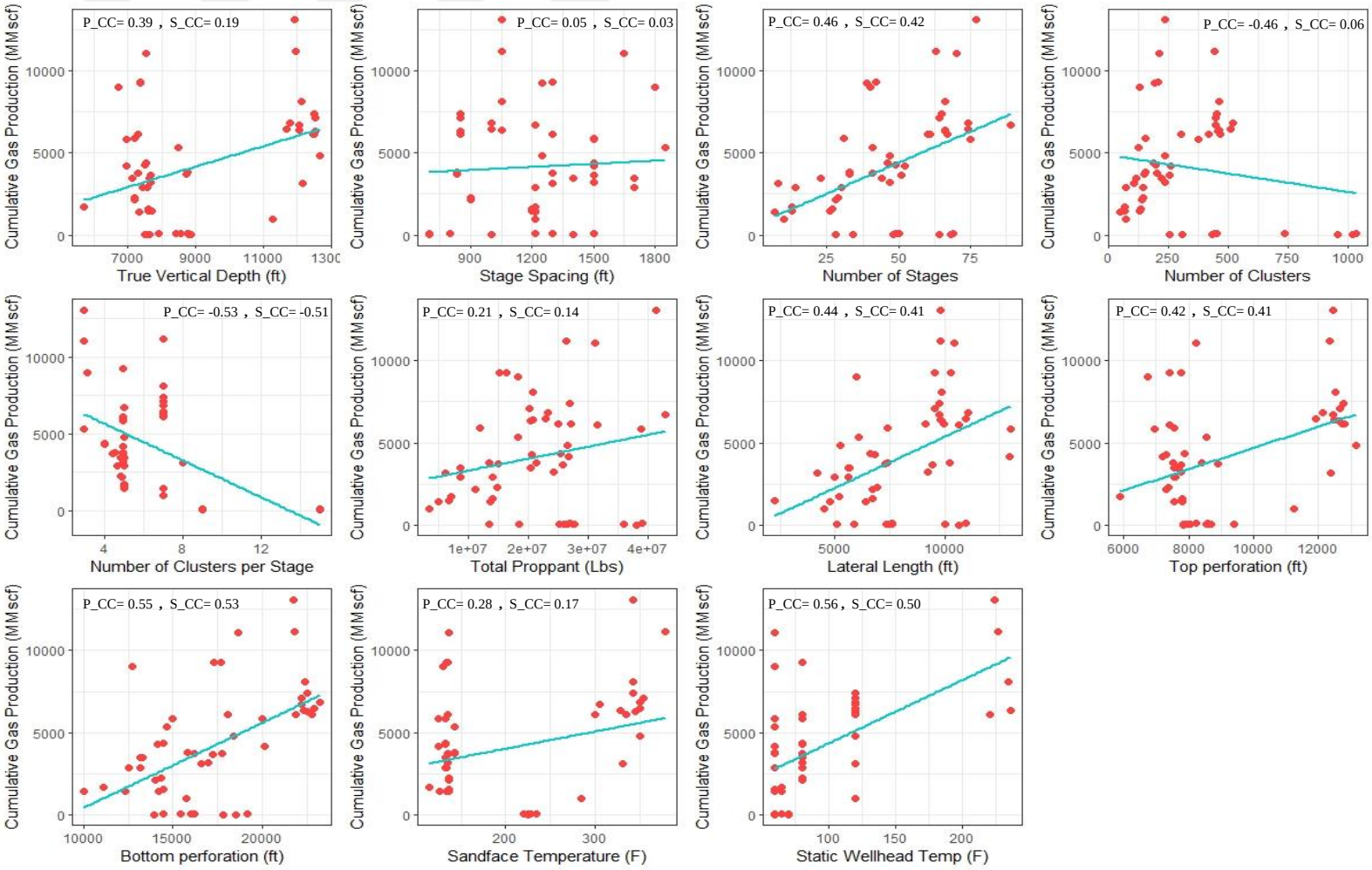


Figure 5. 8 : Operational parameters scatterplots

5.3.3 Correlation coefficients test

In this study, two types of correlation coefficient (CC) were used to test and quantify the relationship between cumulative gas production and reservoir and operational parameters. These correlation coefficients are Pearson and Spearman correlation coefficients. As discussed earlier in the methodology chapter, that Pearson CC may be influenced by outlier and clustered data points, while Spearman correlation coefficient is better for analyzing nonlinear relationships and is more robust to outliers and clusters within the data. Thus, in this study both of them are considered and calculated.

In Table 5.3, the correlation between reservoir parameter and the performance metric which is the cumulative gas production is introduced. It is observed that there is a positively strong correlation between, gas saturation, porosity, and cumulative gas production with their corresponding values of 0.70 and 0.66 for Pearson CC and 0.69 and 0.71 for Spearman CC respectively.

Table 5. 3 : CC of reservoir properties with the cumulative gas production

Reservoir properties	Pearson CC	Spearman CC
Initial Pressure (Pi)	0.38	0.20
Reservoir Temperature (Tr)	0.29	0.17
Net Pay Thickness (h)	0.33	0.32
Porosity (Ø)	0.66	0.69
Water Saturation (Sw)	-0.38	-0.40
Gas Saturation (Sg)	0.70	0.71
Gas Specific Gravity (Yg)	-0.52	-0.32
CO_2	0.39	0.36
N_2	-0.19	0.18

Furthermore, a moderate positive correlation between initial pressure, net pay thickness, reservoir temperature and the cumulative gas production is observed. In contrast, a moderate negative correlation is observed for both water saturation and concentration of nitrogen. A slightly strong negative correlation is also observed between gas specific gravity and cumulative gas production with Pearson and Spearman correlation coefficients of -0.52 and -0.32 respectively. The large difference between the values of Pearson and Spearman CC for gas specific gravity is due to the few extreme outlier data points. This difference can be clearly observed in Nitrogen

and number of clusters parameters where we see a change from negative to positive value of Pearson and Spearman coefficients and this is due to the nonlinearity distribution of these parameters and the extreme outlier values. In this case, the Spearman CC is more robust and reliable. These outlier values are clearly illustrated in Figure 5.1 and Figure 5.7 using the boxplots and scatterplots respectively.

To discuss the Pearson and Spearman correlation between operational parameter and cumulative gas production Table 5.4 is illustrated. Slightly strong positive correlation between bottom perforation, lateral length, top perforation, and static well head temperature is observed. As well as a slightly strong negative correlation is observed between number of clusters per stage and cumulative gas production with corresponding value of 0.-53 and-0.51 for Pearson and Spearman CC respectively. The other operational parameters are depicting whether a moderate, modest or weak correlation with the performance metric. In addition, approximately no correlation is observed between stage spacing and cumulative gas production since the value of Pearson and Spearman for stage spacing is 0.05 and 0.03 respectively. This confirm the visual discussion of scatterplots earlier in this chapter. However, this does not indicate that stage spacing is not a significant parameter for gas production, since the relationship might be a quadratic or nonlinear type for this parameter. Therefore, the significance of both reservoir and operational parameters will be addressed in the upcoming sections of this chapter, particularly in the variable importance discussion.

Table 5. 4 : CC of operational properties with cumulative gas production

Operational properties	Pearson CC	Spearman CC
True Vertical Depth (TVD)	0.39	0.19
Stage Spacing	0.05	0.03
Number of stages	0.46	0.42
Number of clusters	-0.16	0.06
Clusters per Stage	-0.53	-0.51
Proppant	0.21	0.14
Lateral Length	0.44	0.41
Top Perforation	0.42	0.21
Bottom Perforation	0.55	0.53
Sandface Temperature (Tsf)	0.28	0.17
Static Wellhead Tem. (Twhs)	0.56	0.50

5.4 Multivariate Analysis

Multivariate data analysis is performed in this study as extension for what we discussed before in the bivariate analysis using scatterplots and correlation coefficient. The main difference here is instead of examining the association between each pair of variables individually, multivariate analysis using correlation matrix helps us to describe the correlation among each two variables at once. For example, Figure 5.9 shows the correlation matrix that depict all parameters in our dataset with their corresponding correlation measure from -1 to 1 to mark the strength of the correlation. The red circle indicates a negative while the blue one imply a positive correlation. The larger the circle the more strong the correlation is. It is observed from Figure 5.9 that there is strong and slightly strong positive correlation (marked with a blue circle) between the cumulative gas production (Gp) and porosity, gas saturation (Sg), bottom and top perforation, lateral length, stages and TVD. A strong negative correlation is also observed between the number of clusters per stages, water saturation, gas specific gravity, and cumulative gas production. A modest and moderate correlation is also observed in the correlation matrix below.

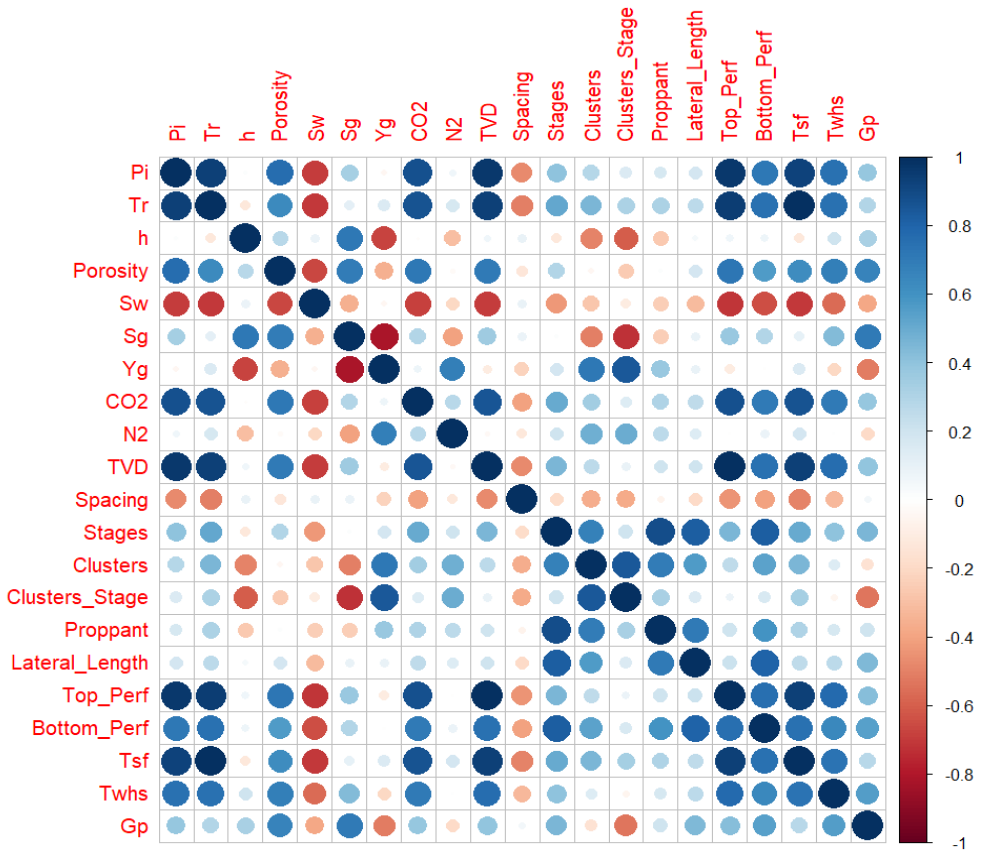


Figure 5. 9 : Correlation matrix

5.5 Machine Learning-based Predictive Modeling

After we gained a valuable insights into our dataset using exploratory data analysis (EDA) in the previous section, which helped us understanding patterns, trend and relationship among variables and between each parameter and cumulative gas production, we can move on to develop predictive model that is able forecast the performance of our shale gas wells (cumulative gas production after one year).

In this study, two primary methods were utilized for model development, namely multiple linear regression (MLR) and tree based approaches. Among the tree-based approaches were regression tree, bagging, random forest (RF), and extreme gradient boosting machine (XGBoost). The use of these predictive models is crucial for accurately forecasting the gas production based on the available dataset.

5.5.1 Multiple linear regression

Given the presence of more than one variable in our dataset, encompassing both reservoir and operational properties, this study employed multiple linear regression to construct a proxy model for predicting the performance of shale gas wells.

The dataset employed in this study consist of 53 shale gas wells with 20 different variables (predictors) distributed between reservoir and operational properties. The performance metric here is the cumulative gas production after one year (Gp) measured by million standard cubic feet (MMscf). Table 5.5 listed the reservoir and operational predictor variables that influence the performance metric (Gp).

In this study, initially the MLR is conducted for the full dataset (53 observation and 20 variables) in order to examine if there is at least one of the predictors listed in Table 5.5 able to make prediction on the response variable namely cumulative gas production Gp. First of all, we start conducting the multiple linear regression model to be able to extract the importance metrics that can be helpful in determining the relationship between predictors and response variable. After fitting multiple linear regression model the summary statistics in Table 5.6 is obtained. As it seen from Table 5.6, F-statistic is 14.43 which is far larger than 1, thus it indicates the presence of an association between cumulative gas production (response) and at least one of the predictors.

Table 5. 5 : Overall dataset characteristics

Parameter Description	Variable	Unit	Type
Cumulative Gas Production	Gp	MMscf	Response
Initial Pressure	Pi	psi	
Reservoir Temperature	Tr	deg F	
Net Pay Thickness	h	ft	
Porosity	Porosity	%	
Water Saturation	Sw	%	
Gas Saturation	Sg	%	
Gas Specific Gravity	Yg	%	
CO_2	CO2	%	
N_2	N2	%	
True Vertical Depth	TVD	ft	Predictor
Stage Spacing	Spacing	ft	
Stages	Stages	#	
Clusters	clusters	#	
Clusters per Stage	Clusters_Stage	#	
Proppant	Proppant	lbs	
Lateral Length	Lateral_Length	ft	
Top Perforation	Top_Perf	ft	
Bottom Perforation	Bottom_Perf	ft	
Sandface Temperature	Tsf	deg F	
Static Wellhead Temperature	Twhs	deg F	

Table 5. 6 : Summary of multiple linear regression model on the full dataset.

Quantity	Value
Multiple R^2	0.90
Adjusted R^2	0.8378
RSE	1343
F-statistic	14.43
p-value	8.156e-011

The residual standard error as well as is observed with a value of 1343, which indicate how the residuals are deviated from the regression model. This values is moderately small hence the model is fitting our dataset well. Furthermore, multiple R-Squared is observed to be 0.90 which help to determine the model goodness of fit. In this case, it is indicating that 90% of the variation in cumulative gas production can be explained by the reservoir and operational properties (predictors).

Finally, it is clearly observed from Table 5.6 that the p-value is 8.156e-011, which is highly significant. This indicates that at least one of the predictors, which is in this study the reservoir and operational parameters, is significantly correlated with cumulative gas production Gp. The adjusted R-squared value is 0.8378, which represent the R-squared value after adjusting the number of contributed predictors to the model. This will be discussed later in this section.

The subsequent step in this analysis involves identifying the subset of predictors that significantly influence the shale gas wells performance in order to fit the multiple regression model with only these predictors.

In order to do identify the main subset of predictors that will be included in the model, best subset section algorithm is performed as shown in (Appendix A.1). The results in (Appendix A.1) illustrate the predictor parameters with asterisk signs. These asterisks represent the optimal parameters that best explain the model. The results indicate (ordered by asterisk sign count) that lateral length, bottom perforation, number of clusters per stage, gas saturation, top perforation, porosity, and thickness are the parameters that have the most significant contribution to the model. Furthermore, for model assessment, R-squared and RSS metrics can be used to determine the number of parameters to include in the model, relying on the training. For example, it is observed from (Appendix A.1) and Figure 5.10 that R-squared value exhibits an

increase from approximately 49% with just one parameter in the model to nearly 90% with the inclusion of all parameters.

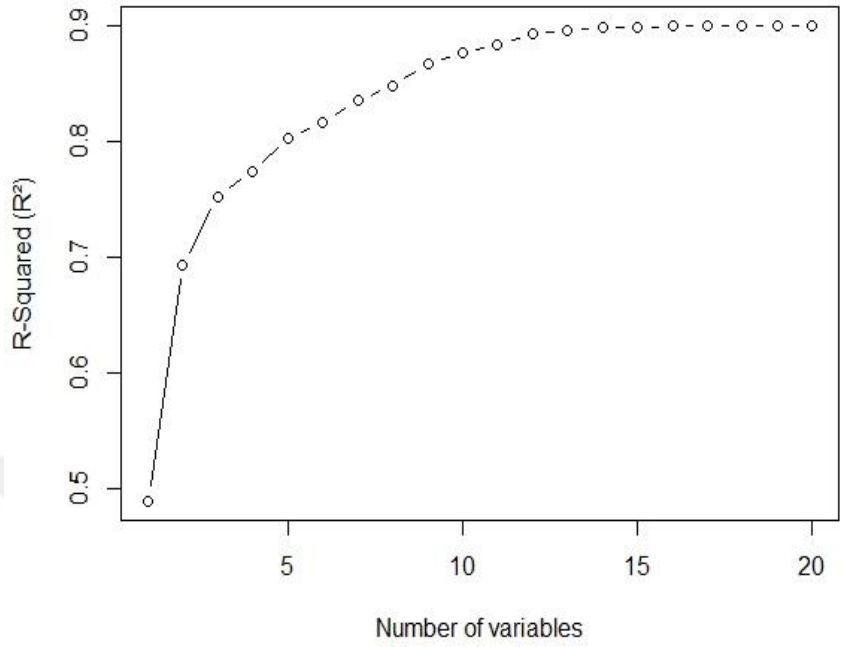


Figure 5. 10 : R-squared plot

Moreover, Figure 5.11 illustrate that while increasing the contributed number of predictors to the model, the RSS decreases. Therefore, both R-squared and RSS are not appropriate metrics to evaluate our regression model and choose the best parameters that influence the response variable (Gp).

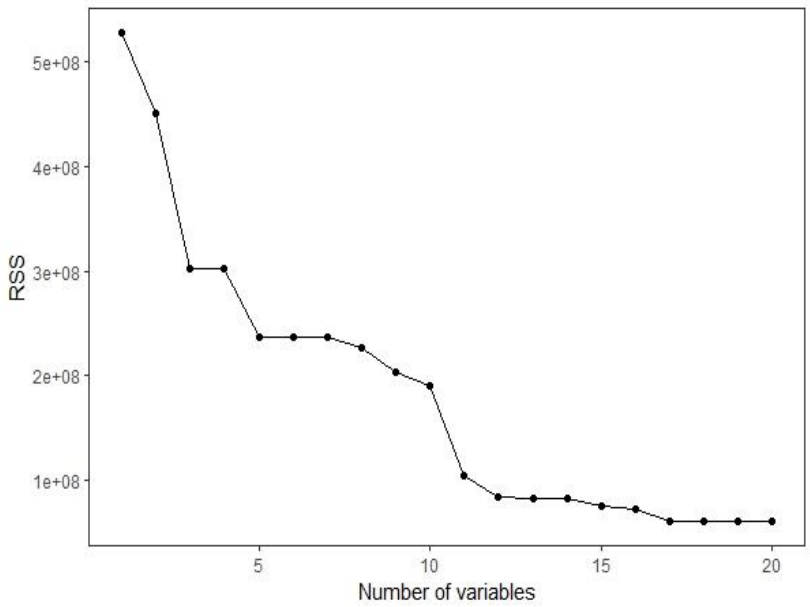


Figure 5. 11 : RSS Plot

While R-squared and RSS metrics are used to suggest a model based on the training error, some other metrics such as C_p , adjusted R-squared and BIC may look more powerful since they choose the model based on the test error which is more important for selecting the appropriate model among multiple models. The preferred model is the one with the lowest Cp or BIC values, while models with the highest adjusted R-squared values are accepted. For instance, Figure 5.12 illustrate the number of predictors that should be optimal while building our regression model. It is observed that the highest value of adjusted R^2 corresponds to 13 parameters. This means when building our regression model rather than including 20 parameters it is preferred to include just 13 parameters in order to get the lowest test error along with avoiding overfitting.

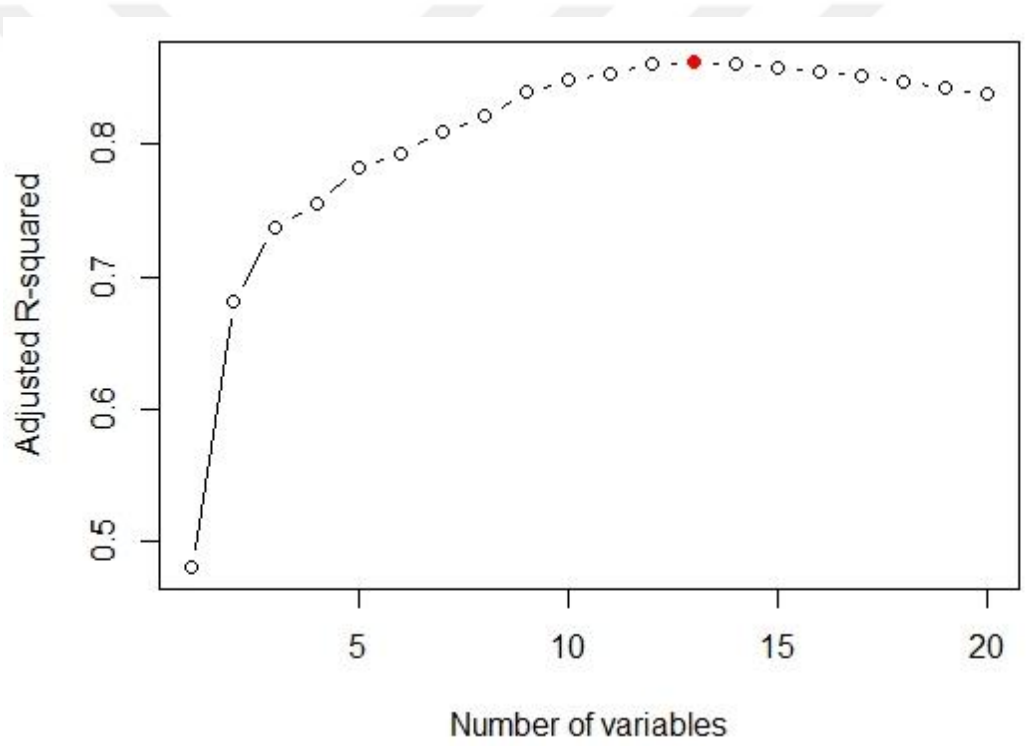


Figure 5. 12 : Adjusted R-squared plot

In the other hand, Figure 5.13 illustrate the Cp values with their corresponding number of variables. It is observed that the optimal model in this case is obtained when the lowest value of Cp corresponds to 12 variables.

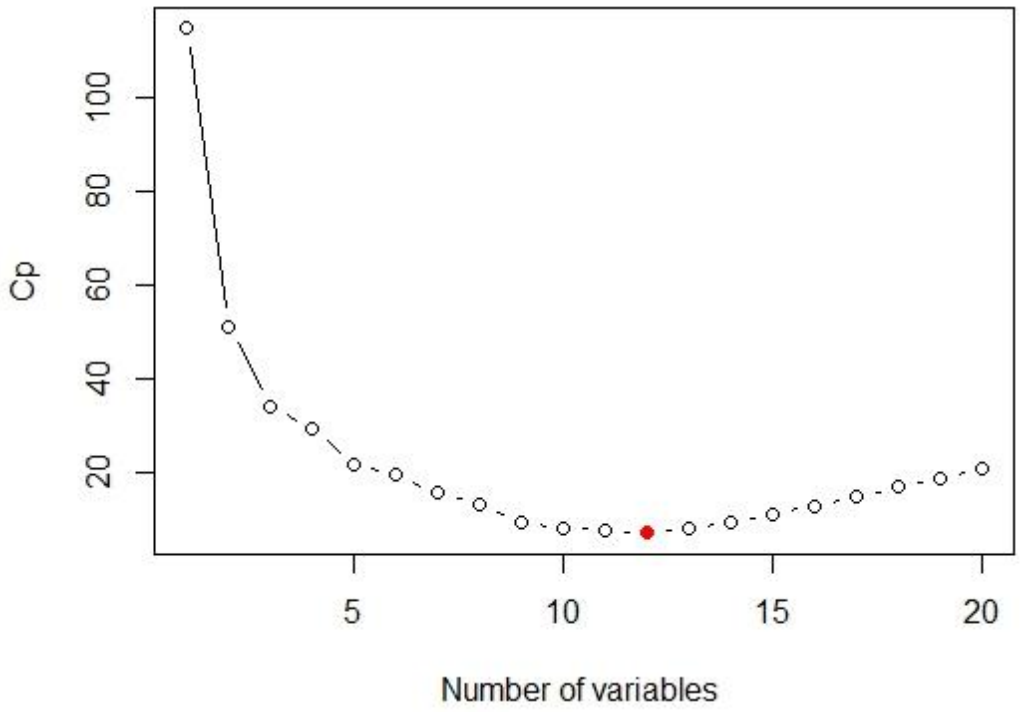


Figure 5. 13 : Cp plot

Finally, BIC metric vs number of parameters is conducted in Figure 5.14 and it is observed that the lowest value of BIC is corresponding to an optimal model with only 10 variables.

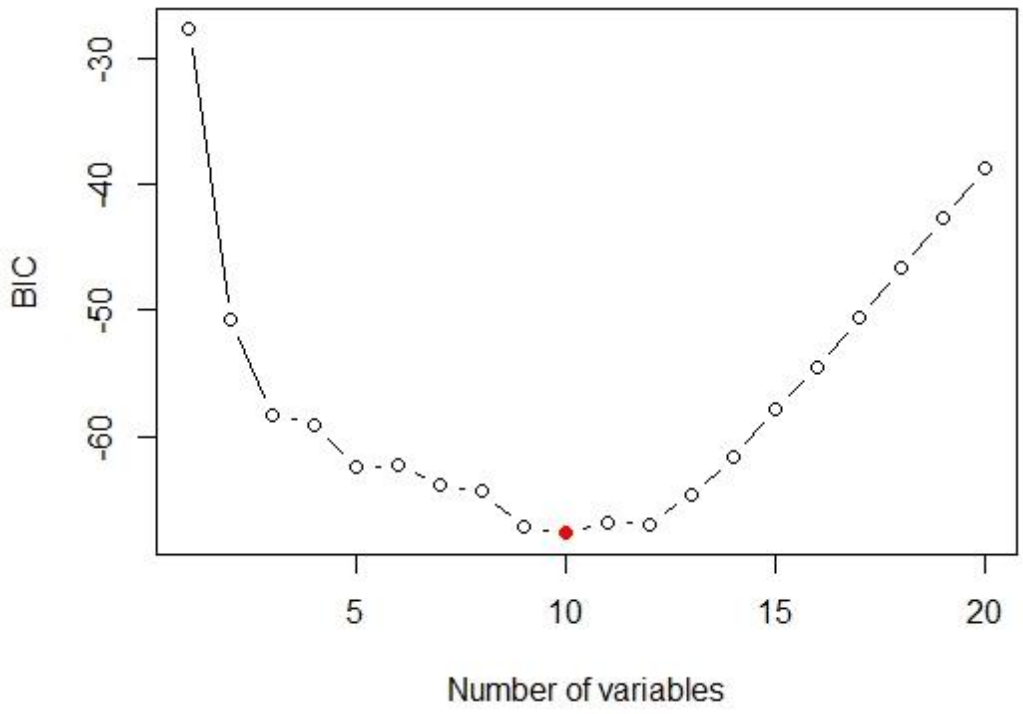


Figure 5. 14 : BIC plot

As seen from the graphical discussion of Cp, BIC, and Adjusted R^2, the metric with the lowest number of variables was BIC. Hence, BIC statistic based optimal multiple linear regression model is conducted after estimating the suggested 10 optimal variables that make up our model. Equation 5.1 is illustrating the 10 optimal variables corresponding model.

$$G_P = \beta_0 + \beta_1 S_g + \beta_2 h + \beta_3 Lateral_{length} + \beta_4 Top_{Perf} + \beta_5 Stages + \beta_6 Bottom_{Perf} + \beta_7 TVD + \beta_8 Spacing + \beta_9 Proppant + \beta_{10} S_w \quad (5.1)$$

Where the regression coefficients are as follows;

$\beta_0 = -10543.896$ $\beta_4 = 15.397$ $\beta_8 = -2.245$

$\beta_1 = 9020.454$ $\beta_5 = 110.148$ $\beta_9 = -0.00017$

$\beta_2 = -21.457$ $\beta_6 = -13.365$ $\beta_{10} = 13924.480$

$\beta_3 = 13.706$ $\beta_7 = -1.886$

It's crucial to note that this multiple linear regression model cannot be generalized to all unconventional shale reservoirs unless a similarity of the input parameters is presence.

There are more methods to select the optimal model among multiple models such as validation set, LOOCV, and k-fold cross-validation. These methods main perspective is splitting the data into training and validation set which allow us to select the model based on the test error rather than selection the model by just using the training data as in Cp, BIC and adjusted R-squared. For instance Figure 5.15 shows the validation set method plot for selecting the optimal parameters that make up our model and it shows that the validation test error is in its minimum when only 16 variables are included in our regression model. The advantage of this method is before implementing the selection of variables, it allow us splitting our dataset to training and test sets, thus the resulting errors are based on the validation set (test set) using mean square error (MSE). Hence, it suggest using 16 variables in our model since it corresponds the lowest test MSE.

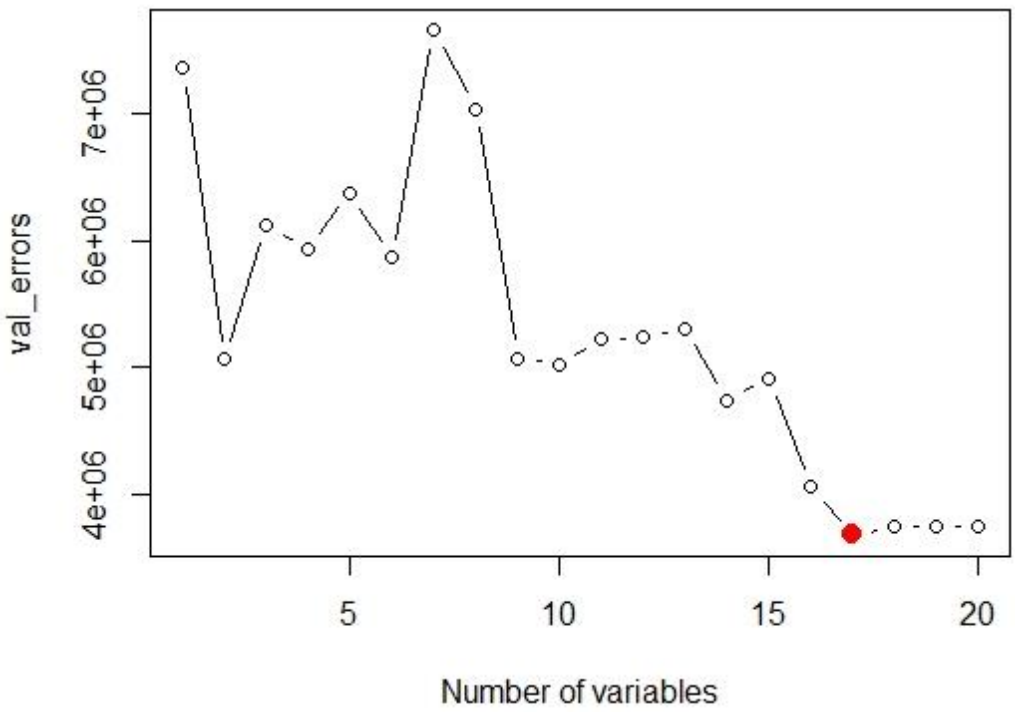


Figure 5. 15 : Validation set approach plot

Furthermore, Figure 5.16 illustrate the leave-one-out cross-validation method which allow us to use repeatedly only one observation for the validation test while the other all observation is used for training. As a result, LOOCV suggest to use 10 variables in our model since it corresponds to the lowest test MSE.

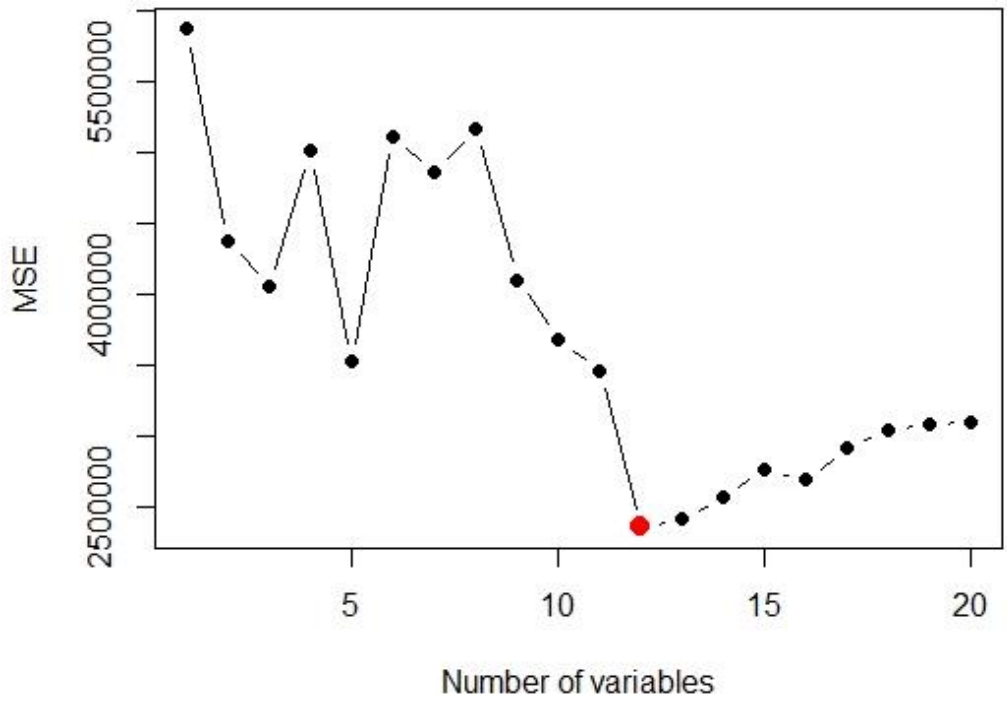


Figure 5. 16 : Leave-one-out cross-validation plot

Finally, we used a well-known so-called k-fold cross-validation method as refinement for the previous methods. In this method we used 10 folds, these folds are splitted into training and test folds. Where one fold was repeatedly used for validation and the remaining 9 folds were utilized for training. The results of k-fold cross-validation is illustrated in Figure 5.17. The method suggest to use 13 variables in our model based on the mean cross-validation errors (mean-cv-errors).

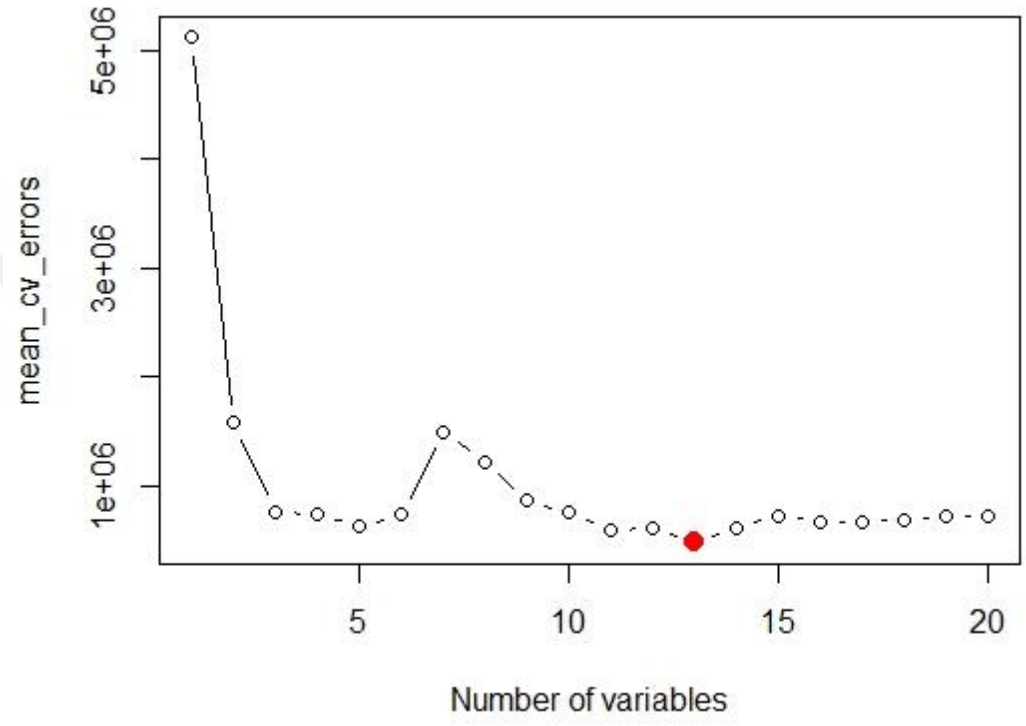


Figure 5. 17 : k-fold cross-validation plot

Lastly, we performed best subset selection algorithm (Appendix A.2) for our full dataset to obtain the optimal 13 variables that will be included in our multiple linear regression proxy model. As seen in (Appendix A.2), the variables with 'TRUE' are included whereas the variables with 'FALSE' are removed from the model.

Finally, equation 5.2 is obtained to represent our multiple linear regression model (Appendix A.3) for the full data set with 13 variables that was determined by k-fold cross-validation.

$$G_P = \beta_0 + \beta_1 S_g + \beta_2 h + \beta_3 Lateral_{length} + \beta_4 Top_{Perf} + \beta_5 Stages$$
$$+\beta_6 Bottom_{Perf} + \beta_7 TVD + \beta_8 Spacing + \beta_9 Proppant \quad (5.2)$$
$$+\beta_{10} S_w + \beta_{11} P_i + \beta_{12} Clusters + \beta_{13} Y_g$$

Where the regression coefficients are as follows;

$\beta_0 = -12249.91$ $\beta_4 = 13.96$ $\beta_8 = 2.17$

$\beta_1 = 8440.14$ $\beta_5 = 132.61$ $\beta_9 = -0.00017$

$\beta_2 = -20.57$ $\beta_6 = -12.22$ $\beta_{10} = 14642.16$

$\beta_3 = 12.64$ $\beta_7 = -2.01$ $\beta_{11} = 0.3932$

$\beta_{12} = -4.63$ $\beta_{13} = 4447.26$

After conducting the multiple linear regression model with 13 variables, it is importance to compare the results with the ones obtained above with all 20 variables. Table 5.7 illustrate the summary statistics of MLR model with 13 variables. It is observed that multiple R-squared slightly decreases while adjusted R-squared depicts an improvement from 83% to 86% which confirm our earlier discussion that multiple R^2 is not appropriate measure to choose a model among multiple models since it is increasing with the increase of variables. In the other hand, we said that adjusted R-squared is taking the number of variables into account when assessing a model and here it confirms that as it depicts an increase when removing the variables that do not contribute well to our model. Furthermore, the residual standard error decreased when only 13 variables are used. As well as the F-statistic is now far larger than the one above which signify a better association between variables and our response variable (cumulative gas production). Finally, the p-value with 13 variables is more significant and less than the p-value with 22 variables which indicates a better correlation between the cumulative gas production and reservoir and operational parameters.

Table 5. 7 : Multiple linear regression with 13 variables

Quantity	Value
Multiple R^2	0.896
Adjusted R^2	0.862
RSE	1239
F-statistic	25.97
p-value	3.526e-015

Consequently, these 13 variables will be used to develop a single predictive model using MLR for the full dataset after splitting the data into 80% and 20% for training and test respectively. As well as the model goodness of fit will be evaluated using RMSE, MSE, and AAE.

Figure 5.18 illustrates the results of multiple linear regression model predicted versus actual values of cumulative gas production after one year (MMscf). It is observed that all points are not fallen in the trend line (model fit line), but in general a good fit is observed between the predicted and observed values of cumulative gas production. Furthermore, the test R^2 value is 88%, which confirms that our model explain a significant amount of the variability in the cumulative gas production.

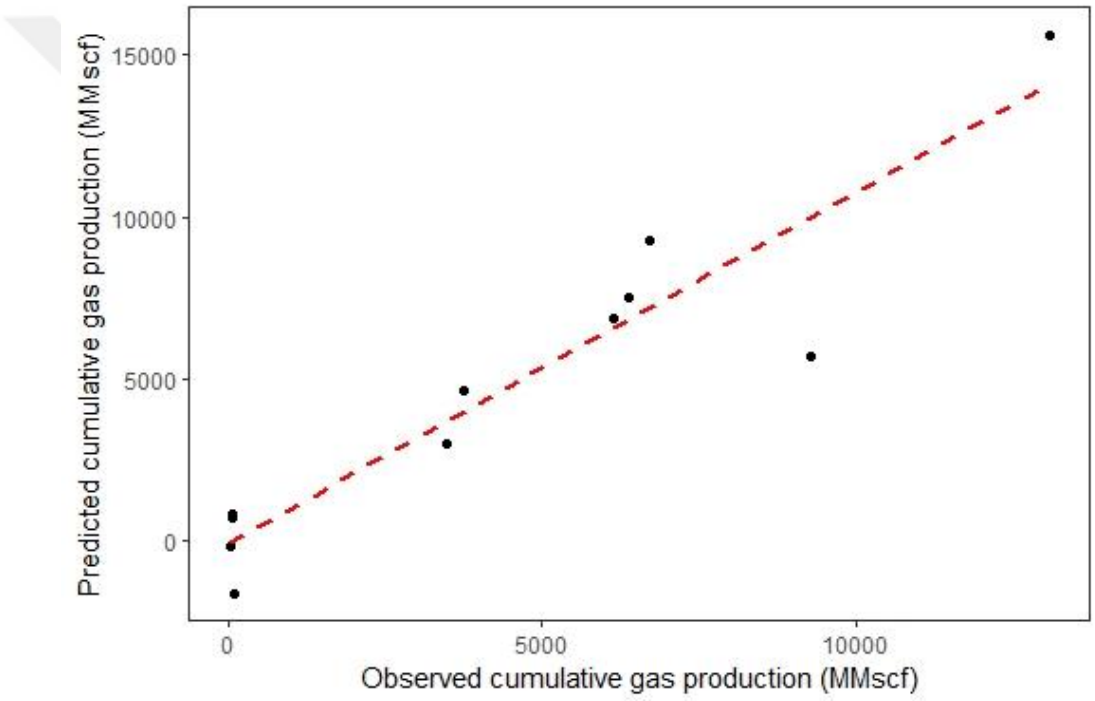


Figure 5. 18 : Observed vs predicted cumulative gas production for MLR model.

Finally, to assess the goodness of fit of the MLR model, we obtained the following metrics: RMSE, MSE, and AAE.

$RMSE = 1710.53\ MMscf$

$MSE = 2925936\ MMscf^2$

$AAE = 1376\ MMscf$

$Test\ R^2 = 88\%$

Later on, these metrics will be compared with the outcomes of tree-based methods to assess and choose the best machine learning-based predictive model that show a better performance in forecasting the cumulative gas production.

The last step of multiple linear regression is to check if the conducted MLR model validates the regression assumptions. To do this, diagnostics plots of our MLR model is illustrated in Figure 5.19.

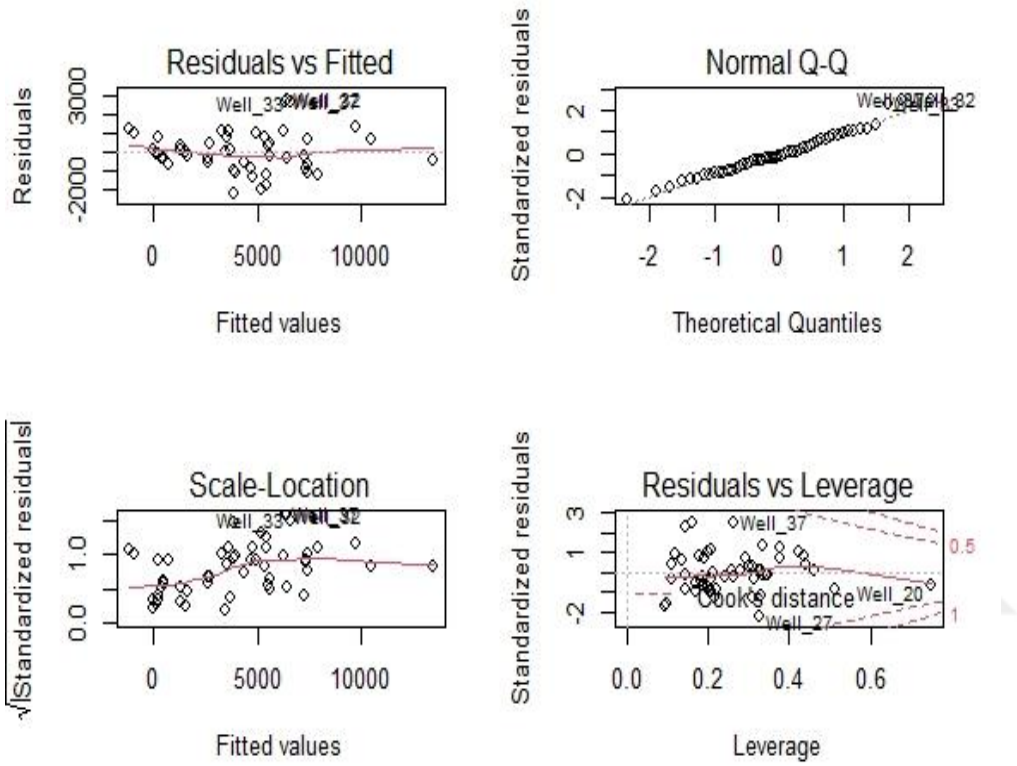


Figure 5. 19 : MLR diagnostic plots.

From Figure 5.19, it is evident that the residuals vs fitted values plot depict approximately horizontal red line which indicate the linearity of the data. For Normal Q-Q plot, it is observed that the residuals followed the dashed line which indicate a normality distribution of the residuals.

In the Scale-Location plot, the residual points are spreading out around the red line which assume homogeneity of variance. Finally, the Residuals vs Leverage plot indicate no outliers since the data point are not exceeding the value of 3 standardized residuals and it assume influential points marked by ‘well-20’ which indicate that this observation may alter the results if we removed it. As a result, we can say that our regression model is satisfying the regression assumptions.

5.5.2 Tree-based Methods

Tree-based algorithms is preferred over MLR since it capture the non-linearity of the data. As shown before in MLR, we had a nearly linearity in our data when performing 13 variables-model, but when we include all 20 variables in the model the linearity assumption is somehow violated. Therefore, tree-based methods is a solution for this since it has the capability of capturing the nonlinearity as well.

In this section we developed machine learning-based predictive models using regression tree, bagging, RF, and XGBoost in order to forecast the performance of our shale gas wells. As we did in multiple linear regression, we started splitting our available dataset for train and test by 80% and 20% respectively. The model was trained using the 80% training data, while the 20% test data is utilized to validate the model, hence being able to evaluate the model's predictive accuracy.

After performing the regression tree for our full dataset Figure 5.20 is obtained. It observed that 5 main reservoir and operational parameters were used to build the tree. This 5 parameters sections the tree until no more splits is observed. Hence, 7 terminal leafs (nodes) is reaches which make the final predictions of the regression tree.

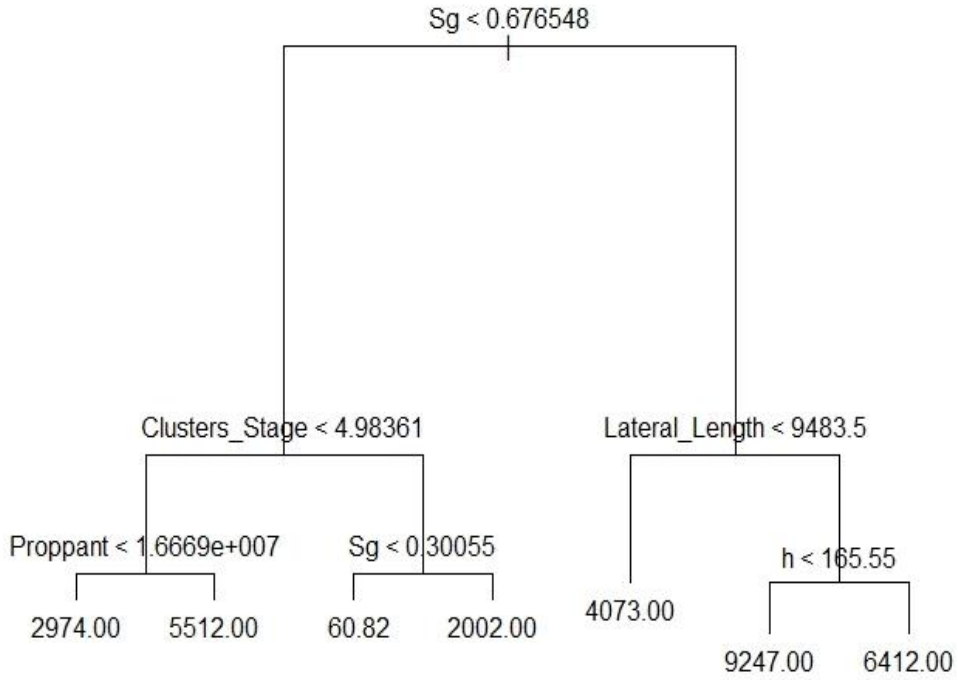


Figure 5. 20 : Regression tree for the full dataset.

Furthermore, in order to build a predictive model, the regression tree should be performed in training data, then develop the predictive model using the test data that remained untouched during the training. Therefore, Figure 5.21 is illustrating the regression tree using only the training data and it depicts that the tree also used 5 out of 20 parameters distributed between reservoir and operational parameters and 6 terminal nodes which is the mean predictions of the model.

As it observed from the tree, the regression model is easy to interpret, since the main split (root node) which is here the gas saturation parameter is the main parameter that influence the cumulative gas production after one year. In general, the parameters around the top of the tree are the influential parameters. In Figure 5.21, gas saturation, lateral length, and the number of clusters per stage are the primary parameters that affect the performance of gas production according to our available dataset.

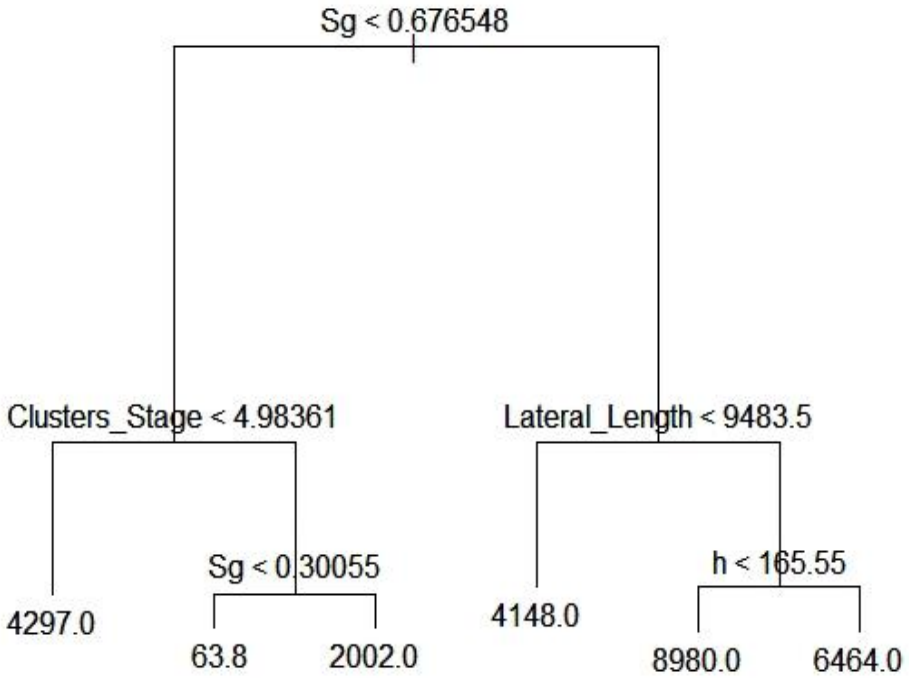


Figure 5. 21 : Regression tree for training data.

As seen from Figure 5.21, the regression tree looks simple with 7 terminal mean predictions and there is no complexity in the tree. However, to be sure, we performed cross-validation to determine the tree size. Figure 5.22 shows that the tree size could be 5 or 6 terminal nodes however the optimal one can be obtained with 4 terminal nodes.

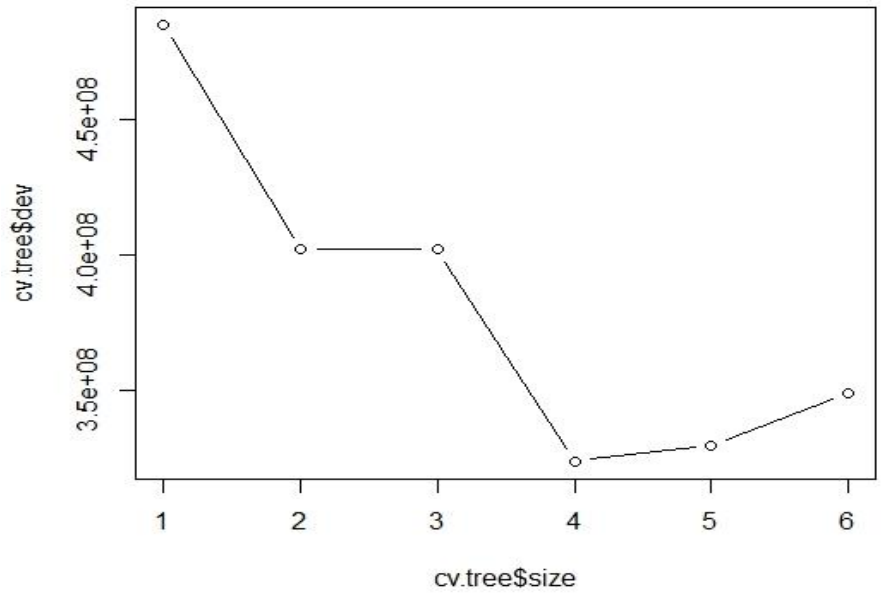


Figure 5. 22 : Cross-validation for pruning the regression tree.

Therefore, the obtained pruned tree is illustrated in Figure 5.23 and it can simply be interpreted. The pruned tree has a portion with a low mean prediction value and a region with a high mean forecast value of cumulative gas production. For example, a high gas production scenario can be obtained when gas saturation is larger than 0.676548 and the lateral length is larger than 9483.5 ft. which yield a mean predicted cumulative production value of 7608 MMscf (Sg > 0.676548 and Lateral_length > 9483.5 ft then the mean Gp=7608 MMscf).

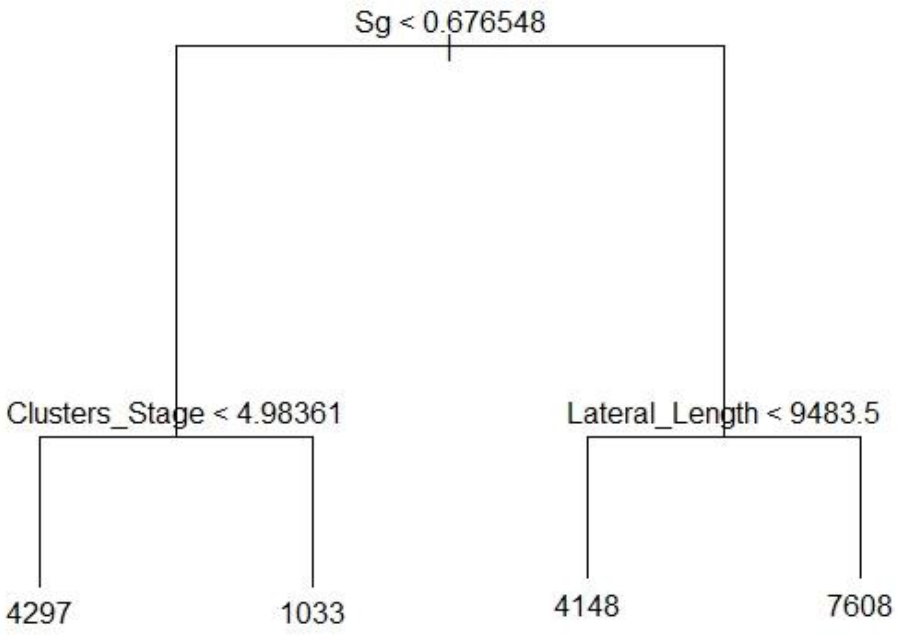


Figure 5. 23 : Pruned regression tree.

In contrast, a low cumulative gas production scenario could be obtained when gas saturation is less than 0.676548 and the number of clusters per stage is larger than 4.98361. This yield a mean cumulative production of 1033 MMscf.

Finally, as a conclusion from regression tree, the gas saturation, number of clusters per stage, and lateral length are the most significant parameters that determine the shale wells performance in unconventional reservoirs. In this case, the gas saturation was the top parameter that control the gas production since the presence of high gas saturation values in our wells data, together with the fact that the oil saturation values was almost nil for all of the wells in our data. Hence, the pore spaces of our reservoir formations were approximately occupied by the gas rather than other fluids. Additionally, a cross-plot of the predicted versus observed values of cumulative gas production from regression tree is used to quantify the model prediction error. This results in the following prediction error:

$RMSE = 1947.61\ MMscf$

$MSE = 3793177\ MMscf^2$

$R^2 = 78\%$

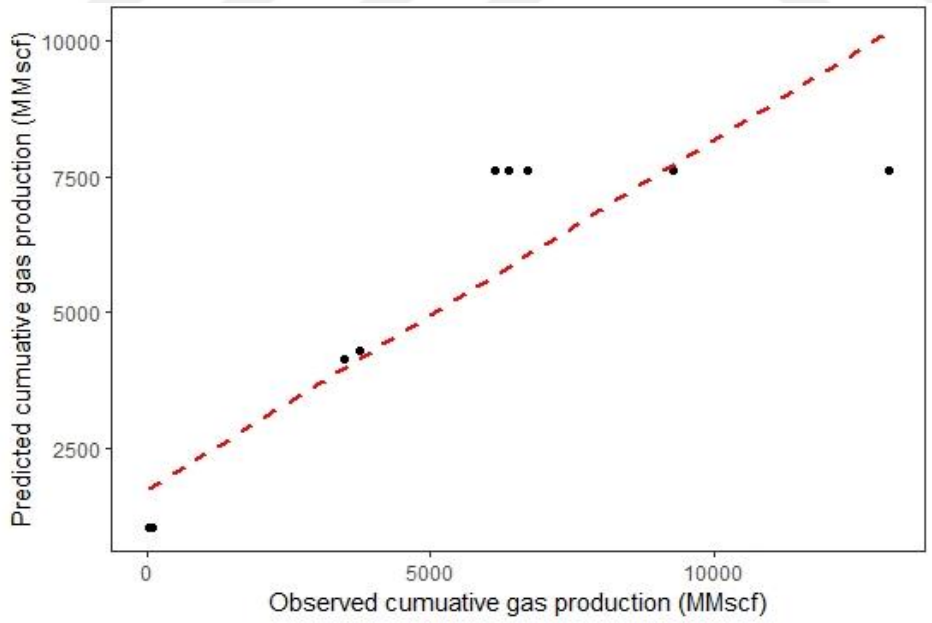


Figure 5. 24 : Observed vs predicted cumulative gas production for regression tree.

More effective machine learning methods, such bagging, RF, and XGBoost were used to improve the results of the prior regression tree. These techniques as mentioned earlier help in reducing the variance and prevent overfitting to the training data as well

as improve the predictive accuracy resulting in a more generalized model. These techniques were performed using R programming software (Appendix A.4). The following cross-validation plots of bagging , random forest and XGBoost were constructed after we developed the models, in order to evaluate the prediction error from each model.

For instance, Figure 2.25 illustrating the predicted vs observed values of cumulative gas production using bagging approach. The bagging model's calculated prediction errors and corresponding R-squared are as follows:

$RMSE = 1547.66\ MMscf$

$MSE = 2395252\ MMscf^2$

$R^2 = 85.9\%$

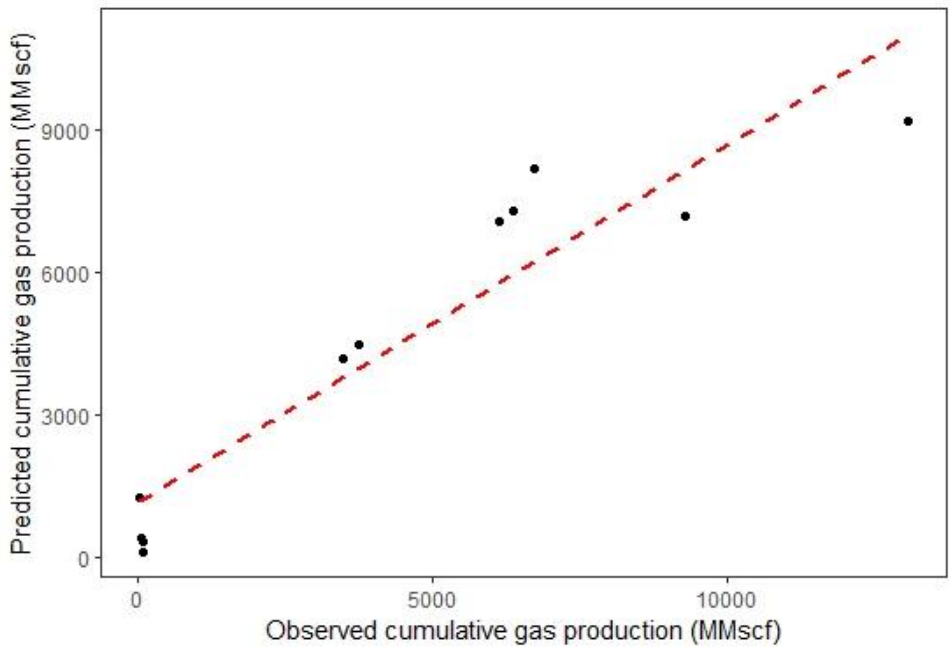


Figure 5. 25 : Observed vs predicted cumulative gas production for bagging.

Figure 2.26 as well illustrating the predicted vs observed values of cumulative gas production using random forest approach. The random forest model's calculated prediction errors and corresponding R-squared are as follows:

$RMSE = 1486.17\ MMscf$

$MSE = 2208688\ MMscf^2$

$R^2 = 87\%$

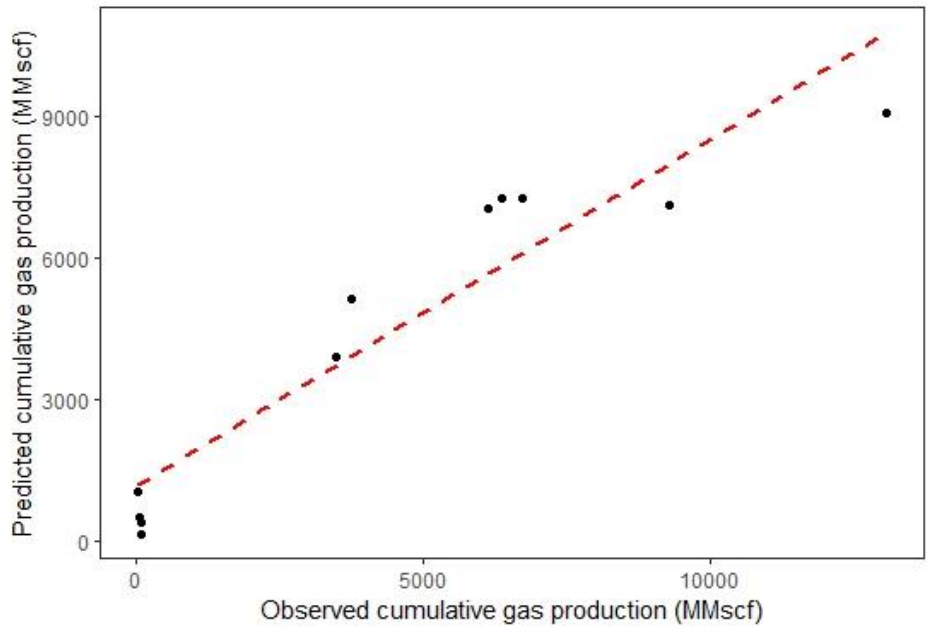


Figure 5. 26 : Observed vs predicted cumulative gas production for RF.

Finally, the cross-validation plot in Figure 2.27 is used to estimate the prediction errors of forecasting the cumulative gas production with XGBoost model. The model's calculated prediction errors and corresponding R-squared are as follows:

$RMSE = 1218.54\ MMscf$

$MSE = 1484842\ MMscf^2$

$R^2 = 91.3\%$

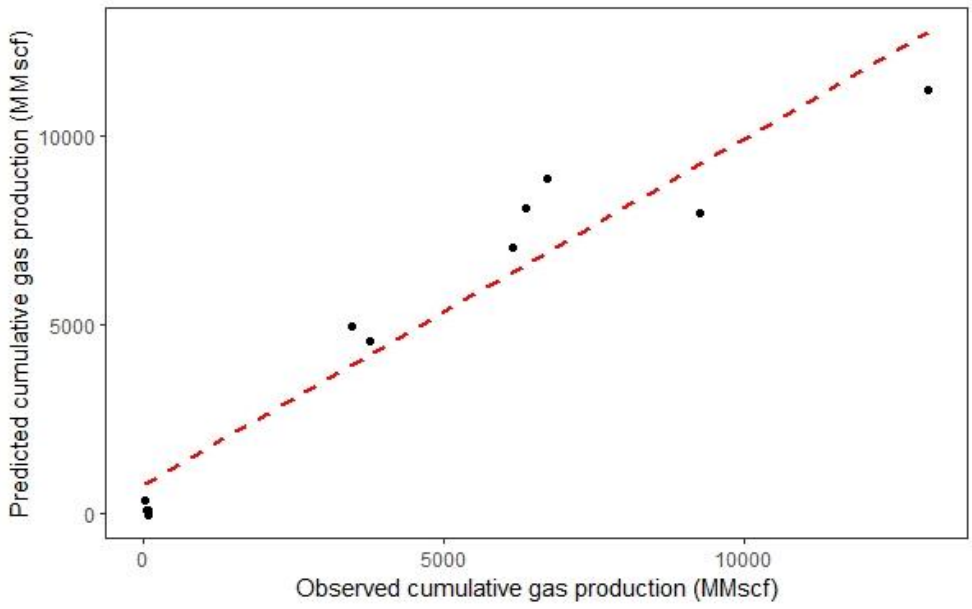


Figure 5. 27 : Observed vs predicted cumulative gas production for XGBoost.

When comparing the results of tree-based models, XGBoost shows the best predictive performance as it yields the highest R-squared and lowest prediction error (RMSE). Furthermore, Table 5.8 illustrate the overall results of predictive models that we used in this study including multiple linear regression.

Table 5. 8 : Overall model results, SPE 2021 shale wells data

Model	R^2	MSE ($MMscf^2$)	RMSE ($MMscf$)
Multiple Linear Regression	89	2925936	1710.53
Regression Tree	78	3793177	1947.61
Bagging	86	2395252	1547.66
Random Forest	87	2208688	1486.17
XGBoost	91	1484842	1218.54

It is observed from Table 5.8 that extreme gradient boosting machine (XGBoost) exhibits a superior performance in predicting the cumulative gas production after one year from unconventional shale reservoirs compared to all other data-driven methods, demonstrating the highest R^2 of 91% and the lowest RMSE of 1218.54 MMscf.

The obtained results are consistent with the theoretical principles of XGBoost and agree to the academia literature that assume XGBoost as one of the most robust machine-learning algorithms. For instance, a study by Pan et al. (2022) to predict the reservoir porosity shows that XGBoost outperform linear regression, support vector regression, and Random forest. Another study by Liu et al. (2019) for predicting water absorption in sublayers shows that the forecasting performance of XGBoost is more suitable than support vector machine (SVM). Tang et al. (2021) as well obtained the best result in predicting shale reservoirs quality with XGBoost when compared to SVM and other ensemble methods like RF. Furthermore, a study by Zhao et al. (2022) utilized 7 different predictive models including RF, SVM , K-Nearest Neighbors (KNN), and XGBoost to forecast the permeability of low-permeable sandstones, XGBoost demonstrates superior forecasting performance among the seven models. Finally, research by Ma et al. (2020) from the field of medicine demonstrates that XGBoost outperforms models like RF, KNN, SVM, etc. in predicting the cancer stage of patients.

5.5.3 Variable Importance

The main purpose of this section is to determine the main driver parameters that influence that cumulative gas production (response) among a wide range of reservoir and operational parameters (predictors). It is important to ask which variables have the highest predictive power. In order to perform the variable importance random forest and XGBoost have inbuilt function that help in conduction the variable importance. This function utilized to plot the increase in MSE of predictions versus the predictors. For instance, to determine the predictive power of each variable in a RF model, the increase in MSE is computed for a specific permutated variable while all other variables remain unchanged. XGBoost calculate the 'gain' metric which use MSE to identify the improvement in prediction accuracy for each variable.

It can be observed from Figure 5.28 which represent random forest variable importance that gas saturation (Sg) is the most influential parameter with the highest impact on the cumulative gas production after one year followed by the number of stages, the number of clusters by stage, porosity, bottom perforation and lateral length.

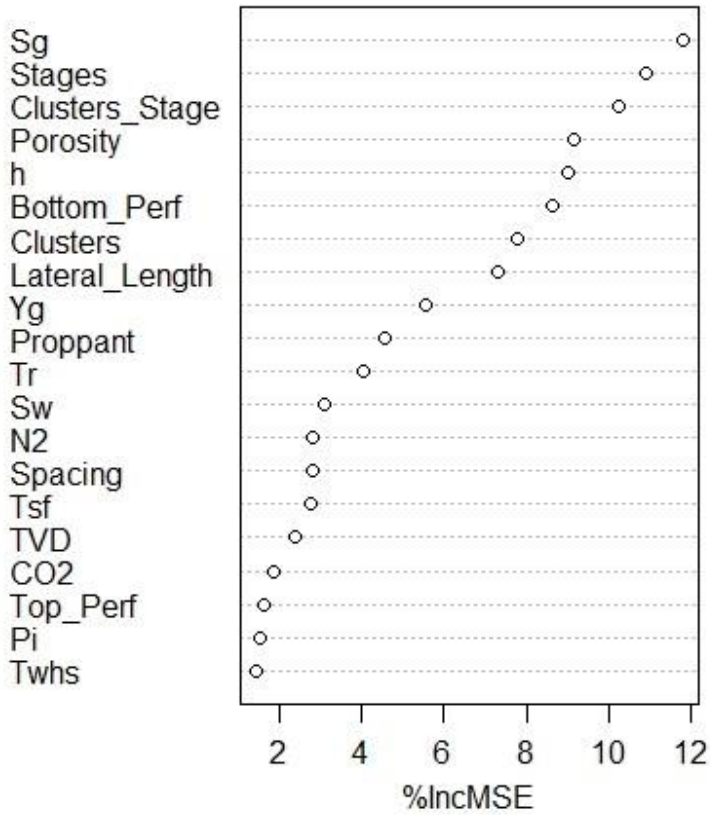


Figure 5. 28 : Importance of variables in the random forest model

Moreover, it is noticeable in Figure 5.29, which represent XGBoost variable relative influence that gas saturation (Sg) is the most influential parameter that significantly influencing the cumulative gas production after one year followed by the number of stages, lateral length, porosity, thickness, bottom perforation and the number of clusters per stage.

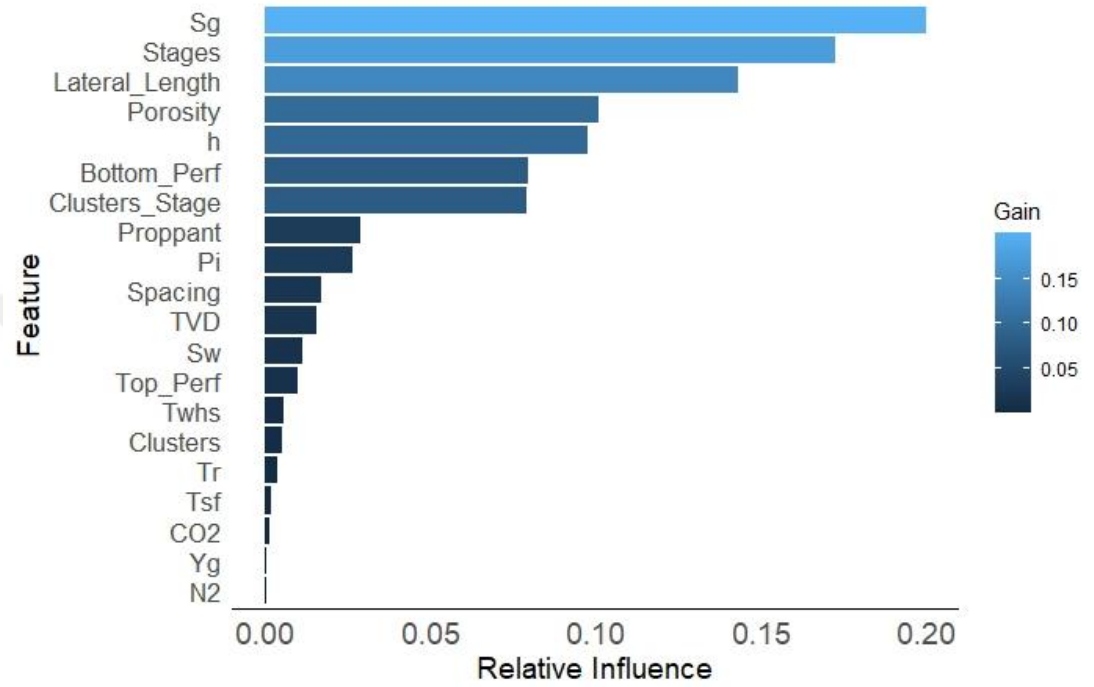


Figure 5. 29 : Relative influence for XGBoost model

It's worth mentioning that both Random Forest and XGBoost models exhibit a good agreement regarding the top 6 influential parameters, where gas saturation (sg) emerges as the most influential parameter. The only exception is the ranking of the third parameter, as XGBoost places lateral length in that position, while Random Forest assigns that position to the number of clusters per stage. These top parameters have the most predictive power in forecasting the cumulative gas production after one year from unconventional shale reservoirs.

Furthermore, XGBoost makes more sense for lateral length to be the third most significant parameter influencing shale gas wells performance. This also agrees with the regression tree, where the main parameters utilized to split the tree were gas saturation and lateral length. The obtained results exhibit notable consistency with findings from a similar study (Schuetter et al., 2018).

5.5.4 Comparison of Results

It's crucial to emphasize that the purpose of this study is to forecast the cumulative gas production after a year. However, since certain wells produced for varying lengths of time, we initially retrieved the data for one year of production time. This enables us to collect the equivalent gas production values from all wells for a full year in order to create a consistent, generalizable model. The models were built upon these data for one year and the results is obtained as shown earlier.

The dataset was then examined as it corresponded to an average of three years of cumulative production. This is an average time of production for all wells otherwise the actual time was different for each well. Then all steps we did above for one year production time were repeated but this time for an average of three years of production time. The obtain results is illustrated in (Appendix B)

The overall results of an average 3 years production time shows in regression tree Figure B.1 (Appendix B) that porosity is the main influential parameter since it located on the top of the tree followed by net pay thickness.

Furthermore, random forest variable importance in Figure B.2 illustrates that porosity is the main influential parameter along with gas saturation and thickness. XGBoost relative influence in Figure B.3 shows a perfect agreement with random forest on the top three influential parameters, porosity, gas saturation, and thickness.

As a comparison between results of 3 year and 1 year of gas production time, it can be concluded that for a long period of production form unconventional gas reservoirs, the reservoir parameters such as porosity, gas saturation, and thickness are more dominant than operational parameters. Which mean the reservoir characteristics matters more for a longer time production. In the other hand, for one year production time, we observed that operational parameters such as lateral length and hydraulic fracturing characteristic along with gas saturation is more dominate. These results is consistent with the need of both horizontal drilling and hydraulic fracturing methods for a better production from unconventional shale reservoirs.

6. CONCLUSIONS

In this study, predictive models based on machine learning were developed using R Statistical Computing Environment to forecast the cumulative gas production after 1 year from unconventional shale reservoirs, including MLR, regression tree, bagging, RF and XGBoost regression models. This study used a recently published dataset which consist of 53 observation (gas wells) and 20 different reservoir and operational parameters (SPE, 2021). Exploratory data analysis (EDA) is employed to check visually the quality of the data and gain a better understanding of the dataset. After examining all reservoir and operational parameters relationship with the cumulative gas production, machine learning based forecasting models are developed. The developed models then are evaluated to select the best model that can provide a more accurate forecast of the cumulative gas production after one year. Finally, variable importance approach is considered to determine the key influential parameters with the most effect on the shale gas wells performance. The study's findings can be concluded in the following manner:

1. Predicting cumulative gas production through data-driven methods is feasible in unconventional shale reservoirs. This approach has proven to be efficient in accurately predicting gas production fast and cost-effective manner.
2. Five statistical and machine learning models, which include MLR, regression tree, bagging, RF, and XGBoost regression models, were utilized to forecast the cumulative gas production after 1 year from unconventional shale reservoirs. Among five models evaluated, XGBoost demonstrated a superior prediction performance, attaining the highest R^2 score of 91% and the lowest RMSE of 1218.54 MMscf. This finding is consistent with many academia literature that assume XGBoost as one of the most robust machine learning algorithms.
3. XGBoost and RF agrees on the top 6 influential parameters that influence the gas production performance from shale wells, except for the ranking of the third parameter, where lateral length is prioritized for XGBoost and the number of clusters per stage for RF. These 6 main influential parameters are:

- Gas saturation (Sg)
- Number of stages
- Lateral length (XGboost), Number of clusters per stage (RF)
- Porosity
- Thickness
- Bottom perforation

4. Gas saturation is the most critical parameter in forecasting the cumulative gas production in unconventional shale reservoirs since it ranked first in XGBoost, RF and Regression tree main split (root node).
5. The comparison of 1 year cumulative production and 3 years cumulative production results revealed that, for long term production, reservoir parameters are more dominant, where porosity, gas saturation, and thickness were the main influential parameters. In the other hand, for short term production operational parameters such as number of stages, lateral length, bottom perforation and number of clusters per stage are more dominate. The short term results are consistent with the findings of similar studies (Schuetter et al., 2018)
6. The multiple linear regression method performed best after reducing the number of contributing parameters to the model from 20 to 13. The number of parameters was identified using k-fold cross-validation and extracted using the best subset selection algorithm.

6.1 Recommendations

The obtained results of this study is based on the available dataset (SPE, 2021). The developed models can be also generalized to a similar projects that have similar reservoir and operational parameters as the ones shown in Table 5.5.

XGBoost model has many optimization parameters that can control the results of the model. Hence these optimizations should be identified carefully to avoid overfitting. Cross validation can be used to determine the best tune of these optimizations.

This study can be extended to include some other critical parameters that were not available in our dataset, such as petrophysical properties (permeability, lithology-particularly the clay content in shale, etc.), and the mechanical properties (Young's Modulus, Poisson's ratio, fracture toughness, natural fracture system, etc.).

REFERENCES

Al-Alwani, M. A., Britt, L., Dunn-Norman, S., Alkinani, H. H., Al-Hameedi, A. T., & Al-Attar, A. (2019). Production Performance Estimation from Stimulation and Completion Parameters Using Machine Learning Approach in the Marcellus Shale. In *53rd U.S. Rock Mechanics/Geomechanics Symposium* (p. ARMA-2019-2034).

Arora, S., & Doshi, P. (2021). A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence*, *297*, 103500. https://doi.org/10.1016/J.ARTINT.2021.103500

Artun, E. (2020). Performance assessment and forecasting of cyclic gas injection into a hydraulically fractured well using data analytics and machine learning. *Journal of Petroleum Science and Engineering*, *195*, 107768. https://doi.org/10.1016/J.PETROL.2020.107768

Beck, F., & Latif, S. (2022). *Introduction to Multivariate Data Visualization*. https://doi.org/10.5281/zenodo.6523081

Belyadi, H., & Haghighat, A. (2021). Supervised learning. *Machine Learning Guide for Oil and Gas Using Python*, 169–295. https://doi.org/10.1016/B978-0-12-821929-4.00004-4

Bhattacharya, S., Ghahfarokhi, P. K., Carr, T. R., & Pantaleone, S. (2019). Application of predictive data analytics to model daily hydrocarbon production using petrophysical, geomechanical, fiber-optic, completions, and surface data: A case study from the Marcellus Shale, North America. *Journal of Petroleum Science and Engineering*, *176*, 702–715. https://doi.org/10.1016/J.PETROL.2019.01.013

Bowker, K. A. (2007). Barnett Shale gas production, Fort Worth Basin: Issues and discussion. *AAPG Bulletin*, *91*(4), 523–533. https://doi.org/10.1306/06190606018

Bruce, P., Bruce, A., & Gedeck, P. (2020). *Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python* (2nd ed.). O'Reilly Media. https://www.oreilly.com/library/view/practical-statistics-for/9781492072935/

Chahar, J., Verma, J., Vyas, D., & Goyal, M. (2022). Data-driven approach for hydrocarbon production forecasting using machine learning techniques. *Journal of Petroleum Science and Engineering*, *217*, 110757. https://doi.org/10.1016/J.PETROL.2022.110757

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Ertekin, T., & Sun, Q. (2019). Artificial Intelligence Applications in Reservoir Engineering: A Status Check. In *Energies* (Vol. 12, Issue 15). https://doi.org/10.3390/en12152897

Esmaili, S., & Mohaghegh, S. D. (2016). Full field reservoir modeling of shale assets using advanced data-driven analytics. *Geoscience Frontiers*, *7*(1), 11–20. https://doi.org/10.1016/J.GSF.2014.12.006

Favero, L. P., Belfiore, P., & de Freitas Souza, R. (2023). *Data Science, Analytics and Machine Learning with R*. Elsevier Science.

Gong, X., Tian, Y., McVay, D. A., Ayers, W. B., & Lee, W. J. (2013). Assessment of Eagle Ford Shale Oil and Gas Resources. In *SPE Unconventional Resources Conference Canada* (p. SPE-167241-MS). https://doi.org/10.2118/167241-MS

Gumus, M., & Kiran, M. S. (2017). Crude oil price forecasting using XGBoost. *2017 International Conference on Computer Science and Engineering (UBMK)*, 1100–1103. https://doi.org/10.1109/UBMK.2017.8093500

Guo, J., Zhang, L., Wang, H., & Feng, G. (2012). Pressure Transient Analysis for Multi-stage Fractured Horizontal Wells in Shale Gas Reservoirs. *Transport in Porous Media*, *93*(3), 635–653. https://doi.org/10.1007/s11242-012-9973-4

Gupta, S., Fuehrer, F., & Jeyachandra, B. C. (2014). Production Forecasting in Unconventional Resources using Data Mining and Time Series Analysis. In *SPE/CSUR Unconventional Resources Conference – Canada* (p. D011S004R008). https://doi.org/10.2118/171588-MS

Hammes, U., & Gale, J. (2014). *Geology of the Haynesville Gas Shale in East Texas and West Louisiana*. American Association of Petroleum Geologists. https://doi.org/10.1306/M1051344

Han, D., & Kwon, S. (2021). Application of Machine Learning Method of Data-Driven Deep Learning Model to Predict Well Production Rate in the Shale Gas Reservoirs. In *Energies* (Vol. 14, Issue 12). https://doi.org/10.3390/en14123629

Hanga, K. M., & Kovalchuk, Y. (2019). Machine learning and multi-agent systems in oil and gas industry applications: A survey. *Computer Science Review*, *34*, 100191. https://doi.org/10.1016/J.COSREV.2019.08.002

Heller, R., Vermylen, J., & Zoback, M. (2014). Experimental investigation of matrix permeability of gas shales. *AAPG Bulletin*, *98*(5), 975–995. https://doi.org/10.1306/09231313023

Holdaway, K. R. (2009). Exploratory Data Analysis in Reservoir Characterization Projects. In *SPE/EAGE Reservoir Characterization and Simulation Conference* (p. SPE-125368-MS). https://doi.org/10.2118/125368-MS

Holdaway, K. R. (2014). *Harness Oil and Gas Big Data with Analytics: Optimize Exploration and Production with Data-Driven Models*. Wiley. https://onlinelibrary.wiley.com/doi/book/10.1002/9781118910948

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: with Applications in R* (2nd ed., Vol. 102). Springer US. https://doi.org/https://doi.org/10.1007/978-1-0716-1418-1

Kalantari-Dahaghi, A. (2022). Machine learning applications in unconventional shale gas systems. *Unconventional Shale Gas Development: Lessons Learned*, 433–443. https://doi.org/10.1016/B978-0-323-90185-7.00015-7

Kaliraj, P., & Devi, T. (2021). *Big Data Applications in Industry 4.0* (1st ed.). CRC Press. https://doi.org/10.1201/9781003175889

Kassambara, A. (2018). *Machine Learning Essentials: Practical Guide in R*. sthda. https://books.google.com.tr/books?id=745QDwAAQBAJ

Kim, T. H., & Lee, K. S. (2015). Pressure-Transient Characteristics of Hydraulically Fractured Horizontal Wells in Shale-Gas Reservoirs With Natural- and Rejuvenated-Fracture Networks. *Journal of Canadian Petroleum Technology*, *54*(04), 245–258. https://doi.org/10.2118/176027-PA

Kong, B., Chen, Z., Chen, S., & Qin, T. (2021). Machine learning-assisted production data analysis in liquid-rich Duvernay Formation. *Journal of Petroleum Science and Engineering*, *200*, 108377. https://doi.org/10.1016/J.PETROL.2021.108377

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. In *Applied Predictive Modeling* (1st ed.). Springer New York, NY. https://doi.org/10.1007/978-1-4614-6849-3

Kulga, B., Artun, E., & Ertekin, T. (2017). Development of a data-driven forecasting tool for hydraulically fractured, horizontal wells in tight-gas sands. *Computers & Geosciences*, *103*, 99–110. https://doi.org/10.1016/J.CAGEO.2017.03.009

LaFollette, R. F., Izadi, G., & Zhong, M. (2013). Application of Multivariate Analysis and Geographic Information Systems Pattern-Recognition Analysis to Production Results in the Bakken Light Tight Oil Play. In *SPE Hydraulic Fracturing Technology Conference* (p. D021S006R003). https://doi.org/10.2118/163852-MS

Lee, J. H., Shin, J., & Realff, M. J. (2018). Machine learning: Overview of the recent progresses and implications for the process systems engineering field. *Computers & Chemical Engineering*, *114*, 111–121. https://doi.org/10.1016/J.COMPCHEMENG.2017.10.008

Leem, J., Mazeli, A. H., Musa, I. H., & Che Yusoff, M. F. (2022). Data Analytics and Machine Learning Predictive Modeling for Unconventional Reservoir Performance Utilizing Geoengineering and Completion Data: Sweet Spot Identification and Completion Design Optimization. In *ADIPEC* (p. D021S062R002). https://doi.org/10.2118/211024-MS

Lesmeister, C. (2019). *Mastering Machine Learning with R: Advanced machine learning techniques for building smart applications with R 3.5, 3rd Edition*. Packt Publishing. https://books.google.com.tr/books?id=lCWGDwAAQBAJ

Liu, W., Liu, W. D., Gu, J., & Shen, X. (2019). Predictive model for water absorption in sublayers using a machine learning method. *Journal of Petroleum Science and Engineering*, *182*, 106367. https://doi.org/10.1016/J.PETROL.2019.106367

Loh, W.-Y. (2011). Classification and regression trees. *WIREs Data Mining and Knowledge Discovery*, *1*(1), 14–23. https://doi.org/https://doi.org/10.1002/widm.8

Lolon, E., Hamidieh, K., Weijers, L., Mayerhofer, M., Melcher, H., & Oduba, O. (2016). Evaluating the Relationship Between Well Parameters and Production Using Multivariate Statistical Models: A Middle Bakken and Three Forks Case History. In *SPE Hydraulic Fracturing Technology Conference* (p. D031S007R003). https://doi.org/10.2118/179171-MS

Ma, B., Meng, F., Yan, G., Yan, H., Chai, B., & Song, F. (2020). Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data. *Computers in Biology and Medicine*, *121*, 103761. https://doi.org/10.1016/J.COMPBIOMED.2020.103761

Martinez, W. L., Martinez, A. R., & Solka, J. (2017). *Exploratory Data Analysis with MATLAB* (3rd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315366968

Michael, A. (2021). Hydraulic Fractures from Non-Uniform Perforation Cluster Spacing in Horizontal Wells: Laboratory Testing on Transparent Gelatin. *Journal of Natural Gas Science and Engineering*, *95*, 104158. https://doi.org/10.1016/J.JNGSE.2021.104158

Mishra, S., & Datta-Gupta, A. (2018). Applied Statistical Modeling and Data Analytics: A Practical Guide for the Petroleum Geosciences. In *Applied Statistical Modeling and Data Analytics: A Practical Guide for the Petroleum Geosciences*. Elsevier Inc. https://doi.org/10.1016/C2014-0-03954-8

Mishra, S., & Lin, L. (2017). Application of Data Analytics for Production Optimization in Unconventional Reservoirs: A Critical Review. In *SPE/AAPG/SEG Unconventional Resources Technology Conference* (p. URTEC-2670157-MS). https://doi.org/10.15530/URTEC-2017-2670157

Mohaghegh, S D, Gaskari, R., & Maysami, M. (2017). Shale Analytics: Making Production and Operational Decisions Based on Facts: A Case Study in Marcellus Shale. In *SPE Hydraulic Fracturing Technology Conference and Exhibition* (p. D031S007R004). https://doi.org/10.2118/184822-MS

Mohaghegh, Shahab D. (2013). A Critical View of Current State of Reservoir Modeling of Shale Assets. In *SPE Eastern Regional Meeting* (p. SPE-165713-MS). https://doi.org/10.2118/165713-MS

Mohammed, M., Khan, M., & Bashier, E. (2016). *Machine Learning: Algorithms and Applications* (1st ed.). CRC Press. https://doi.org/10.1201/9781315371658

Moore, D. S., Notz, W. I., & Fligner, M. A. (2017). *The Basic Practice of Statistics* (8th ed.). W. H. Freeman.

Mousavi, S. M., Jabbari, H., Darab, M., Nourani, M., & Sadeghnejad, S. (2020). Optimal Well Placement Using Machine Learning Methods: Multiple Reservoir Scenarios. In *SPE Norway Subsurface Conference* (p.

D021S008R005). https://doi.org/10.2118/200752-MS

Nejad, A. M., Sheludko, S., Shelley, R. F., Hodgson, T., & McFall, R. (2015). A Case History: Evaluating Well Completions in the Eagle Ford Shale Using a Data-Driven Approach. In *SPE Hydraulic Fracturing Technology Conference* (p. D021S004R004). https://doi.org/10.2118/SPE-173336-MS

Nguyen-Le, V., Shin, H., & Little, E. (2020). Development of shale gas prediction models for long-term production and economics based on early production data in Barnett reservoir. *Energies*, *13*(2). https://doi.org/10.3390/en13020424

Nisbet, R., Miner, G. D., & Yale, K. (2018). *Handbook of Statistical Analysis and Data Mining Applications* (2nd ed.). Elsevier Science. https://www.sciencedirect.com/book/9780124166325/handbook-of-statistical-analysis-and-data-mining-applications

Niu, W., Lu, J., & Sun, Y. (2022). Development of shale gas production prediction models based on machine learning using early data. *Energy Reports*, *8*, 1229–1237. https://doi.org/10.1016/J.EGYR.2021.12.040

NTNG. (2016). *An Energy Revolution : 35 Years of Fracking in the Barnett Shale*.

Ogunleye, A., & Wang, Q.-G. (2020). XGBoost Model for Chronic Kidney Disease Diagnosis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *17*(6), 2131–2140. https://doi.org/10.1109/TCBB.2019.2911071

Pan, S., Zheng, Z., Guo, Z., & Luo, H. (2022). An optimized XGBoost method for predicting reservoir porosity using petrophysical logs. *Journal of Petroleum Science and Engineering*, *208*, 109520. https://doi.org/10.1016/J.PETROL.2021.109520

Pope, C. D., Palisch, T., & Saldungaray, P. (2012). *Improving Completion And Stimulation Effectiveness In Unconventional Reservoirs - Field Results In The Eagle Ford Shale Of North America*. *All Days*. https://doi.org/10.2118/152839-MS

Railroad Commission of Texas. (2013). Eagle Ford Shale Information & Statistics. *Railroad Commission of Texas*. https://www.rrc.texas.gov/

Sahai, R. (2022). Field development and asset management. *Unconventional Shale Gas Development: Lessons Learned*, 1–31. https://doi.org/10.1016/B978-0-323-90185-7.00019-4

Saldaña, M., Gálvez, E., Navarra, A., Toro, N., & Cisternas, L. A. (2023). Optimization of the SAG Grinding Process Using Statistical Analysis and Machine Learning: A Case Study of the Chilean Copper Mining Industry. In *Materials* (Vol. 16, Issue 8). https://doi.org/10.3390/ma16083220

Schuetter, J., Mishra, S., Zhong, M., & LaFollette (ret.), R. (2018). A Data-Analytics Tutorial: Building Predictive Models for Oil Production in an Unconventional Shale Reservoir. *SPE Journal*, *23*(04), 1075–1089. https://doi.org/10.2118/189969-PA

Shahkarami, A., Ayers, K., Wang, G., & Ayers, A. (2018). Application of Machine

Learning Algorithms for Optimizing Future Production in Marcellus Shale, Case Study of Southwestern Pennsylvania. In *SPE/AAPG Eastern Regional Meeting* (p. D043S005R004). https://doi.org/10.2118/191827-18ERM-MS

Smith, G. (2011). *Essential Statistics, Regression, and Econometrics*. Elsevier Science. https://www.elsevier.com/books/essential-statistics-regression-and-econometrics/smith/978-0-12-382221-5

SPE. (2021). *SPE Data Repository: Data Set: 1. From URL." https://www.spe.org/datasets/ (accessed 10 July 2021)*.

Speight, J. G. (2007). CHAPTER 8 - Emissions Control and Environmental Aspects. In J. G. Speight (Ed.), *Natural Gas* (pp. 193–208). Gulf Publishing Company. https://doi.org/https://doi.org/10.1016/B978-1-933762-14-2.50013-3

Speight, J. G. (2020). Chapter 2 - Resources. In J. Speight (Ed.), *Shale Oil and Gas Production Processes* (pp. 65–138). Gulf Professional Publishing. https://doi.org/https://doi.org/10.1016/B978-0-12-813315-6.00002-6

Suriamin, F., & Ko, L. T. (2022). Geological characterization of unconventional shale-gas reservoirs. *Unconventional Shale Gas Development: Lessons Learned*, 33–70. https://doi.org/10.1016/B978-0-323-90185-7.00006-6

Tan, C., Yang, J., Cui, M., Wu, H., Wang, C., Deng, H., & Song, W. (2021). Fracturing Productivity Prediction Model and Optimization of the Operation Parameters of Shale Gas Well Based on Machine Learning. *Lithosphere*, *2021*(Special 4), 2884679. https://doi.org/10.2113/2021/2884679

Tang, J., Fan, B., Xiao, L., Tian, S., Zhang, F., Zhang, L., & Weitz, D. (2021). A New Ensemble Machine-Learning Framework for Searching Sweet Spots in Shale Reservoirs. *SPE Journal*, *26*(01), 482–497. https://doi.org/10.2118/204224-PA

Team, R. C. (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.r-project.org/

Tukey, J. W. (1977). *Exploratory Data Analysis* (Issue v. 2). Addison-Wesley Publishing Company. https://books.google.com.tr/books?id=UT9dAAAAIAAJ

U.S. Energy Information administration. (2011). Review of Emerging Resources: U.S. Shale Gas and Shale Oil Plays. *U.S. Energy Information Administration*, *July*, 105 pp. www.eia.gov

Utts, J. M. (2014). *Seeing Through Statistics* (4th ed.). Cengage Learning. https://books.google.com.tr/books?id=MerKAgAAQBAJ

Wei, T., & Simko, V. (2021). *An Introduction to corrplot Package*. https://taiyun.github.io/corrplot/

Wilcox, R. R. (2021). *Introduction to Robust Estimation and Hypothesis Testing* (5th ed.). Elsevier Science. https://www.elsevier.com/books/introduction-to-robust-estimation-and-hypothesis-testing/wilcox/978-0-12-820098-8

Zhao, X., Chen, X., Huang, Q., Lan, Z., Wang, X., & Yao, G. (2022). Logging-data-driven permeability prediction in low-permeable sandstones based on machine learning with pattern visualization: A case study in Wenchang A Sag, Pearl River Mouth Basin. *Journal of Petroleum Science and Engineering*, *214*, 110517. https://doi.org/10.1016/J.PETROL.2022.110517

Zhong, M., Schuetter, J., Mishra, S., & LaFollette, R. F. (2015). Do Data Mining Methods Matter? : A Wolfcamp "Shale" Case Study. In *SPE Hydraulic Fracturing Technology Conference* (p. D021S004R002). https://doi.org/10.2118/SPE-173334-MS

Zhou, X., & Ran, Q. (2023). Optimization of Fracturing Parameters by Modified Genetic Algorithm in Shale Gas Reservoir. In *Energies* (Vol. 16, Issue 6). https://doi.org/10.3390/en16062868

Zoback, M. D., & Kohli, A. H. (2019). *Unconventional Reservoir Geomechanics*. Cambridge University Press. https://doi.org/10.1017/9781316091869

APPENDICES

APPENDIX A: Steps of Developing a Predictive Model

APPENDIX B: Additional 3 Years Result Graphs

APPENDIX A: Steps of developing a predictive model

APPENDIX A.1 - Best subset selection

Conducting subset selection enables us to identify the variables that make more significant contribution to the model.

```
library(leaps)

one_yearData <-read_excel("1-yearData.xlsx",
                          sheet = 'Sheet1',
                          range = 'A1:V54')

data_1year<-one_yearData %>% remove_rownames %>% column_to_rownames
(var="Names")

regfit.full=regsubsets(Gp~.,data_1year)
summary(regfit.full)

##
## Call: regsubsets.formula(Gp ~ ., data_1year)
## 20 Variables  (and intercept)
##                Forced in Forced out
## Pi                 FALSE      FALSE
## Tr                 FALSE      FALSE
## h                  FALSE      FALSE
## Porosity           FALSE      FALSE
## Sw                 FALSE      FALSE
## Sg                 FALSE      FALSE
## Yg                 FALSE      FALSE
## CO2                FALSE      FALSE
## N2                 FALSE      FALSE
## TVD                FALSE      FALSE
## Spacing            FALSE      FALSE
## Stages             FALSE      FALSE
## Clusters           FALSE      FALSE
## Clusters_Stage     FALSE      FALSE
## Proppant           FALSE      FALSE
## Lateral_Length     FALSE      FALSE
## Top_Perf           FALSE      FALSE
## Bottom_Perf        FALSE      FALSE
## Tsf                FALSE      FALSE
## Twhs               FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
```

```
##          Pi  Tr  h   Porosity Sw  Sg  Yg  CO2 N2  TVD Spacing St
ages Clusters
## 1  ( 1 ) " " " " " " " " " "      " " "*" " " " " " " " " " " "     "
"    " "
## 2  ( 1 ) " " " " " " " " " "      " " " " " " " " " " " " " " "     "
"    " "
## 3  ( 1 ) " " " " " " " " "*"      " " " " " " " " " " " " " " "     "
"    " "
## 4  ( 1 ) " " " " " " " " "*"      " " " " " " " " " " " " " " "     "
"    " "
## 5  ( 1 ) " " " " " " " " "*"      " " " " " " " " " " " " " " "     "
"    " "
## 6  ( 1 ) " " " " " "*" " " "      " " "*" " " " " " " " " " " "     "
"    " "
## 7  ( 1 ) " " " " " "*" " " "      "*" "*" " " " " " " " " " " "*"     "
"    " "
## 8  ( 1 ) " " " " " "*" " " "      " " "*" " " " " " " " " " " " "     "*
"    "*"
##          Clusters_Stage Proppant Lateral_Length Top_Perf Bottom_
Perf Tsf Twhs
## 1  ( 1 ) " "            " "      " "            " "      " "
" " " "
## 2  ( 1 ) "*"            " "      " "            " "      "*"
" " " "
## 3  ( 1 ) "*"            " "      "*"            " "      " "
" " " "
## 4  ( 1 ) "*"            " "      "*"            " "      " "
" " "*"
## 5  ( 1 ) "*"            " "      "*"            "*"      "*"
" " " "
## 6  ( 1 ) "*"            " "      "*"            "*"      "*"
" " " "
## 7  ( 1 ) " "            " "      "*"            "*"      "*"
" " " "
## 8  ( 1 ) " "            "*"      "*"            "*"      "*"
" " " "
```

In order to check the contribution of each variable based on R-squared, the following code is implemented.

```
Reg_full<-regsubsets(Gp~.,data=data_1year,nvmax=20)
reg_summary<-summary(reg_full)
reg_summary$rsq
```

```
## [1] 0.4901123 0.6936556 0.7530214 0.7743149 0.8036208 0.8171479
0.8355141
## [8] 0.8486500 0.8669330 0.8776631 0.8845854 0.8931219 0.8964397
0.8984356
## [15] 0.8987136 0.9000131 0.9001178 0.9001754 0.9001960 0.9001991
```

APPENDIX A.2 – Extracting best variables

Selecting the best variables that will be included in multiple linear regression model. The variables with 'TRUE' are included whereas the ones with 'FALSE' is removed from the model.

```
library(leaps)

one_yearData <-read_excel("1-yearData.xlsx",
                          sheet = 'Sheet1',
                          range = 'A1:V54')

data_1year<-one_yearData %>% remove_rownames %>% column_to_rownames
(var="Names")

bestsub.model<-regsubsets(Gp~ Pi+Tr+h+Porosity+Sw+Sg+Yg+CO2+
                          N2+TVD+Spacing+Stages+Clusters+
                          Clusters_Stage+Proppant+Lateral_Length+
                          Top_Perf+Bottom_Perf+Tsf+Twhs,
                          data=data_1year,
                             nvmax = 20)

summary_best_subset<-summary(bestsub.model)
summary_best_subset$which[13,]
```

```
##    (Intercept)             Pi             Tr              h       Porosity
##           TRUE           TRUE          FALSE           TRUE          FALSE
##             Sw             Sg             Yg            CO2             N2
##           TRUE           TRUE           TRUE          FALSE          FALSE
##            TVD        Spacing         Stages       Clusters Clusters_Stage
##           TRUE           TRUE           TRUE           TRUE          FALSE
##       Proppant Lateral_Length       Top_Perf    Bottom_Perf            Tsf
##           TRUE           TRUE           TRUE           TRUE          FALSE
##           Twhs
##          FALSE
```

APPENDIX A.3 – Multiple linear regression

R code for performing multiple linear regresision model.

```
one_yearData <-read_excel("1-yearData.xlsx",
                          sheet = 'Sheet1',
                          range = 'A1:V54')

data_1year<-one_yearData %>% remove_rownames %>% column_to_rownames
(var="Names")

MLR<-lm(Gp~ Pi+h+Sw+Sg+Yg+TVD+Spacing+Stages+Clusters
              +Proppant+Lateral_Length+Top_Perf+Bottom_Perf,
               data=data_1year)

summary(MLR)

##
## Call:
## lm(formula = Gp ~ Pi + h + Sw + Sg + Yg + TVD + Spacing + Stages
+
##     Clusters + Proppant + Lateral_Length + Top_Perf + Bottom_Per
f,
##     data = data_1year)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2188.0  -655.9  -118.0   665.6  2826.2
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.225e+04  6.528e+03  -1.877 0.068076 .
## Pi              3.932e-01  2.710e-01   1.451 0.154736
## h              -2.057e+01  5.261e+00  -3.909 0.000359 ***
## Sw              1.464e+04  4.666e+03   3.138 0.003234 **
## Sg              8.440e+03  1.901e+03   4.440 7.19e-05 ***
## Yg              4.447e+03  3.979e+03   1.118 0.270494
## TVD            -2.006e+00  1.026e+00  -1.955 0.057760 .
## Spacing         2.168e+00  9.558e-01   2.268 0.028960 *
## Stages          1.326e+02  3.392e+01   3.910 0.000359 ***
## Clusters       -4.631e+00  2.034e+00  -2.276 0.028392 *
## Proppant       -1.677e-04  5.383e-05  -3.115 0.003443 **
## Lateral_Length  1.264e+01  2.433e+00   5.196 6.73e-06 ***
## Top_Perf        1.396e+01  2.447e+00   5.705 1.33e-06 ***
## Bottom_Perf    -1.222e+01  2.429e+00  -5.031 1.14e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1239 on 39 degrees of freedom
## Multiple R-squared:  0.8964, Adjusted R-squared:  0.8619
## F-statistic: 25.97 on 13 and 39 DF,  p-value: 3.526e-15
```

APPENDIX A.4 – Tree-based methods

R code for performing tree-based predictive models including regression tree, bagging, RF, and XGBoost.

```
library(tree)

one_yearData <-read_excel("1-yearData.xlsx",
                          sheet = 'Sheet1',
                          range = 'A1:V54')

data_1year<-one_yearData %>% remove_rownames %>% column_to_rownames
(var="Names")

##Splitting the dataset##

set.seed(123)
indexSet<-sample(2,nrow(data_1year),replace = T,prob = c(0.8,0.2))
Train<-data_1year[indexSet==1, ]
Test<-data_1year[indexSet==2, ]

##Regression Tree##

model_tree<-tree(Gp ~ ., data= Train)
summary(model_tree)

#Pruning#
pruned_model<- prune.tree (model_tree, best = 4)
summary(pruned_model)


## Variables actually used in tree construction:
## [1] "Sg"             "Clusters_Stage" "Lateral_Length"
## Number of terminal nodes:  4
## Residual mean deviance:  3193000 = 121300000 / 38
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -3425.0  -997.6  -482.9     0.0   684.9  4720.0

Predictions_tree <- predict (pruned_model , newdata = Test)

##Bagging model##
library(randomForest)

set.seed(123)

bagging_model <- randomForest(Gp ~ ., data = Train,
                              mtry=20,importance=TRUE)


bagging_preds <- predict(bagging_model, newdata = Test)
```

```r
##Random forest model##

set.seed(123)

Rf_model <- randomForest(Gp ~ ., data = Train ,
                         mtry=10,importance=TRUE)


Rf_preds <- predict(Rf_model, newdata = Test)



##XGBoost model##
library(xgboost)

library(caret)

XGBoost_model <-train(Gp ~., data=Train, method ="xgbTree" ,
              trControl =trainControl("cv"  , number =10 ))

XGBoost_model$bestTune

##    nrounds max_depth eta gamma colsample_bytree min_child_weight
subsample
##      150         3 0.3     0              0.8                1
1

set.seed(123)

train_d <- xgb.DMatrix(data = as.matrix(Train[, -ncol(Train)]), lab
el = Train[, ncol(Train)])
 test_d  <- xgb.DMatrix(data = as.matrix(Test[, -ncol(Test)]), labe
l = Test[, ncol(Test)])

params <- list(
  objective = "reg:squarederror",
  eta = 0.3,
  max_depth = 3,
  subsample = 0.6,
  colsample_bytree = 0.8,
  min_child_weight = 1
)

xgboost_model <- xgb.train(
  params = params,
  data = train_d,
  nrounds = 150,
  early_stopping_rounds = 10,
  watchlist = list(train = train_d, test = test_d),
  verbose = 0


xgb_preds <- predict(xgboost_model, test_d)
```

```r
##Results of all models##

          ## MSE ##
cat("Regression Tree MSE:",  tree_mse, "\n")

## Regression Tree MSE: 3793177

cat("Bagging  MSE: ", bagging_mse, "\n")

## Bagging  MSE:  2395252

cat("Random Forest MSE:", Rf_mse, "\n")

## Random Forest MSE: 2208688

cat("XGBoost model MSE:", XGBoost_mse, "\n")

## XGBoost model MSE: 1484842

     ##  R Squared   ##
cat("Regression Tree R Squared:", tree_rsquared, "\n")

## Regression Tree R Squared: 0.778181

cat("Bagging R Squared:", bagging_rsquared, "\n")

## Bagging R Squared: 0.8599294

cat("Random Forest R Squared:", rf_rsquared, "\n")

## Random Forest R Squared: 0.8708394

cat("XGBoost R Squared:", xgb_rsquared, "\n")

## XGBoost R Squared: 0.9131688
```

APPENDIX B: Additional 3 Years Result Graphs

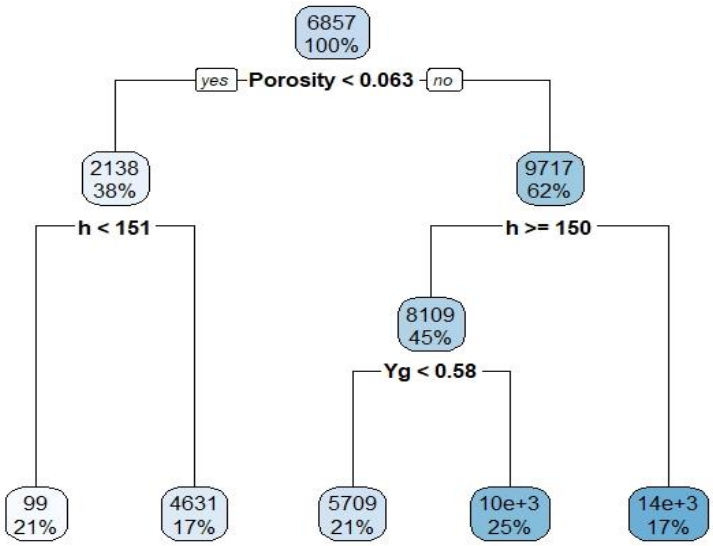


Figure B. 1 : Regression tree model, 3 years-production

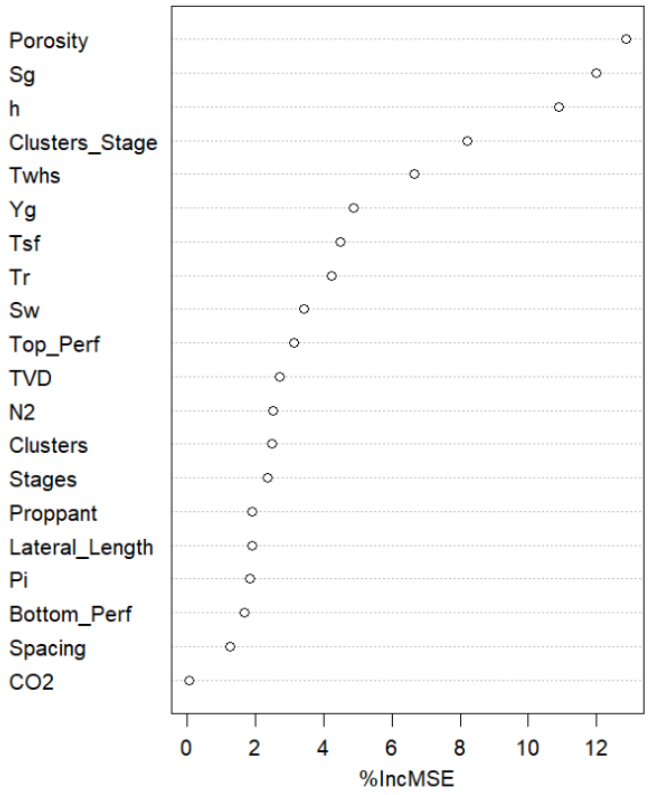


Figure B. 2 : RF variable importance, 3 years-production

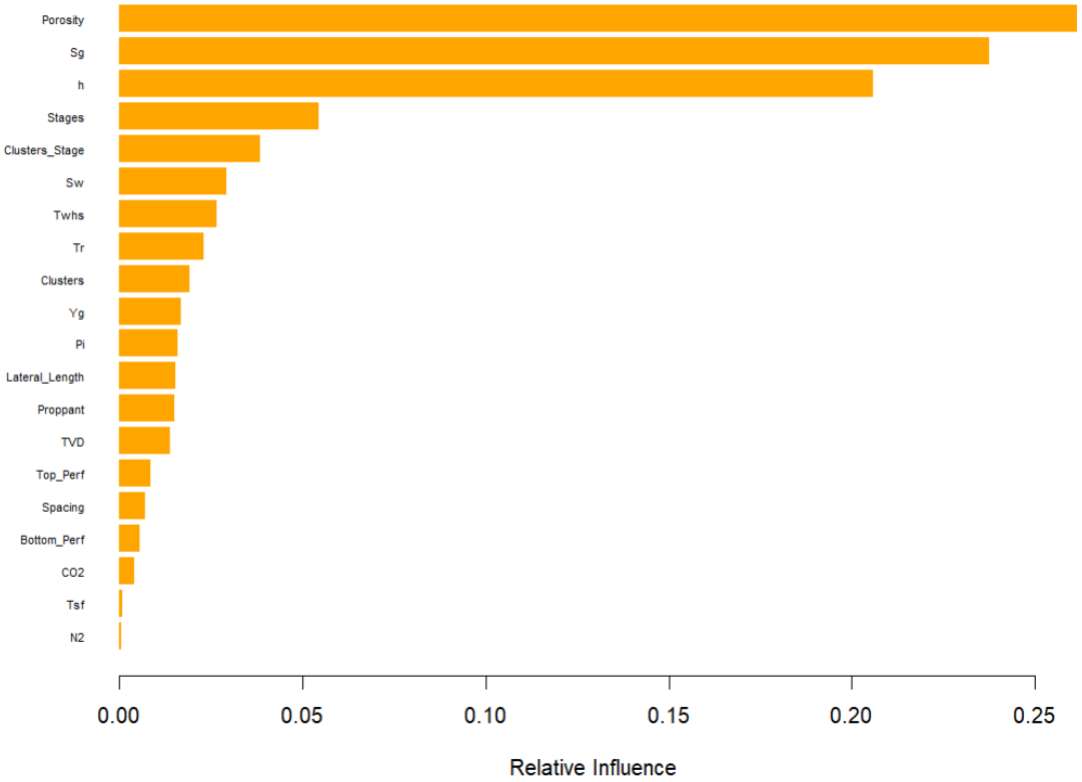
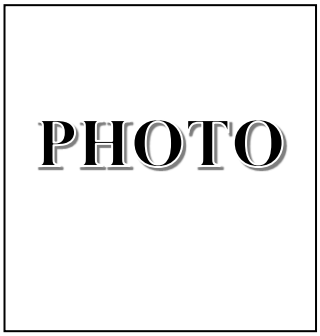


Figure B. 3 : XGBoost relative influence, 3 years-production

CURRICULUM VITAE



Name Surname : Mohammed SHEDAIVA

Place and Date of Birth :

E-Mail :

EDUCATION :

- **B.Sc. :**

PROFESSIONAL EXPERIENCE AND REWARDS:

PUBLICATIONS, PRESENTATIONS AND PATENTS ON THE THESIS:

OTHER PUBLICATIONS, PRESENTATIONS AND PATENTS: