ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

A NOVEL APPROACH TO HOUSE PRICE PREDICTION USING SEGMENTED BAYESIAN OPTIMIZATION BASED ENSEMBLE MODELS

M.Sc. THESIS

Artun PINAR

Department of Data Engineering and Business Analytics

Big Data and Business Analytics Programme

JANUARY 2026

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

A NOVEL APPROACH TO HOUSE PRICE PREDICTION USING SEGMENTED BAYESIAN OPTIMIZATION BASED ENSEMBLE MODELS

M.Sc. THESIS

Artun PINAR
(528221067)

Department of Data Engineering and Business Analytics

Big Data and Business Analytics Programme

Thesis Advisor: Assoc. Prof. Tuncay ÖZCAN

JANUARY 2026

ISTANBUL TEKNİK ÜNİVERSİTESİ ★ LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

SEGMENTASYON İLE BAYES OPTİMİZASYON TABANLI ENSEMBLE MODELLERİ KULLANARAK EV FİYATLARI ÖNGÖRÜSÜNE YENİ BİR YAKLAŞIM

YÜKSEK LİSANS TEZİ

Artun PINAR
(528221067)

Veri Mühendisliği ve İş Analitiği Anabilim Dalı

Büyük Veri ve İş Analitiği Programı

Tez Danışmanı: Doç. Dr. Tuncay ÖZCAN

OCAK 2026

Artun Pınar, a M.Sc./a Ph.D. student of İTU Graduate School student ID 528221067 successfully defended the thesis/dissertation entitled “A NOVEL APPROACH TO HOUSE PRICE PREDICTION USING SEGMENTED BAYESIAN OPTIMIZATION BASED ENSEMBLE MODELS”, which he/she prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor :	**Assoc. Prof. Tuncay ÖZCAN** İstanbul Technical University	.....
Jury Members :	**Prof. Dr. Tolga KAYA** Istanbul Technical University	.....
	Prof. Dr. Tarık KÜÇÜKDENİZ Istanbul University-Cerrahpasa	.....

Date of Submission : 29 December 2025
Date of Defense : 23 January 2026

To my supporting family,

FOREWORD

I would like to express my sincere gratitude to my advisor, Assoc. Prof. Tuncay Ozcan, for his expert guidance, continuous support, and valuable feedback throughout this study. It has been a privilege to work under his mentorship.

I would like to thank my friends for their constant support and motivation during this journey.

Finally, I would like to express my deepest appreciation to my family for their unconditional love, patience, and belief in me. This thesis would not have been possible without their support.

January 2026 Artun PINAR

TABLE OF CONTENTS

Page

ABBREVIATIONS

ANN	**:** Artificial Neural Network
CatBoost	**:** Categorical Boosting
CRISP-DM	**:** Cross-Industry Standard Process for Data Mining
DNN	**:** Deep Neural Networks
GBDT	**:** Gradient Boosting
GBDT	**:** Gradient Boosting Decision Tree
GLM	**:** Generalized Linear Model
GPU	**:** Graphics Processing Unit
IQR	**:** Interquartile Range
LightGBM	**:** Light Gradient Boosting Machine
MAE	**:** Mean Absolute Error
MAPE	**:** Mean Absolute Percentage Error
MRA	**:** Multiple Regression Analysis
MSE	**:** Mean Squared Error
PCA	**:** Principal Component Analysis
SEM	**:** Spatial Econometrics
SMAPE	**:** Symmetrical Mean Absolute Percentage Error
SRA	**:** Stepwise Regression Analysis
STN	**:** Spatial Transformer Network
SVC	**:** Spatially Varying Coefficient Models
SVM	**:** Support Vector Machine
SVR	**:** Support Vector Regression
RAM	**:** Random Access Memory
RM	**:** Random Forest
RM	**:** Malaysian Ringgit
RMSE	**:** Root Mean Squared Error
RMSLE	**:** Root Mean Squared Logarithmic Error
XGBoost	**:** Extreme Gradient Boosting
WCSS	**:** Within-cluster Sum of Squares
USA	**:** United States of America

LIST OF TABLES

Page

LIST OF FIGURES

Page

A NOVEL APPROACH TO HOUSE PRICE PREDICTION USING SEGMENTED BAYESIAN OPTIMIZATION BASED ENSEMBLE MODELS

SUMMARY

The housing market is a multifaceted system that not only satisfies the fundamental human need for shelter but also functions as a critical instrument for wealth accumulation, investment planning, and financial decision-making. Over the past decades, residential properties have evolved into complex economic assets, whose valuation is affected by a wide array of factors ranging from structural characteristics and location to broader economic conditions and market sentiment. The global financial crises, including the 2008 mortgage collapse and the subsequent disruptions caused by the Covid-19 pandemic, have vividly demonstrated the volatility and sensitivity of housing prices to macroeconomic shocks, inflationary pressures, and shifting consumer expectations. These dynamics are particularly pronounced in developing economies such as Türkiye, where economic fragility and rapid urbanization create an environment of heightened uncertainty, making reliable, data-driven methods for predicting property values indispensable for buyers, sellers, investors, financial institutions, and policy regulators.

This study addresses these challenges by developing a comprehensive, machine learning-based framework for predicting housing prices within the Istanbul real estate market. The dataset utilized includes a broad spectrum of property-specific, locational, and neighborhood-level attributes, such as net and gross square meter areas, number of rooms, building age, floor level, total floors, heating type, listing category, and district and neighborhood-level price indicators. To ensure data quality and analytical reliability, an extensive preprocessing pipeline was applied. This pipeline involved cleaning and standardization of entries, resolution of inconsistent or missing values, and mitigation of outliers through Winsorization, which reduces the influence of extreme values while preserving natural distributional characteristics. Categorical features, including district and neighborhood identifiers, were target-encoded in a manner designed to prevent information leakage, while the target variable itself underwent a log1p transformation to stabilize variance and enhance model learning efficiency.

A critical component of the study was the development of engineered features to better capture underlying relationships that are not directly observable in raw data. Features such as floor ratios, area-per-room metrics, bathroom-to-area ratios, and binary indicators for properties within residential complexes were introduced to quantify nuanced aspects of property structure and usability. Additionally, average price-based encodings at district and neighborhood levels allowed the models to incorporate local market expectations, which are known to strongly influence individual property valuations. Correlation-based feature selection and redundancy elimination were further applied to retain only the most informative predictors, improving both model efficiency and interpretability.

A novel aspect of the methodology was the segmentation of the dataset into distinct market segments, recognizing that Istanbul's housing market is inherently heterogeneous. Properties were clustered based on a combination of price ranges, structural characteristics, and neighborhood-level indicators, allowing each segment to reflect unique market dynamics. Segment-specific modeling enabled the identification of differentiated feature importance patterns and improved predictive performance, as individual models could adapt to the structural, locational, and economic nuances of their respective groups. This approach also provides meaningful insights into how property characteristics influence pricing differently across luxury, mid-range, and smaller-scale residential segments.

Five ensemble-based regression algorithms—XGBoost, LightGBM, CatBoost, Gradient Boosting, and Random Forest—were trained and optimized using Bayesian hyperparameter search through Optuna. Evaluation employed multiple complementary metrics, including R^2, MAE, RMSE, and SMAPE, offering a comprehensive assessment of both accuracy and predictive stability. Segment-specific hyperparameter optimization further strengthened each model's adaptability, while an R^2-weighted ensemble of the optimized models was constructed to combine their predictive strengths. This ensemble consistently outperformed individual models, demonstrating enhanced generalization, stability across heterogeneous property groups, and improved resistance to overfitting.

In conclusion, this research presents a robust, interpretable, and highly adaptable machine learning framework for housing price prediction in dynamic urban markets. By combining rigorous preprocessing, advanced feature engineering, segment-specific modeling, and ensemble learning, the study not only delivers high predictive accuracy but also provides actionable insights into the determinants of property values. The proposed methodology offers practical benefits for buyers, sellers, financial institutions, and policymakers by enabling data-driven decision-making and fairer market assessments. Future extensions may involve incorporating temporal and macroeconomic indicators, spatial econometric techniques, and deeper hierarchical segmentation strategies, which could further enhance the framework's capacity to capture evolving market behaviors and complex nonlinear relationships in real estate pricing.

SEGMENTASYON İLE BAYES OPTİMİZASYON TABANLI ENSEMBLE MODELLERİ KULLANARAK EV FİYATLARI ÖNGÖRÜSÜNE YENİ BİR YAKLAŞIM

ÖZET

Konut piyasası, yalnızca temel insan barınma ihtiyacını karşılamakla kalmayıp aynı zamanda servet birikimi, yatırım planlaması ve finansal karar alma süreçlerinde kritik bir araç olarak da işlev gören çok yönlü bir sistemdir. Son on yıllarda konutlar, değerlemeleri yapısal özelliklerden ve konumdan daha geniş ekonomik koşullara ve piyasa duyarlılığına kadar çok çeşitli faktörlerden etkilenen karmaşık ekonomik varlıklara dönüşmüştür. 2008'deki ipotek krizi ve ardından gelen Covid-19 pandemisinin neden olduğu aksamalar da dahil olmak üzere küresel finansal krizler, konut fiyatlarının makroekonomik şoklara, enflasyonist baskılara ve değişen tüketici beklentilerine karşı oynaklığını ve hassasiyetini açıkça ortaya koymuştur. Bu dinamikler, ekonomik kırılganlığın ve hızlı kentleşmenin artan belirsizlik ortamı yarattığı ve alıcılar, satıcılar, yatırımcılar, finans kurumları ve politika düzenleyicileri için güvenilir, veriye dayalı mülk değerlerini tahmin etme yöntemlerini vazgeçilmez kıldığı Türkiye gibi gelişmekte olan ekonomilerde özellikle belirgindir.

Bu çalışma, İstanbul gayrimenkul piyasasındaki konut fiyatlarını tahmin etmek için kapsamlı, makine öğrenimine dayalı bir çerçeve geliştirerek bu zorlukları ele almaktadır. Kullanılan veri seti, net ve brüt metrekare alanları, oda sayısı, bina yaşı, kat seviyesi, toplam kat sayısı, ısıtma türü, listeleme kategorisi ve ilçe ve mahalle düzeyinde fiyat göstergeleri gibi mülke özgü, konumsal ve mahalle düzeyinde geniş bir özellik yelpazesini içermektedir. Veri kalitesini ve analitik güvenilirliği sağlamak için kapsamlı bir ön işleme süreci uygulanmıştır. Bu süreç, girdilerin temizlenmesi ve standartlaştırılmasını, tutarsız veya eksik değerlerin çözümlenmesini ve doğal dağılım özelliklerini korurken uç değerlerin etkisini azaltan Winsorization aracılığıyla aykırı değerlerin azaltılmasını içermektedir. İlçe ve mahalle tanımlayıcıları da dahil olmak üzere kategorik özellikler, bilgi sızıntısını önleyecek şekilde hedef kodlanmış, hedef değişkenin kendisi ise varyansı sabitlemek ve model öğrenme verimliliğini artırmak için bir log1p dönüşümünden geçirilmiştir.

Çalışmanın kritik bir bileşeni, ham verilerde doğrudan gözlemlenemeyen temel ilişkileri daha iyi yakalamak için tasarlanmış özelliklerin geliştirilmesiydi. Kat oranları, oda başına alan metrikleri, banyo/alan oranları ve konut komplekslerindeki mülkler için ikili göstergeler gibi özellikler, mülk yapısının ve kullanılabilirliğinin nüanslı yönlerini nicelleştirmek için tanıtıldı. Ayrıca, ilçe ve mahalle düzeyindeki ortalama fiyat tabanlı kodlamalar, modellerin bireysel mülk değerlemelerini güçlü bir şekilde etkilediği bilinen yerel pazar beklentilerini de dahil etmesini sağladı. Korelasyon tabanlı özellik seçimi ve fazlalık giderme, yalnızca en bilgilendirici öngörücüleri korumak için daha da uygulandı ve hem model verimliliği hem de yorumlanabilirliği artırıldı.

Metodolojinin yenilikçi bir yönü, İstanbul'un konut piyasasının doğası gereği heterojen olduğu gerçeğini göz önünde bulundurarak, veri setinin farklı pazar segmentlerine ayrılmasıydı. Mülkler, fiyat aralıkları, yapısal özellikler ve mahalle düzeyindeki göstergelerin bir kombinasyonuna göre kümelendi ve her segmentin benzersiz pazar dinamiklerini yansıtması sağlandı. Segmente özgü modelleme, farklılaştırılmış özellik önem modellerinin belirlenmesini ve bireysel modeller ilgili gruplarının yapısal, konumsal ve ekonomik nüanslarına uyum sağlayabildiğinden, tahmin performansının iyileştirilmesini sağladı. Bu yaklaşım, lüks, orta sınıf ve daha küçük ölçekli konut segmentlerinde mülk özelliklerinin fiyatlandırmayı nasıl farklı şekilde etkilediğine dair anlamlı içgörüler de sağlar.

XGBoost, LightGBM, CatBoost, Gradient Boosting ve Random Forest olmak üzere beş topluluk tabanlı regresyon algoritması, Optuna aracılığıyla Bayes hiperparametre araması kullanılarak eğitilmiş ve optimize edilmiştir. Değerlendirme, R^2, MAE, RMSE ve SMAPE dahil olmak üzere birden fazla tamamlayıcı metrik kullanarak hem doğruluk hem de öngörü kararlılığı açısından kapsamlı bir değerlendirme sunmuştur. Segmente özgü hiperparametre optimizasyonu, her modelin uyarlanabilirliğini daha da güçlendirirken, optimize edilmiş modellerin R^2 ağırlıklı bir topluluğu, öngörü güçlerini birleştirmek için oluşturulmuştur. Bu topluluk, bireysel modellerden sürekli olarak daha iyi performans göstererek, gelişmiş genelleme, heterojen mülk grupları arasında kararlılık ve aşırı uyum sağlamaya karşı gelişmiş direnç göstermiştir.

Sonuç olarak, bu araştırma, dinamik kentsel pazarlarda konut fiyatı tahmini için sağlam, yorumlanabilir ve son derece uyarlanabilir bir makine öğrenimi çerçevesi sunmaktadır. Titiz ön işleme, gelişmiş özellik mühendisliği, segmente özgü modelleme ve topluluk öğrenimini bir araya getiren çalışma, yalnızca yüksek tahmin doğruluğu sağlamakla kalmıyor, aynı zamanda özellik değerlerinin belirleyicilerine ilişkin eyleme geçirilebilir içgörüler de sağlıyor. Önerilen metodoloji, pratik faydalar sunuyor. Veriye dayalı karar alma ve daha adil piyasa değerlendirmeleri sağlayarak alıcılar, satıcılar, finans kurumları ve politika yapıcılar için daha fazla olanak sağlayacaktır. Gelecekteki genişletmeler, zamansal ve makroekonomik göstergelerin, mekansal ekonometrik tekniklerin ve daha derin hiyerarşik segmentasyon stratejilerinin dahil edilmesini içerebilir; bu da çerçevenin, gayrimenkul fiyatlandırmasındaki değişen piyasa davranışlarını ve karmaşık doğrusal olmayan ilişkileri yakalama kapasitesini daha da artırabilir.

Bununla birlikte, çalışmanın bulguları makine öğrenimi tabanlı değerleme modellerinin yalnızca teknik doğruluk açısından değil, aynı zamanda piyasa şeffaflığı ve bilgi asimetrisinin azaltılması açısından da önemli katkılar sunduğunu göstermektedir. Özellikle karmaşık ve hızla değişen kentsel piyasalarda, geleneksel değerleme yaklaşımlarının sınırlı kaldığı durumlarda, veri odaklı modeller karar alma süreçlerini destekleyici tamamlayıcı araçlar olarak öne çıkmaktadır. Bu bağlamda geliştirilen çerçeve, konut piyasasının yapısal karmaşıklığını nicel olarak ele alabilen esnek bir altyapı sunmakta ve farklı paydaşların risk değerlendirmesi, fiyat beklentisi oluşturma ve stratejik planlama süreçlerinde daha tutarlı ve öngörülebilir çıktılar elde etmesine olanak tanımaktadır. Çalışma, akademik literatüre yöntemsel bir katkı sağlamanın yanı sıra, Türkiye özelinde gayrimenkul piyasasının analitik olarak incelenmesine yönelik uygulanabilir ve ölçeklenebilir bir referans modeli ortaya koymaktadır.

Ayrıca, çalışmada izlenen yaklaşımın ölçeklenebilir yapısı, farklı şehirler, bölgesel alt piyasalar ve değişen veri yapıları için kolaylıkla uyarlanabilir olduğunu göstermektedir. Model mimarisinin modüler olarak kurgulanmış olması, yeni değişkenlerin eklenmesine, alternatif segmentasyon stratejilerinin test edilmesine ve farklı öğrenme algoritmalarının entegrasyonuna olanak tanımaktadır. Bu esneklik, özellikle veri erişiminin zaman içinde genişlediği veya piyasa koşullarının hızla değiştiği ortamlarda modelin güncelliğini ve geçerliliğini korumasını sağlamaktadır. Dolayısıyla önerilen çerçeve, yalnızca tek seferlik bir tahmin çalışması olmanın ötesine geçerek, sürekli güncellenebilir ve karar destek sistemlerine entegre edilebilir dinamik bir değerleme altyapısı sunmaktadır.

1. INTRODUCTION

Background Information:

The housing market, in its simplest form ,serves the one of the most crucial needs of human beings which is having a shelter to live. Housing need was always a important aspect of life since the beginning of time. First, it was caves; now it is mostly apartments, houses or skyscrapers that cater the modern-day housing needs. Surely, the housing market became more than just shelter for people, it became a wealth preservation and creation tool as an investment and that made the progression much faster due to increased demand during development ages, emerging capitalist ideologies, financial tools, such as mortgage systems, made it easier to capture interest of many people all around the world.

In recent years, the world experienced a major crisis, which emerged from the derivative tools related to housing market loans and their valuation underscored the risks inherent in housing-related financial products. (2008-Mortgage crisis, housing bubble) Then, the Covid-19 pandemic caused global inflation burst, and developing countries were more severely impacted due to their economic fragility. Naturally, these developments affected the housing market, and pricing became unclear due to inflation. Prices changed rapidly as a result of both economic conditions and people's perception of the markets and long term loans such as mortgages.

Therefore, it is safe to say that house prices get affected by many variables, and selection of the variables that ultimately affect the price of a house-such as location, year built, room number, overall size etc.- will help to better identify and select the optimum house and its possible fair transfer price.

Problem Statement:

As stated above, the housing market is influenced by various conditions and it is important to address the right components that have meaningful relationship with the appropriate value of a property during the fluctuating market conditions and social perceptions. Investigating and selecting the variables that drive prices in the most

influential way, will help related parties in the housing market to better evaluate and make decisions about properties more efficiently.

Significance of the Study:

The ultimate goal would be to increase efficiency in the housing market as a whole by enabling fairer transactions due to reliable price prediction models, and possibly making government assessments simpler with their data-driven nature to better create and implement regulations for the housing market. Additionally, the finance sector, especially banks, would benefit from more realistic valuations to provide long-term loans such as mortgages, and they could price their advanced derivative financial products more efficiently. Overall, this study contributes to the market by enhancing insights, improving the decision-making process and creating data backed analyses.

Research Objectives/Hypothesis:

The main objectives of this research are to identify and select the most effective factors that affect pricing in the housing market by developing a real-world data-driven price prediction model with the help of machine learning tools. Accurate, reliable and relevant results will help all parties-including buyers, sellers, financial institutions, and policymakers- better analyze and build various solutions by the detection of meaningful factors and their effects on pricing during the dynamic market conditions.

The remainder of this study is structured as follows. Chapter 2 (Literature Review) includes review of the studies similar to the research subject on housing market and prices. Chapter 3 (Methodology) describes the algorithms and models evaluted for the study. Chapter 4 (Application) includes comprehensive dataset introduction, preprocessing, model evaluation, selection and results. Chapter 5 (Conclusions) briefly discusses the results and gives guidence for future studies.

2. LITERATURE REVIEW

In this section, literature studies on housing market and the analyses of the house prices and its drivers are reviewed.

This study mainly focuses on prediction models about house prices with the aim of identifying key attributes that affects the price significantly. Housing market has many drivers deriving from macroeconomic factors, global political environment, local regulations and incentives, production costs such as land, labor, building materials, machinery etc. Therefore, it is critical to keeping in mind that studies across different continents, countries and cultures can differ greatly. For example, United States has housing market mostly consists of single story houses whereas in Europe most of the people live in apartments. In this aspect, the relevant studies have been studied. There are many different approaches in this specific subject.

2.1 Literature Studies About House Price Prediction

Satish et al. (2019) discussed about the use of machine learning algorithms in prediction of future housing prices. The relevant features to predict the future price of a house found as number of bedrooms, square footage, location and condition of the house. Their dataset included 21,613 observations of house sale transactions in King County, Washington in United States between May 2014 to May 2015. Various prediction methods have been explored and compared to select the better ones among the tested approaches. Linear Regression, Multiple Regression Analysis, Lasso Regression, Gradient Boosting Algorithm analysed to create a holistic Cost Function of a house. Due to its ability to handle high-dimensional data and also preventive habit about overfitting through L1 regularization, Lasso Regression method selected as the primary method of the research. Also, Gradient Boosting techniques implemented to reach better accuracy in prediction by using combination methods to combine weak learners.

Erkek et al. (2020) studied housing market data of three big cities of Turkey which are Istanbul, Izmir and Ankara. They used five-year period from January 2015 to December 2019. It is stated that mainly 15 explanatory variables of a house were used to describe almost every aspect of a given house. Three different data models created by using algorithms that are Support Vector Machine, Feedforward Neural Networks and Generalized Regression Neural Networks algorithms to use as prediction models. The comparison between these models focused on accuracy scores and absolute deviations of test results using Keras library within Python programming language. At the end, it is found that Feedforward Neural Network Model has provided better results in terms of accuracy and consistance than other two models. Also, most important variable of the 15 explanatory variables of a house became location of the house and least important variable turned out to be size of the terrace of a house.

Truong et al. (2020) studied about advanced machine learning techniques about housing price prediction and composed both traditional and advanced machine learning approaches to explore the difference among several models. Housing Price in Beijing, China dataset which consists of 300,000 data with 26 variables as features of a house ranging from 2009 to 2018 from Kaggle is used and several methods applied to validate the performance of the approaches. Random Forest, Extreme Gradient Boosting, Light Gradient Boosting Machine, Hybrid Regression (65% Lasso and 35% XGBoost), Stacked Generalization Regression methods tested. Although having the worst time complexity due to its K-fold cross-validation feature in its mechanism, lowest Root Mean Squared Logarithmic Error (RMSLE) is found in Stacked Generalization approach, so it was applied on the model to optimize predicted values.

Thamarai and Malarvizhi (2020) conducted research on house price prediction modeling with the use of machine learning. The real time real estate dataset from Tadepalligudem, West Godavari District, India is used and consisted of 50 properties and features such as number of bedrooms, age of the house, transportation utilities, schools nearby and shopping facilities. Attribute selection conducted with three steps that are Information Gain, Gain Ratio and lastly Gini Index. The proposed method utilizes the Decision Tree Classifer to predict avaliability of houses according to user requirements, Decision Tree Regression and Multiple Linear Regression methods used to predict the prices of the filtered houses. Scikit Learn which is a Python toolbox for machine learning is used to implement the proposed model. The performance metrics

used for evaluation are Mean Absolute Error, Mean Squared Error and Root Mean Squared Error. The importance of the study that is it studies with real time data and implement two stepped model to the data, first to classify whether the desired requirements are satisfied and there are houses avaliable and then to predict those selected houses with regression analysis. The results showed that Multiple Linear Regression outperformed the Decision Tree Regression by yielding lower error values.

Zulkifley et al. (2020) carried out an survey of literature about forecasting house prices by analyzing most relevant attributes and to create an efficient model that can accurately predict house prices. The study conducted conceptual model and questionnaires on existing house prices in Jakarta, Indonesia. The focus on finding the relevant attributes revealed three main categories as locational, structural and neighborhood condition. Then, machine learning model process started with evaluating traditional methods such as Multiple Linear Regression and Stepwise Regression and advanced valuation methods such as Hedonic Pricing Tool, Support Vector Machine, Artificial Neural Network (ANN), Gradient Boosting Machine and spatial critical methods. According to research, Multiple Regression Analysis was the most used predictive model in the surveyed literatures, then the Artificial Neural Network studies followed. In general, study suggested that Support Vector Regression, Artificial Neural Network and XGBoost are the most efficient predictive methods to use on house price prediction models according to Root Mean Squared Error values.

Deaconu et al. (2021) conducted a study in Romania including a sample data of 900 apartments selling transactions with various locational, physical and neighbourhood related attributes , it is aimed to test the utility and performance of the Artificial Neural Networking (ANN) versus the Generalized Linear Model (GLM) as comparison of artifical intelligence model and regression model. For GLM, four different conditions are tested as M1 being normal distribution with identity link for prices with the area variable, M2 being normal distribution with identity link for prices with the rooms variable, M3 being Gamma distribution with log link with the area variable and M4 being Gamma distribution with log link with the rooms variable. The Likelihood Ratio Chi-Square, R-Square and Mean Absolute Error indicators are compared and M1 performed better in prediction accuracy and overall being better fit. GLM showed good price prediction but failed with non-linear relationships.For ANN, 22 predictor variables which 2 of them continuous and 20 of them being categorical

variables are used as input nodes. One hidden layer with 8 nodes and 1 output node the estimated selling price were used as backpropagation learning algorithm results indicated. ANN identified the most important price determinants as useful area, distance from the city center and finishing quality of the house.The findings revealed that ANN model was better at predicting selling prices of apartments in research data with physical and social attributes. Also, ANN model provided stability over results and greater ability to show significances of different attributes about real estate sample data.

Wang et al. (2021) conducted a research about house price prediction deep learning model with heterogeneous data including joint self-attention mechanism. Satellite maps, public facilities as parks, schools and transportation zones are considered to enhance the study. The dataset consists of 93,696 transactions of 2017-2018 Taipei, Taiwan real estate transactions, public facility data from Taipei government and satellite map information from Google Maps. The proposed prediction model has three main components as public facility data feature representation method to supplement transaction data with public services nearby, image feature extractor including Spatial Transformer Network (STN) to extract crucial features of a given house from satellite maps and joint self-attention mechanism to identify most influential variables that effect housing prices by deriving the weights of each attribute using queries, keys and values from previous steps. Mean Absolute Percentage Error (MAPE) is used to evaluate the performance of the proposed model. Linear Regression, XGBoost and Light Gradient Boosting Machine methods also applied on dataset to compare proposed models' prediction performance. Based on 15 different regions at the dataset, on average of MAPE scores, the proposed model had the lowest MAPE, which explains better overall prediction accuracy.

Abdul-Rahman et al. (2021) focused on house price prediction by searching for the most effective factors that are related to a house to determine their price accurately. Their research was based on the property listings in Kuala Lumpur, Malaysia in 2019 and the dataset had 21,984 observations with 11 variables that are considered as factors of a house. According to Department of Statistics of Malaysia, average price of a house in 2018 was 409k RM and 2009 it was 199k RM. Thus, the dramatic increase in house prices made the subject critically important for people due to it usually being the most expensive transaction in their lives and it seems that it gets harder compared to prior

years. The paper aims to guide in predicticting future house prices as assistance and for better establishment of real estate policies in Malaysia. Machine learning algorithms which are Light Gradient Boosting Machine (LightGBM) and Extreme Gradient Boosting (XGBoost) compared with traditional appoaches which are multiple regression and ridge regression by performance utilization metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and adjusted R-Squared value. Prediction model to gather the target variable which is price; independent variables as location, bedroom, bathroom, car park, size, furnishing status, property type, shopping mall, scholl, hospital and public transport were considered to pick the most influential factors. Pearson Correlation Matrix applied on variables to identify importance and one hot encoding technique applied to express categorical variables in numerical form.XGBoost had the lowest MAE, RMSE and adjusted R-Square value being closest to 1, as it indicted better prediction accuracy it is used to deploy the prediction model.

Mora-Garcia et al. (2022) focused on machine learning algorithms to facilitate best use of them in house price prediction and also quantifying the impact of Covid-19 pandemic on housing market of a Spanish city known as Alicante. They emphasize on detailed data preparation, feature engineering, hyperparameter training and optimization, model evaluation, selection and interpretation. Gradient Boosting Regression, Extreme Gradient Boosting, Light Gradient Boosting Machine as boosting; Random Forest and Extra-Trees Regressor as bagging are bases for ensemble learning algorithms in research, and compared with a linear regression model. The study is developed with georeferenced data with pandemic derivation, other complementary sources such as cadastre, socio-demographic and economic indicators, and satellite images. Traditional linear models fall behind of machine learning algorithms due to better adaptation to nonlinearities of complex data. Bagging algorithms showed overfitting problems and boosting algorithms had better performance and lower overfitting problems. They indicate the importance of avoiding data overfitting for better analysis with house price prediction models.

Kayakuş et al. (2022) used Decision Tree Regression, Artificial Neural Networks and Support Vector Machines methods to estimate the housing sales by monthly data between the 2013-2022 period. Economical variables such as individual mortgage loans volume, interest rates, consumer price index, stock market and

currency movements and also housing sales price per square meter are used in the research. Normalization method was used with linear transformation ranging from 0-1 to improve performance of the machine learning methods and also the results accuracy. Gini Index used as the quality measure of the Decision Trees Regression method with the condition of each node having at least 2 minimum records. As for the Support Vector Regression method; Polynomial, Hyper Tangent and Radial Basis Function techniques tested for the core functions. Radial Basis Function was preferred as core function due to its better performance results in the concluded tests. Overlapping penalty values tested for optimization and the optimum value turned out as 10. Artifical Neural Networks method modelled as to have 100 iterations, 4 hidden layers and 4 neurons per hidden layer in the study. According to coefficient of determination, MSE, RMSE, MAE, MAPE values all the methods analysed using graphical solutions. In the examination, all methods have provided adequate results. In conclusion, it is said that the study results can guide future plans an solutions to the housing market and related sectors as well as creating various financial products suitable for the banking system.

Tekin and Sari (2022) examined the real housing market data with over 36,000 lines including 13 variables to determine how different machine learning algorithms can improve the evaluation process about the real estate prices. Vast data is cleaned and then Linear Regression, Polynomial Regression, Decision Trees, Random Forests, XGBoost are used to create predictive models. CRISP-DM framework used to compare the performances of created predictive models. The researchers created over 100 models with the given 5 different algorithms. Linear Regression, Polynomial Regression and Decision Trees underperformed compared to other two algorithms which are Random Forests and XGBoost. The two better performer models provided similar mean absolute percentage error scores about 20%. It is stated that the significance of optimized parameters which improved the results by 4% to 6% in tested models. As future work recommendations, authors said that a dynamic model which is updated each month with more features and more data should be considered.

Mostofi et al. (2022) studied the Deep Neural Networks (DNN) models application on small-sized real-estate price prediction which was limited due to their reduced accuracy and high dimensioned dataset. The paper uses Principal Component Analysis (PCA) to enhance accuracy of the DNN model driven decisions to influence

small real-estate agencies to use on their local dataset. Dimensionality reduction, dataset transformation and localisation of influential price features are the key improvements of PCA benefits. The dataset consisted of 1,244 house price records in Trabzon, Türkiye. Evaluation metrics used in the research were Mean Squared Error (MSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE). The results showed that the combination of PCA and DNN models achieved better accuracy and generalisation ability than other benchmark predictors such as Stepwise Regression Analysis (SRA). Most impactful variables found as spatial features and building age.

Adetunji et al. (2022) focused on Random Forest machine learning technique for the house price prediction subject. The paper started with explaining why the House Price Index (HPI) which is the repeat-sale index tracking the average price changes in repeated transactions of the assets is ineffective at prediction a single house price due to it being a rough predictor focused on all transactions. The dataset used in the research to propose a prediction model gathered from UCI machine learning repository Boston, Massachusetts housing dataset of the year 1978 with 506 entries featuring 14 features. Performance evaluation metrics selected as Mean Absolute Error (MAE), Coefficient of Determination (R-Squared), and Root Mean Squared Error (RMSE) to assess models' success. In the end; physical conditions, styles, and location of a house found as the three main influential factors of the house price to conduct healthy prediction with the proposed model.

Basysyar and Dwilestari (2022) noted that predicting a price range rather than a single value forecasting such as a house price is more realistic in many real-world applications as House Price Index which consists of mean price variation in repeat sales and refinancing of same assets. Moreover, due to it being dependent on all transactions it lacks the ability to predict the future price of a single house. Linear Regression, Ridge Regression, Lasso Regression, Elastic Net Regression and machine learning for feature selection investigated and used in this study to prepare a prediction model. The dataset used in the research had 1,460 records and 81 features about houses in Ames, Iowa in United States. Recursive Feature Elimination method used to eliminate unnecessary features related to house prices in the dataset. In the result part, after the feature optimization Lasso regression outperformed the other models by

modified alpha, managing reducing the relevant features to only 13 most significant variables with lowest Mean Squared Error score.

Begum et al. (2022) studied various machine learning algorithms like Linear Regression, Decision Tree, Random Forest to create a prediction model for the housing prices. The dataset consisted of 506 samples with 13 features in Boston, United States from January 2015 to November 2019 taken from the Carnegie Mellon University in Pennsylvania, United States. Correlation Analysis used to investigate relationship between quantitative variables among features. Then, min-max scaling as normalization and standardization methods used to employ feature scaling. Three proposed methods which are Linear Regression, Decision Tree and Random Forest Regression evaluated to measure their housing price prediction performances. Mean Squared Error and Cross-Validation performance models are used to identify better performer method and Random Forest method gave the best results in terms of accuracy by lower Mean Squared Error score.

Yağmur et al. (2023) conducted research about house price prediction with mentioning macroeconomic variables and micro-variables which are features of the house. They specified on a area that the housing market is mostly driven by foreigners demand in Antalya, Turkey. 900 house for sale advertisements of August 2022 period are used as dataset. Data were linearly normalized using the min-max method for algorithms. Input variables are stated as; location, usable area, number of rooms, age of residence, floor and social facilities. In the research; lower, middle and upper income groups are all studied and used Artificial Neural Network, Support Vector Regression and Multiple Linear Regression models. Comparison between these models are considered regarding to their R-Square, Mean Square Error and Root Mean Square Error techniques. It is stated that ANN results was more meaningful with better predictions than other models and said that the model can be used to regulate the market in fluctuating house prices conditions.

Geerts et al. (2023) scrutinized 93 articles published between 1992 and 2021 about predicting house prices in order to score their proposed models and data novelties. They mapped the property valuation domain and identification of trends with cluster analysis. The research follows the systematic guidelines of the PSALSAR framework which consists of Protocol, Search and Appraisal, Synthesis, Analysis steps for literature review. In the articles, conventional methods and traditional input data

seemed relevant, but the adaptation of advanced techniques and innovative data sources are present in the newer publications of the dataset. The modeel-based categorization of the research showed that the most 3 used models were SEM (Spatial Econometrics), SVC (Spatially Varying Coefficient Models) and MRA (Multiple Regression Analysis).The opportunities given are to include more advanced input data types which are unstructured data and complex spatial data with deep learning and methods tailored to this spesific area of research.

Zhang et al. (2023) studied about textual description data concepts about house price prediction. The study differs from most researches about house price prediction by not only using numerical attributes of a house transaction, but to use the description text of the house in the house postings and advertisements. Text processing applied with TF-IDF, Word2Vec and BERT word embedding techniques. The dataset gathered from Canadian real estate company website by implementing a web crawler to get housing data of five different cities in Ontario, Canada including 10,418 listings and 57 features between May and June 2021. Feature importance analysis conducted with the SelectKBest feature selection method which is a univariate linear regression that derive a score for each feature. Then, the word embedding step applied as TF-IDF, Word2Vec and BERT to convert house textual descriptions to numerial vector with dimension. Support Vector Regression (SVR), Random Forest, Gradient Boosting and Deep Neural Network learning algorithms selected to conduct tests on the both textual and numerical data. R-Squared and Root Mean Squared Error (RMSE) used as performance evaluation metrics of the learning algorithms. All features considered, Gradient Boosting algorithm matched with TF-IDF word embedding method brings out the best result among all tested algorithms.

Sharma et al. (2024) discusses that the value of a property is not only influenced by physical variables but also significantly influenced by the neighbourhood. They addressed the house price prediction problem which is an important requirement for various sectors, as an example of regression and applied various machine learning techniques to define variables. The datased consisted of 2,930 properties in Ames City, Iowa in United States for 5 years period. Algorithms such as XGBoost, Support Vector Regression, Random Forest Regression, Multilayer Perceptron and Multiple Linear Regression are compared to find the most suitable one to use. For comparison, R-Square, Mean Square Error, Root Mean Square Error, Mean Absolute Error and Cross

Validation Scores are used to create consolidated model scores. In the research, it is stated that XGBoost was the best model for the house price prediction subject as ensemble trees enable easier identification of the effective variables. The top important variables in predicting house prices are “Overall quality of the house”, “Ground floor of the living area”, “Garage for the cars” and “Total basement square feet” according to XGBoost ensemble trees model findings.

Tandon (2024) researched about housing prices trends with 2019 Kuala Lumpur, Malaysia housing market dataset with 21,995 observations. The dataset had eight variables that are price, location, size, kind of house, rooms, bathrooms, and furnishings. Four more variables added with the help of Geocoder and Google Places scripts, distances to closest hospital, school, retail centre and public transportation. Regresison models in machine learning were the scope of the project regarding building a adequate prediction model. Random Forest, Support Vector Machine, Light Gradient Boosting Mchine, Gradient Boosting Regressors models are studied as classification models and compared regarding to their respective results. Root Mean Squared Error (RMSE), Mean Absolute Error (MAE),and R-Squared used as performance measures to evaluate models. Gradient Boosting Regressor was the best overall model with lowest MSE score among all the models tested, so it proved to be most accurate and efficient model for house price prediction model. Socio-economic factors, environmental features and time dynamics offered as future-work suggestions to enhance the framework.

Anelli et al. (2025) focused on the normalization process effects on the complex regression modeling regarding real estate price predictions. They used the real housing market dataset of 117 observations in the city of Rome, Italy during the 2023 period. The data also includes several external variables such as social degration ratio, per capita income of the area, and urbanization levels. As normalization process is critical for a regression model to significantly improve the accuracy, several techniques are tested on the dataset to ensure better prediction performance overall. Min-Max normalization method which squeezes the data between 0 and 1 values, Z-Score standardization method which focuses to get 0 as average and to get 1 as the variance value of the dataset. Lastly, Robust Scaling with median and IQR (interquartile range) processed to get rid of meaningless outliers in the dataset. Evolutionary Polynomial Regression method which is a Genetic Algorithm based

model is used to develop the prediction model rather than generic linear regression methods, due to EPR being more flexible and explanatory model with optimization ability compared to other regression methods. Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-Squared used as performance evaluation metrics and the results showed that EPR with Z-Score and Robust Scaling performed better than Multiple Linear Regression and Support Vector Regression models.

Pastukh and Khomyshyn (2025) studied real estate price prediction using ensemble methods of machine learning on the real dataset of 1,200 properties with nearly 50 variables each, located in city of Ternopil in Ukraine during January 2022 and July 2024 period. Only the direct offerings taken into account as the estate agencies often include commission or fees on top of the property price listings, hence to avoid distortion these offers are not taken into account for the related data collection process of the study. Feature importance analysis is used to define necessary variables to predict prices; location, size and room count being the most important features. Gradient Boosting, Extra Trees, Hist Gradient Boosting, Random Forest Regressor and Linear Regression and several ensemble methods such as Stacking Regressor and Voting Regressor are compared in respect of prediction performance with the given dataset. Common performance evaluation metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-Squared are used to measure performance of the given models. Gradient Boosting Regressor provided the best scores and best accuracy among total of 8 prediction models tested. Extra Trees, Hist Gradient Boosting and Random Forest Regressor followed as their performances became modest compared to Gradient Boosting Regressor. The classification of the literature studies on house price prediction is given in Table 2.1.

Table 2.1 : The Classification of the Literature Studies on House Price Prediction.

Study	Methods	Location (Country/City)	Data size	Performance Metrics
Satish et al. (2019)	Linear Regression, Multiple Regression, Lasso Regression, Gradient Boosting	King County, Washington (USA)	1 year period, 21613 properties	R-Squared
Erkek et al. (2020)	Support Vector Machines, Feedforward Neural Networks, Generalized Regression Neural Network	Istanbul, Izmir and Ankara (Turkey)	5 year period	Mean Absolute Percentage Error
Truong et al. (2020)	Random Forest Regression, Extreme Gradient Boosting, Light Gradient Boosting Machine, Hybrid Regression, Stacked Generalization Regression	Beijing (China)	9 year period, 300000 properties	Root Mean Squared Logarithmic Error
Thamarai and Malarvizhi (2020)	Decision Tree Regression, Multiple Linear Regression	West Godavari (India)	Real time, 50 properties	Mean Absolute Error, Mean Squared Error, Root Mean Squared Error
Zulkifley et al. (2020)	Multiple Linear Regression, Stepwise Regression, Support Vector Machines, Artificial Neural Network, Gradient Boosting Machine	Jakarta (Indonesia)	14 articles	Root Mean Squared Error

Table 2.1 (continued) : The Classification of the Literature Studies on House Price Prediction.

Study	Methods	Location (Country/City)	Data size	Performance Metrics
Deaconu et al. (2021)	Artificial Neural Network, Generalized Linear Model	Cluj-Napoca (Romania)	1 year period, 900 properties	Mean Standard Error, Root Mean Squared Error, Root Mean Square Percentage Error, Goodness of Fit, Mean Raw Residuals
Wang et al. (2021)	Linear Regression, Extreme Gradient Boosting, Light Gradient Boosting Machine	Taipei (Taiwan)	2 years period, 93996 properties	Mean Absolute Percentage Error
Abdul-Rahman et al. (2021)	Light Gradient Boosting Machine, Extreme Gradient Boosting, Multiple Regression, Ridge Regression	Kuala Lumpur (Malaysia)	1 year period, 21984 properties	Mean Absolute Error, Root Mean Squared Error, Adjusted R-Squared
Mora-Garcia et al. (2022)	Linear Regression, Random Forest Regression, Extra Trees Regression, Gradient Boosting Regression, Extreme Gradient Boosting, Light Gradient Boosting Machine	Alicante (Spain)	3 years period, 49875 properties	Mean Absolute Error, Mean Square Error, Root Mean Squared Error, Goodness of Fit
Kayakuş et al. (2022)	Decision Tree Regression, Artificial Neural Network, Support Vector Machines	Turkey	95 months period	Coefficient of Determination, Mean Squared Error, Root Mean Square Error, Mean Absolute Error, Mean Absolute Percentage Error

Table 2.1 (continued) : The Classification of the Literature Studies on House Price Prediction.

Study	Methods	Location (Country/City)	Data size	Performance Metrics
Tekin and Sari (2022)	Linear Regression, Polynomial Regression, Decision Trees Regression, Random Forest Regression, Extreme Gradient Boosting	Istanbul (Turkey)	1 month period, 36759 lines of data	Mean Absolute Percentage Error, R-Squared
Mostofi et al. (2022)	Deep Neural Network	Trabzon (Turkey)	1 year period, 1244 properties	Mean Squared Error, Mean Absolute Error, Mean Absolute Percentage Error
Adetunji et al. (2022)	Rnadom Forest Regression	Boston, Massachusetts (USA)	1 year period, 506 properties	Mean Absolute Error, R-Squared, Root Mean Squared Error
Basysyar and Dwilestari (2022)	Linear Regression, Ridge Regression, Lasso Regression, Elastic Net Regression	Ames, Iowa (USA)	1 year period, 1460 properties	Mean Squared Error
Begum et al. (2023)	Linear Regression, Decision Tree Regression, Random Forest Regression	Pennsylvania (USA)	4 years period, 500 properties	Mean Squared Error

Table 2.1 (continued) : The Classification of the Literature Studies on House Price Prediction.

Study	Methods	Location (Country/City)	Data size	Performance Metrics
Yağmur et al. (2023)	Artificial Neural Network, Multiple Linear Regression, Support Vector Machines	Antalya (Turkey)	1 month period, 900 properties	Coefficient of Determination, Mean Squared Error, Root Mean Square Error
Geerts et al. (2023)	Spatial Econometrics, Spatially Varying Coefficient Models, Multiple Regression Analysis	Various Locations	93 articles	Model-based categorization
Zhang et al. (2023)	Support Vector Regression, Random Forest Regression, Gradient Boosting, Deep Neural Network	Ontario (Canada)	2 months period, 10418 properties	R-Squared, Root Mean Squared Error
Sharma et al. (2024)	Linear Regression, Multi-Layer Perceptron, Random Forest Regression, Support Vector Machines, Extreme Gradient Boosting	Ames City, Iowa (USA)	5 years period, 2930 properties	R-Squared, Adjusted R-Squared, Mean Absolute Error, Mean Square Error, Root Mean Squared Error
Tandon (2024)	Random Forest Regression, Support Vector Machine, Light Gradient Boosting Machine, Gradient Boosting Regressors	Kuala Lumpur (Malaysia)	1 year period, 21995 properties	Root Mean Squared Error, Mean Absolute Error, R-Squared

Table 2.1 (continued) : The Classification of the Literature Studies on House Price Prediction.

Study	Methods	Location (Country/City)	Data size	Performance Metrics
Anelli et al. (2025)	Evolutionary Polynomial Regression, Multiple Linear Regression, Support Vector Regression	Rome (Italy)	6 months period, 117 properties	Root Mean Squared Error, Mean Absolute Error, R-Squared
Pastukh and Khomyshyn (2025)	Linear Regression, Random Forest Regression, Extra Trees Regression, Gradient Boosting Regression, Hist Gradient Boosting Regression	Ternopil (Ukraine)	31 months period, 1200 properties	Root Mean Squared Error, Mean Absolute Error, R-Squared

3.METHODOLOGY

In this section a general overview of studied algorithms and models in this study is provived.

3.1 Machine Learning

Machine learning refers to a set of computational methods that make predictions or decisions by automatically learning patterns from data, without being explicitly programmed. It includes a variety of algorithms designed for supervised and unsupervised learning tasks such as regression, classification, and clustering. These models iteratively improve their performance as they are exposed to more data, enabling accurate and scalable predictive analysis (James, Witten, Hastie, & Tibshirani, 2013).

3.2 Unsupervised Learning

Unsupervised learning is a exploratory data analysis category that works with a set of observations as input ($x_1,...,x_n$), but no target values to learn from for the model. The focus is to find the hidden similarity patterns or distinct groups in the input dataset to have better understanding about the data. Main areas for the unsupervised learning are segmentation and noise filtering about the vast datasets. Clustering, association, anomaly detection and dimensionality reduction falls into unsupervised learning methods. The main advantage is the flexibility among big data discoveries due to not needing any labeled data (Hastie, Tibshirani, & Friedman, 2009).

3.2.1 Clustering

Clustering is a unsupervised learning method that is widely used for mostly segmentation projects by finding distinct subgroups in a given dataset $\{x_i\}_{\{i=1..n\}}$ with no labeled target values. The similarity between inputs are defined as intra-cluster similarity which indicates that points that falls into the same cluster are close to each other, and inter-cluster dissimilarity which indicates that points that are not in the same clusters are very distant from each other. Moreover, the point is to minimize distance in the same cluster to define accurate subgroups. The aim is to assign inputs to clusters as $c:\{1,...,n\}\rightarrow\{1,...,K\}$ with the below minimizing cost function where μ_k is the center of cluster.

$$\min_c \sum_{k=1}^{K} \sum_{i:c(i)=k} \|x_i - \mu_k\|^2 \tag{3.1}$$

There are different distance and similarity measures can be used for different cases and the choice has an affect on the overall cluster design and model performance. Most used ones are Euclidean distance, Manhattan distance and Cosine similarity.

Euclidean distance is widely used for spherical clusters;

$$\| x_i - x_j \|_2 \tag{3.2}$$

Manhattan distance is commonly used for one dimension outliers;

$$\| x_i - x_j \|_1 \tag{3.3}$$

Cosine similarity is used for either text or sparses;

$$1 - \frac{x_i^\top x_j}{\| x_i \|_2 \| x_j \|_2} \tag{3.4}$$

Most known clustering partitioning algorithms are K-means algorithm and hierarchical clustering, but day by day new methods are being added to literature as density-based methods, model-based clustering, spectral clustering alternatives being improved. Evaluation of the clustering methods are mostly done by internal scoring systems such as Silhouette score and relative methods such as elbow method also known as knee method (James et al., 2013).

3.2.2.1 K-means

K-means algorithm is can be seen as one of the most used clustering method as stated in the Clustering chapter. It focuses on minimizing the distance by using the dissimilarity measure as Euclidean distance and optimizing the within-cluster sum of squares (WCSS). K-means method is simple and fast compared to other clustering methods on medium sized data, and usually gives clear cluster definitions. But on the downside *K* cluster number must be defined to perform and often converges to local minimum values for large amount of data. The "elbow" method which is plotting WCSS and finding the knee break point for ideal *K* cluster number to perform K-means is widely used and to evaluate often Silhouette score is used (Hastie et al., 2009).

The steps of K-means are as follows:

$$\min_c \sum_{k=1}^{K} \sum_{i:c(i)=k} \| x_i - \mu_k \|^2 \tag{3.5}$$

where the point i cluster assignment,

$$c(i) \in \{1, \ldots, K\} \tag{3.6}$$

$$\mu_k = \frac{1}{\left|\{i : c(i) = k\}\right|} \sum_{i:c(i)=k} x_i \tag{3.7}$$

is the centroid of the cluster k.

3.2.2.2 Fuzzy c-means

Fuzzy C-means is an clustering algorithm that derives from the K-means algorithm. It states that the values can be members of various clusters with some part of similarity degree which is nominated as usually percentage values to more than one cluster that is defined. The membership values can take values from 0 to 1 to assign to the clusters. The aim is to increase flexibility of clustering process and better define features of a given input value when the dataset has very similar members or the uncertainty is high to define a single cluster for a given input value as mentioned. The algorithm can be

helpful with the advanced segmentation needs as the features can vary among complex datasets.

The aim is to minimize the function as follows:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{K} u_{ij}^{m} \cdot \| x_i - c_j \|^2 \tag{3.8}$$

Where;

Dataset:

$$X = \{x_1, x_2, ..., x_N\},\ x_i \in \mathbb{R}^d \tag{3.9}$$

Clusters:

$$C = \{c_1, c_2, ..., c_K\} \tag{3.10}$$

$u_{ij} : x_i$ values' membership degree to cluster

$$c_j , u_{ij} \in [0,1] \tag{3.11}$$

$m > 1$: fuzziness parameter,

$|| xi - cj ||$ **:** distance between the data point and the cluster center

3.2.4 Association rule mining

Association rule mining is a unsupervised data mining method that focuses on finding relationships or dependencies between variables in the dataset. Frequent co-occurence patterns that are found on itemsets, subsequences or substructures are imply the correlations without needing the predefined target variables. Mostly used in market basket analysis problems to find out which products tend to be bought together and to develop cross selling opportunities, recommendation systems to better analyze the customers and structure better marketing strategies, to find out symptoms or patterns in healtcare analytics to develop pharmaceuticals accordingly, financial fraud detection by analyzing purchasing habits to identify dicrepancies. The measurements for the evaluation of the quality occurences are support which is the frequency, confidence which is the probability and lift which is the strength of the correlation between related variables. Apriori and FP-Growth algorithms are well-known pattern

mining algorithms that are widely used in real world applications (Han, Kamber, & Pei, 2011).

3.3 Supervised Learning

Supervised learning is a statistical learning category which is based on primarily a labeled dataset as target values (y_i) and prediction features (x_i) as input values. Main framework focus is a loss function that learns on minimizing the discrepancy between predictions and targets to have better prediction model mappings. The aim is that either accurately predicting the target values or to have better understanding the relationship also known as inference between input and output values which are also known as predictors and response. Both classical and modern methods embrace supervised learning such as regression and classification methods (James et al. 2013).

3.3.1 Classification

Classification is one of the fundamental methods of supervised learning. The main purpose is to set accurate categorical class labels to unkown inputs with the help of learning from the pre defined categorized dataset values. So overall, it is a two step process as the first being building a model deriving from the available training dataset (learning step) and then determining accuracy of the model if it is suitable for predicting classes for new input data values (classification step). A test dataset is used to check accuracy of the proposed model and also because of the classifiers tend to overfit the data, necessary partitions and data preprocessing actions must be done to avoid the overfitting problem. Logistic regression, decision trees, random forest, Bayes classification are the most known examples of the classification algorithms. The use areas are often in audience targeting for marketing purposes, fraud detection for banking applications and disease diagnosis for medical field (Han et al. 2011).

The function that predicts y class tag for the x input:

$$f : R^d \rightarrow \{1,...,K\} \tag{3.12}$$

Where;

The training dataset:

$$\{(x_i, y_i) \mid i = 1, 2, \ldots, n\} \tag{3.13}$$

d sized feature vector:

$$x_i \in R^d \tag{3.14}$$

Class label:

$$y_i \in \{1, 2, \ldots, K\} \tag{3.15}$$

n: example number

d: feature number

K: total class number

3.3.2 Regression

Regression methods are used for the quantitative problems in the supervised learning field of especially in economy, engineering and scientific researches that are conducted with the help of machine learning algorithms. The basic purpose of the regression methods are to predict continuous quantitative target outcome based on the observations in the dataset. In the literature, target is also known as dependent variable and features are mentioned as independent variables. The linear regression is almost the foundational regression type, but there are also regularized regression which is focused on reducing overfitting with Lasso and Ridge methods, non-linear and tree-based regression models focused on complex relationships with large datasets (Hastie et al., 2009).

The linear regression basically has a formula of:

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j \tag{3.15}$$

Where;

the input vector as predictor features which are quantitative values : $X^T = \left(X_1, X_2, ..., X_p\right)$ to predict a continuous output which has real value as Y

The β_0 is a constant term as intercept, when all predictors are zero the predicted value is β_0

The β_j are the unknown coefficients or parameters that needs to be estimated

The p is the total predictors number

3.3.2.1 Extreme gradient boosting

XGBoost (eXtremeGradientBoosting) is a well optimized, high performance ensemble scalable tree boosting system enhancement that is widely used in the machine learning applications such as regression and classification problems, based on the original gradient boosting algorithm. The winning factor of the XGBoost is that it is almost ten times faster than the similar solutions in all scenarios proven at the many challenges among sales prediction, motion detection, categorization, risk prediction machine learning fields. The tree learning algorithm, parallel and distributed learning and out-of-core computing are the features that makes XGBoost considerably faster as mentioned. The idea behind the XGBoost is that sequentially combining weak learners with decision trees and every new tree focuses on correcting the errors that are called residuals of previous trees. The key features are regularization to avoid overfitting, tree pruning to improve scores, handling missing values by automatic learning, parallel computation for speed and customization of loss function to be flexible solution for many different cases (Chen & Guestrin, 2016).

The calculation steps are as follows:

1.Initial prediction is the mean of target y

2.Calculating residuals (errors) as, residual = actual values – predicted values

3.Decision tree fitting to residuals

4. Updating predictions with new tree outputs:

$$\hat{y}_{\text{new}} = \hat{y}_{\text{old}} + \eta \cdot f_m(X) \tag{3.16}$$

Where;

η is the learning rate,

$f_m(X)$ is the m-th tree,

5. Repeating calculations for M rounds which is the number of iterations.

Objective function that XGBoost minimizes:

$$\text{Obj} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{3.17}$$

l is the loss function,

$\Omega(f_k)$ is the regularization for the tree complexity,

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{3.18}$$

T is the leaves number in the tree,

wj is the score assigned to leaf j,

γ and λ controls complexity to avoid overfitting.

3.3.2.2 Light gradient boosting machine

Light Gradient Boosting Machine is a modified boosting algorithm that is derived from the Gradient Boosting Decision Tree (GBDT) algorithm and largely developed by Microsoft with the Distributed Machine Learning Toolkit project started at 2017. The GBDT method combines weak learners to build a better learner model, but due to the regression tree constraint, each tree carries the residuals of the all previous trees. LightGBM solves the accuracy and efficiency problem in high volume data by employing the residual of the previous tree as target of the current tree to form the final prediction. The efficient use of resources such as RAM, ability to handle large datasets and high processing speed makes LightGBM a highly efficent version of boosting algorithm alternatives. The method derives from gradient boosting decision trees, but rather than depth-wise tree growth LightGBM uses leaf-wise tree growth to speed up the learning process. The leaf-wise tree growth also gives better accuracy because of finding better splits overall. The histogram-based split feature offers faster computation and the support for categorical features without the need of one-hot encoding makes LightGBM a versatile tool for prediction applications. The highlights of the method are the gradient based one side sampling which makes focusing only on the important data that gives information gain for the model, and exclusive feature bundling that reduces the variable count by combining similar features to lower the complexity of the dataset to increase efficiency of the proposed prediction model (Ke et al., 2017).

The sample prediction function as follows:

$$F_t(x) = F_{t-1}(x) + \eta \cdot f_t(x) \tag{3.19}$$

Where;

$F_t(x)$ is the prediction after t trees,

$F_{t-1}(x)$ is the previous prediction,

$f_t(x)$ is the new decision tree,

η is the learning rate which controls the update size.

Split Selection with the Gain formula:

When splitting a leaf into child nodes, LightGBM handles the split with the highest gain.

$$\text{Gain} = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] - \gamma \tag{3.20}$$

G_L, G_R are the sum of gradients in nodes,

H_L, H_R are the sum of Hessians (second derivatives) in left/right nodes,

λ is the regularization term for leaf weights,

γ is the cost of creating a new leaf.

3.3.2.3 Random forest

Random Forest algorithm is essentially an ensemble algorithm that combines many decision tree predictions and builds up a learning method with the aim of reducing overfitting risk to improve variance in comparison to singular decision tree learning model. So, the algorithm decorrelates the decision trees with a small configuration change from the bagged trees methods as an improvement. The algorithm can be used in both classification problems by taking the greater vote of all the decision trees, and regression problems by averaging the predictions of the decision trees in the training model. The overfitting is avoided with the features of Random Forest algorithm as bootstrap sampling which means using different and random subsets for training each decision trees and replacing them, also random feature selection at the split points to

node to consider a subset of features. So, the impact of outliers also reduced by reducing the noise by bootstrap sampling and random feature selection as it offers diversity in the means of features, enhancements of random forest algorithm. As random forests take into account less attributes at each split, the algorithm is very effective on high complexity and large databases. Hence, bagging and boosting methods fall behind of the speed of internal estimates of variable importance of random forest algorithm (Han et al., 2011).

The steps of the Random Forest model are as follows:

The dataset which includes N samples with x_i features and the target:

$$y_i: \quad D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N} \tag{3.21}$$

For each b tree the bootstrap sample from the dataset is created: $D_b \sim \text{Bootstrap}(D)$

$h_b(\mathbf{x})$ = prediction of tree b

For Regression problems, average of all the predictions of B trees is defined as solution equation as follows:

$$\hat{y}(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} h_b(\mathbf{x}) \tag{3.22}$$

For Classification problems, the votes of the each class k are counted and the greatest one is picked as solution equation as follows:

$$\hat{y}(\mathbf{x}) = \arg\max_k \sum_{b=1}^{B} \mathbb{I}\left[h_b(\mathbf{x}) = k\right] \tag{3.23}$$

3.3.2.4 Categorical boosting

CatBoost algorithm is an unbiased boosting method based on gradient boosting which is already a powerful machine learning technique with the implementation of categorical features and it is developed by the Yandex Team at the year of 2017 to enhance the gradient boosting with categorical variables. The problem with the gradient boosting algorithms' target leakage as a result of the learning steps which includes boosting targets from the training examples, leads to prediction shift and also categorical features would not be preprocessed correctly due to the leakage issue of

the standard algorithm. The algorithm basically founded on decision trees but rather than ussing asymmetric trees like XGBoost and LightGBM methods, it uses symmetric trees as splitting with same features at every level which helps with processing speed and overall better generalization, and applies gradient boosting method with a few important enhancements with avoiding overfitting focus as an important side benefit. Ordered boosting, also known as ordering principle, solves the issue with derivated gradient boosting algorithm that avoids target leakage problems with using only past observations to create predictions, and a new algorithm is proposed to preprocess categorical features which does not require one-hot encoding of the categorical variables. The automatic processing of categorical variables, endurance to overfitting risk, supporting GPU hardware acceleration to enhance processing speed and easy to config parameter settings make CatBoost an important alternative for machine learning problems. On the other side, due to symmetrical tree system the flexibity of the algorithm is limited and the processing speed on very large datasets can be challenging due to the ordered boosting feature in the calculation steps (Prokhorenkova et al., 2018).

The boosting steps are as follows:

1. Starting prediction which is usually the average of the targets.
2. Residuals (errors) are calculated
3. A new decision tree formed to minimize the residuals with the loss function
4. The result of the decision tree then multiplied with the learning rate and added to predictions.
5. Process continues until the given iteration number which is nominated as m.

The steps of the Categorical Boosting model are as follows:;

$$\hat{y}^{(m)} = \hat{y}^{(m-1)} + \eta \cdot f_m(X) \tag{3.24}$$

$\hat{y}^{(m)}$ is the prediction at the m. İteration,

η is the learning rate which can be changed with parameter optimiztion,

$f_m(X)$ is the tree that is learned by the model at the m. ,teration.

Target function:

$$L = \sum_{i=1}^{n} \ell\left(y_i, F_{m-1}(x_i) + \eta \cdot f_m(x_i)\right) + \Omega(f_m) \tag{3.25}$$

ℓ is the loss function which can be RMSE, Logloss etc.

F_{m-1} is the predictions of the previous model

f_m is the new added decision tree

Ω regularization coefficient as complexity penalty

3.3.2.5 Gradient boosting

Gradient Boosting is a very popular machine learning algorithm acting as solution among both regression and also classification problems. The algorithm is a ensemble model which focuses on weak learners as base learner estimators which are genereally the decision trees, to sequentially combine each weak learners' predictions and adding new models afterwards to reduce residuals or errors of previous predictors and improving robustness to create a strong learner estimator model. As a result with the continuous iteration process the gradient boosting algorithm aims to reduce the overall error of the proposed ensemble model. Also gradient boosting algorithm has the regularization component embedded to avoid or minimize overfitting risk and enhancing generalization feature for the unseen data. The penalization process which is for to gather a balanced outcome between prediction accuracy and model complexity, focuses on overly complex models to reduce their impact on the ensemble model. Morover, the enhanced penalization feature makes the gradient boosting algorithm more accurate for the new data predictions with the iterative, sequentially combining modelling system. The calculation steps starts with building a basic model which predicts the target variable and for regression problems it generally would be the targets average, for classification problems it generally would be the log-odds value is used. Then, the residuals are calculated by computing the difference between real target values and predictions. The residuals then are accounted for the next sequential model as the information that needs to be learned by the next weak learner model. The models' outcomes then taken into account with a learning rate scaling coefficient which lower values indicate slower but more trustable learning, higher values indicate faster but riskier overfitting problems. The iteration continues until the

given iteration number or the targeted residual removing is achived by the gradient boosting model (Hastie et al., 2009).

Initial prediction as follows:

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma) \tag{3.26}$$

For each iteration m;

Compute the negative gradient:

$$r_{im} = -\left[\frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)} \right] \tag{3.27}$$

Fit a new model $h_m(x)$ to r_{im},

Update the model,

$$F_m(x) = F_{m-1}(x) + \eta \cdot h_m(x) \tag{3.28}$$

3.4 Validation Models

3.4.1 Hold-out validation

Hold-out validation method is widely used among machine learning applications due to being simple, easy to use and effective on validation process of machine learning models about regression and classification problems. There are randomly splitted subsets which are defined as training set, validation set and lastly the test set. In many cases, training set also used for validation set to save time as training set is used for fitting the parameters for the model, validation set is to optimize the parameters for the selected models or appliying pruning steps for early stopping and finally the test set used to test the models' performance with the unknown new data as it is critical for real world applications of the proposed model. The general use of training and testing split falls into higher percentage of data usage at training set, generally being 70%-80% at the machine learning researches to make the model learn from the consistent and meaninful amount of data with the large training dataset, but also avoiding overfitting risk by not giving all the dataset to the model. The main advantages of the Hold-out validation method are being efficient about the resource usage due to only one model need to be trained and evaluated, easy to use for varying cases from the

beginner or complex problems regarding machine learning, and ability to work with both small and large datasets without issues with the help of its' basic random splitting technique which randomness can be a disadvantage as it may overestimate or underestimate the performance of the model (James et al., 2013).

3.4.2 Cross-validation

Cross-validation method is a resampling based advanced validation method compared to traditional hold-out method as by the algorithms' framework is more robust with the focus of variance reducing between random splits in the dataset. The most known cross-validation variant is k-fold cross-validation which the dataset split into given *k* sized folds, usually 5 or 10 is chosen for *k* number. The model operates by first get trained on *k-1* folds, then the model is fully tested at the remaining fold until all the *k* fold iterations are completed. The average metric off all folds is the outcome of cross-validation as the final performance metric. Hence, the performance of the validation is more reliable and stable compared to hold-out method in respect of generalization. Also, by applying dataset split with a balanced bias and variance, it reduces the overfitting risk of a single subset of dataset (Kohavi, 1995).

3.5 Hyperparameter Optimization

Hyperparameter optimization focuses on finding the most suitable parameter set for the machine learning models to increase their efficencies about predicting values of the unseen data. Almost all modern machine learning algorithms have configurable set of input parameters such as learning rate, maximum depth or number of trees to optimize models based on the relevancy of given problem, input data structure or desired time duration to process calculations. There are many searching algorithms for finding optimal values; Grid Search which has a defined range for all possible combinations, Random Search for randomization and advanced methods as Bayesian Optimization and Genetic Algorithms are used to increase the accuracy of the overall model performance with intelligent search abilities.

3.5.1 Bayesian optimization with optuna framework

Bayesian Optimization is a hyperparameter optimization method that focuses on advanced probabilistic reasoning to avoid searching unnecessary areas of the search

space by brute force methods. Hence, by avoiding evaluation of the every possible combination and using unnecessary amount of resources, Bayesian Optimization rather works more efficient than other optimization methods. Usually, the optimization model includes a Gaussian Process or Tree-structured Parzen Estimator to find relationships between parameter values and increases overall scores by iteration process of acquisition function that determines the next exploration point in the search region. For neural networks and advanced prediction model optimizations, the benefit of Bayesian based parameter optimization is significant at the model prediction performances (Snoek, Larochelle, & Adams, 2012).

Optuna is a hyperparameter optimization framework which has serious advantages in optimization such as fast computation, easy to setup and versatile architecture and dynamic ability to being deployable for scalable and various purposed machine learning problems. The framework has the main purpose of finding the best parameter set available for any machine learning library or even custom created functions with very little manual intervention. First, the process starts with defining the search space to find the optimal variable value set. Then, the objective function gets the parameter values, trains and evaluates the model with a score that can be accuracy, R-square or any defined performance metric for the model. The optimization strategy of Optuna framework includes a default Bayesian optimization approach called as Tree-structured Parzen Estimator, so Optuna learns from the previous trials and priorities the important areas to work with in terms of combinations. It also has pruning feature which makes it available for early stopping if there is no improvement area to go for in the modeling to save time and resources. The main advantages of the Optuna are being automatic and efficient about finding good parameters for almost every function that are avaliable, pruning support to save time and adaptive search ability to focus on best performing areas to find parameters (Akiba et al., 2019).

3.6 Performance Measures

Model evaluation is critical for assessing performance of the proposed solution to machine learning problems. Both for accuracy for prediction models and also generalization success should be evaluated to suggest which improvements are needed for the proposed models. There are many succesful performance measures which can be seen among various machine learning related researches and papers; the coefficient

of determination (R^2), mean absolute error (MAE), root mean squared error (RMSE) and lastly the symmetric mean absolute percentage error (SMAPE) are used in the related research.

3.6.1 Coefficient of determination (R2)

The coefficient of determination is a widely used measure in regression related machine learning tasks. The score indicates the ability to explain the variances of the dependent variable by independent variables. The aim is to compare the regression models' performance with a base model which both forecasts the target variables' mean value. The score can be between *0* and *1* with *1* being the perfect fit that can explain all the variance. Overall, coefficient of determination is an useful measure for comparing variance explanation ability of models, but can be misleading with large target value variations due to not taking into account error magnitudes.

The calculation formula as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{3.29}$$

Where;

y_i is the actual value,

$\hat{y}_i$ is the predicted value,

$\bar{y}$ is the mean of actual values,

n is the number of observations.

3.6.2 Mean absolute error (MAE)

Mean Absolute Error simply calculates the absolute variance averages between the models' predictions and the actual values. As the calculation method works on averages of absolute differences, the performance metric becomes resistant to outliers

while not considering the direction of the differences. The main use areas are for demand forecasting or allocation of resources problems which has priorization over errors where small errors are more acceptable than the large errors. Moreover, due to the averaging method embedded the metric is falling behind about estimating the outliers impact compared to other averaging methods such as Root Mean Squared Error.

The calculation formula as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.30}$$

Where;

y_i is the actual value,

$\hat{y}_i$ is the predicted value,

$\bar{y}$ is the mean of actual values,

n is the number of observations .

3.6.3 Root mean squared error (RMSE)

Root Mean Squared Error is a heavier deviation penalizing performance metric due to calculation sequence of squaring the errors before taking the average of all errors. On the other side, even a few high deviated outliers can considerably increase the root meean squared error scores and that can lead to misleading evaluation of the model. As large deviations especially not wanted in financial use cases such as profit and loss forecasting, consumption or demand predictions or valuations such as merge and acqusitions; the root mean squared error can be useful to use as comparison metric for model evaluations.

The calculation formula as follows:

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

(3.31)

Where;

y_i is the actual value,

$\hat{y}_i$ is the predicted value,

$\bar{y}$ is the mean of actual values,

n is the number of observations.

3.6.4 Symmetric mean absolute percentage error (SMAPE)

Symmetric Mean Absolute Percentage Error is a enhanced version of the Mean Absolute Percentage Error to avoid asymmetry issues such as where the target values are too small the MAPE can lead to infinity and seem pointless and to make sure of the boundedness as *0%* is the perfect prediction ability and *200%* is the possible maximum error, so interpretion is easier compared to MAPE. The calculation is that average of the absolute value between the actual and the predicted values and expressing the forecast value in terms of percentage form. The scale independent form makes SMAPE comperable with respect to dataset and problem contexts. The use cases of SMAPE for performance evaluation is generally time series forecasting, financial valuations and macroeconomic researches.

The calculation formula as follows:

$$SMAPE = \frac{100}{n} \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{\frac{|y_i| + |\hat{y}_i|}{2}} \tag{3.32}$$

Where;

y_i is the actual value,

$\hat{y}_i$ is the predicted value,

$\bar{y}$ is the mean of actual values,

n is the number of observations.

4. APPLICATION

In this section the dataset introduction, preprocessing steps, model evaluation, selection, optimization and result interpretation are conducted.

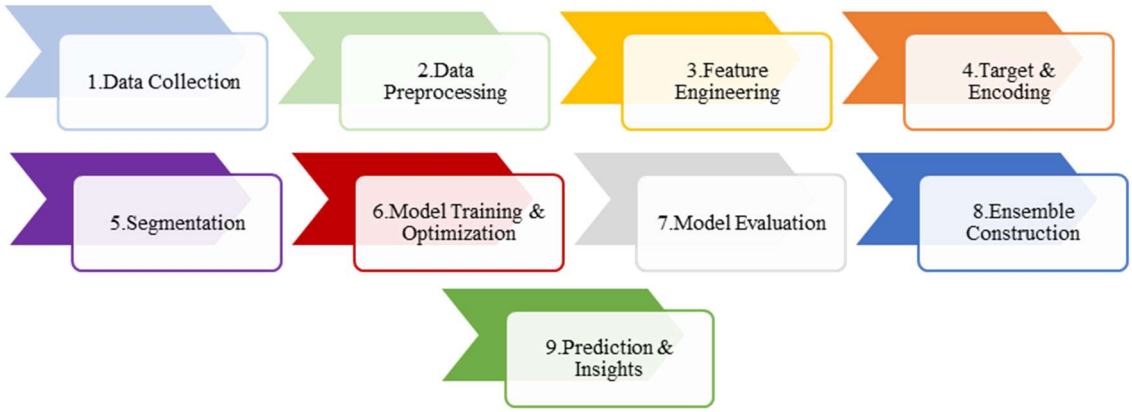

Figure 4.1 : Flowchart of the study.

4.1 Dataset

The particular study is based on real life residential property listings dataset from the various districts which covers almost all of the population of the city of Istanbul, Turkey. The dataset offers detailed specifications of each of the listing records both by categorical and numerical information providing crucial insights about the real estate market of one of the most populated cities in the whole world. The key features of each of the properties includes living area, number of rooms, building's age, floor level, heating type and, other significant attributes. The dataset includes 1508 property listings which have 52 features including categorical and numerical attributes.

Derived from the real data that collected in the April 2025 period, there are new engineered features such as floor ratio, area per room, neighbourhood average that makes the data more suitable for prediction model foundation.

4.2 Data Preprocessing

The preprocessing period started with cleaning and fixing some of the problematic columns due to the dataset being real life data and most of the listings are done

manually by real estate agencies or private property owners there were missing values, wrong entries or format issues in numerical attributes. Also, unreliable or non-informative variables such as listing ID, creation date, cadastral information are removed to enhance data quality by preventing data leakage and providing simplicity. Missing or empty values in listing entries were handles as median values for numeric fields and "Unknown" mark for categorical fields. Label Encoding applied to categorical attributes as to offer compatibility to prediction models by transforming them to numeric indices. Smoothed target encoding employed at district and neighborhood levels to reduce variance between small groups and the global mean of the listings with an empirical Bayes prior ($a=10$) The prediction target which is the listing price defined as natural log of listing price for mitigation of right-skewness and to stabilize variance of error metrics during the model accuracy testing phase.

$$target = \log(1 + price) \tag{4.1}$$

Outlier effects mitigated by unrealistic prices removal of below 100,000 TL and above 20,000,000 TL entries. Standardization process applied to few textual attributes such as floor level, room count, square meter area measurements to increase data handling ability of the proposed prediction models.Conducted outlier analysis for numerical attributes by checking their Z-scores whether their values are higher than "3" profiled as distributional quirks in the features to detect and manage the outliers.

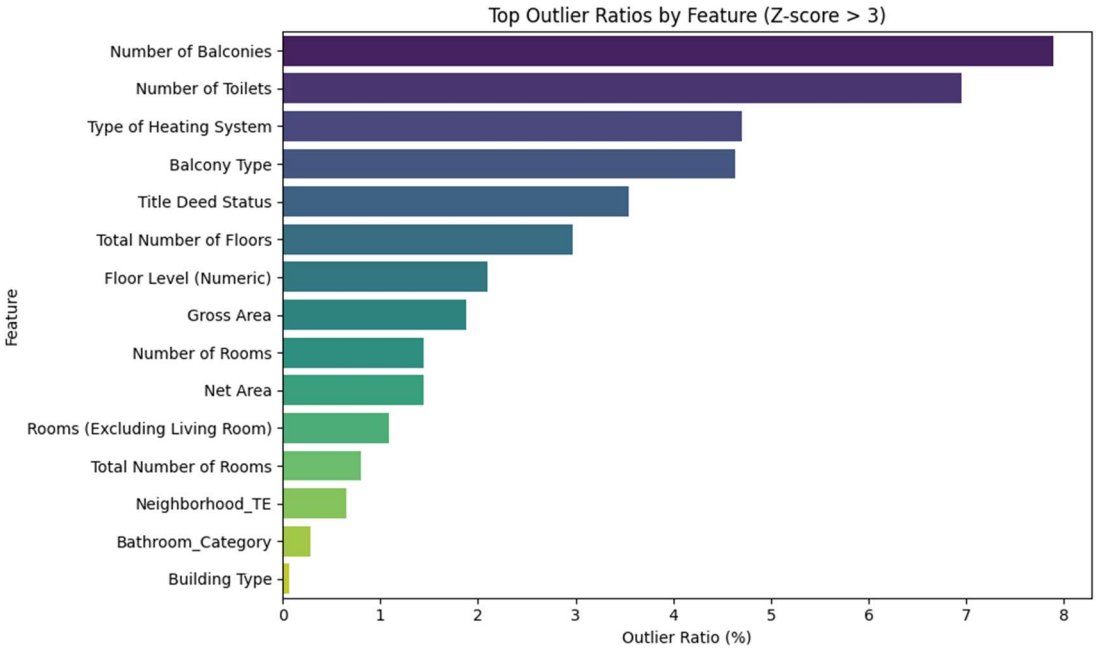

Figure 4.2 : Top Outlier Ratios.

As outlier presence can highly distort the prediction accuracy of the models due to interference with the learning step especially of the regression based prediction models, mitigation or lowering the effects of the outlier values are critical for the sake of success of prediction model abilities. Hence, Winsorization is selected to handle outliers which is a reputable statistical method of reducing outlier effects but also keeping the data points by capping all values beyond the specified threshold limits to nearest boundary values. As the data integrity preserved, also the negative effects of outliers are minimized by handling the extreme values. Especially for regression based prediction models, Winsorization performs better than the Interquartile Range (IQR) method because of data preservation, stabilization of data distribution and increases generalization ability of the regression based models such as XGBoost and Random Forest. Overall, %1 Winsorization applied at 1st and 99th percentiles to handle extreme numerical values for the numerical columns that has the most outlier ratio as the outcome of Z-score analysis.

“Number of Balconies”, “Number of Toilets”, “Net Area”, “Gross Area”, “Total Number of Floors” and “Number of Rooms” features are selected as being the numerical features that have high outlier ratios to apply Winsorization to even out the dataset and increase prediction accuracy. The effects of the outlier handling process with The Winsorization method are graphically shown as box plot visualizations for both before and after Winsorization application for the outlier handling process. So that the data distribution for the related numerical features and the influence of the process can be easily traceble and comparable. Values outside of lower 1% and upper 99% limits are defined as outliers and increpancies treated with capping to nearest boundary values to stabilize the dataset. This approach guarentees the minimization of negative effects of robustness and predictive accuracy of the prediction models.

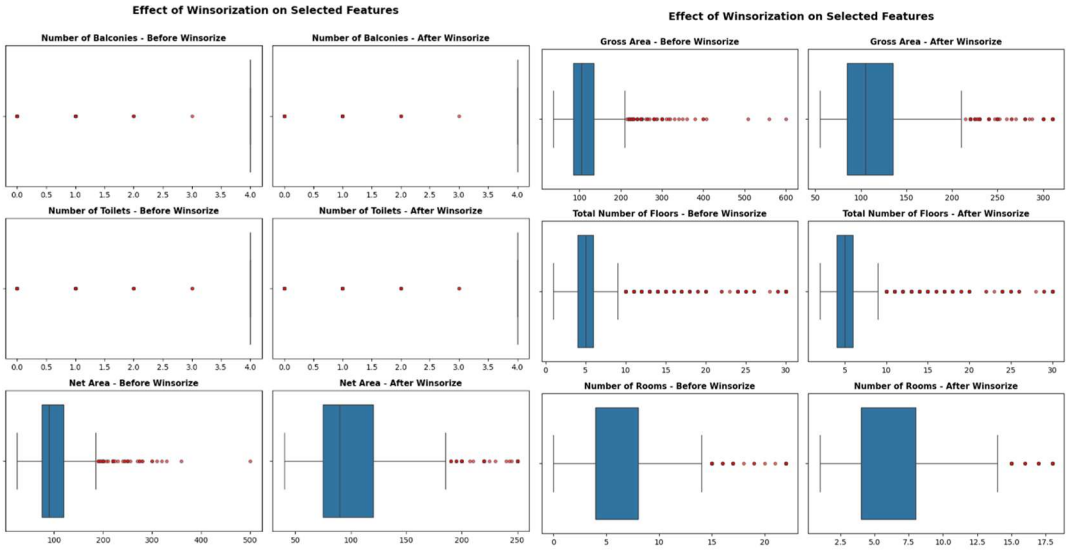

Figure 4.3 : Effect of Winsorization on Selected Features.

Also, new features that imply meaningful relationships such as "floor_ratio" which indicates the ratio of the floor that the property is in and buildings' total floor number, "sqm_per_room" which indicates the net area per room average, "bath_area_ratio" which indicates the bathroom area to total net living area of the property are created with the feature engineering method. Due to structure of the discontinous data distribution nature of "Number of Balconies" and "Number of Toilets" features as it was seen at the Winsorization stage, the outlier ratios remained high for these features.

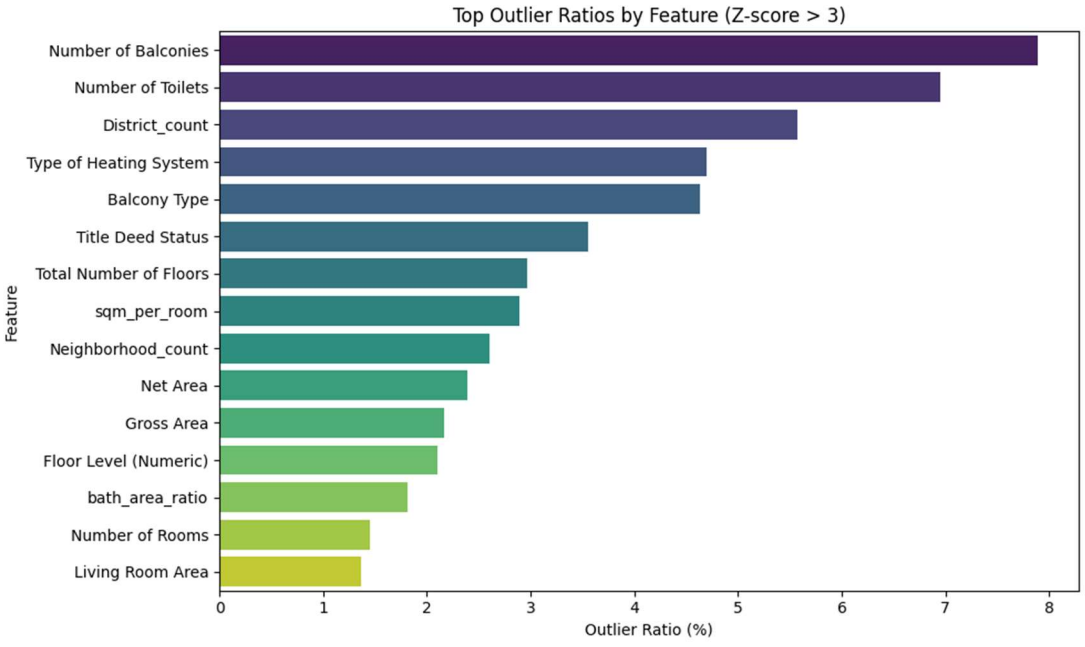

Figure 4.4 : Top Outlier Ratios after Winsorization.

Pearson correlation heatmap drawn after the winsorization for the selected features to better analyze the features relationship with the ultimate target of the research that is accurately predicting the price of the property listing according to given features. The Pearson correlation coefficient (r) is used for measurement of strength and also direction of relationships between features. The range differs from -1 to +1 where -1 means total negative relationship, +1 means perfect positive correlation and 0 implies there are no relationship between the two features subject to coefficient calculation. Color intensity makes easier to see and analyze the relationships between similar features, the darkest tone of red implies +1 value and perfect positive correlation, the darkest tone of blue implies -1 value and perfect negative correlation. As the real estate listing dataset being real and highly detailed, there are many features that can be seen at the Figure 4.4 below with some of them having positive, some of them having negative correlation with each other. Also, it is important to note that few features have the risk of multicollinearity problem due to being almost identical to each other in terms of features. “Net Area”, “Number of Rooms” are easily distinguishable with their positive relationship with the target visually shown on the heatmap. So, it is critical that defining the most fluential features by analyzing the relationship with the “target” variable as the ultimate goal of the study is to accurately predict the price.

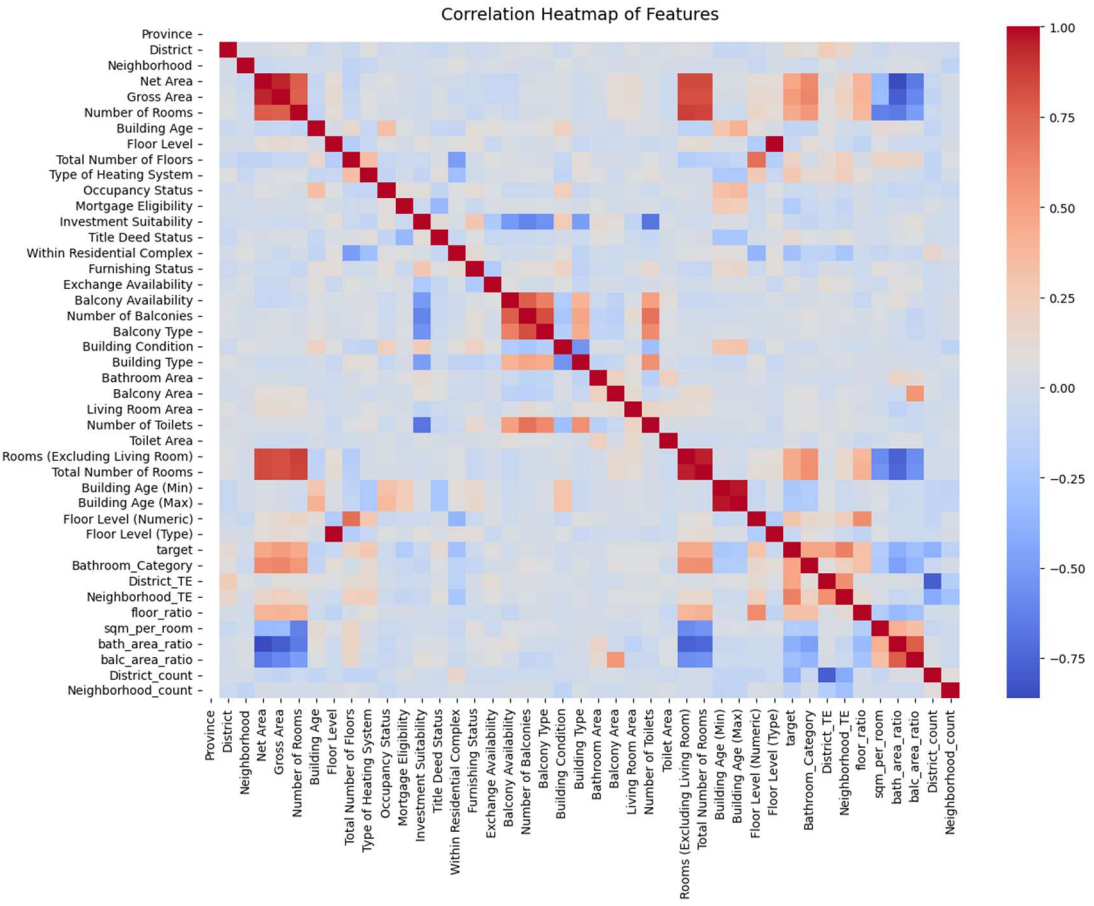

Figure 4.5 : Correlation Heatmap.

As prediction models such as XGBoost regressor automatically focuses on most influential features and discards non meaningful ones, top 15 most related features to target are analyzed with highest Pearson correlation values to target variable. So that a more comprehensive feature analysis could be conducted due to constantrated focus on most important features. The clearer interpretability is achieved with reduced complexity and avoiding overfitting by rermoving unnecessary features from the prediction model dataset with the target-oriented insight provided by the focused correlation matrix.

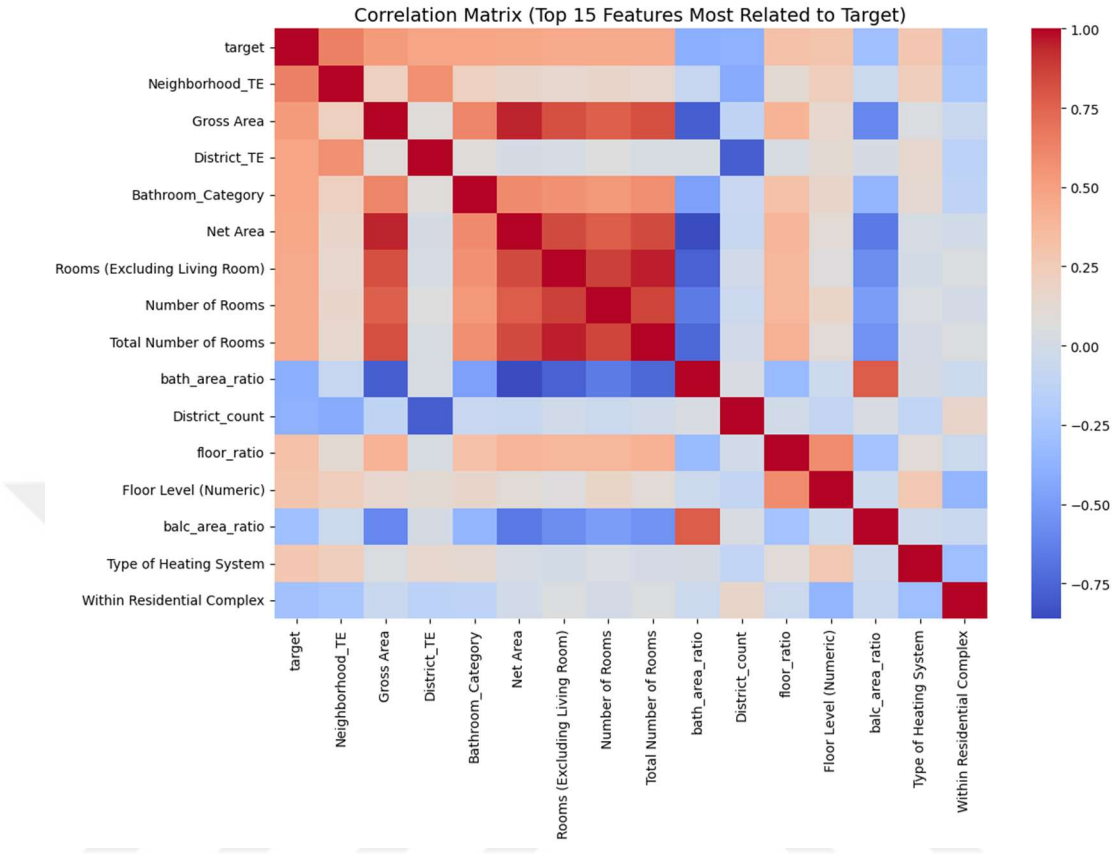

Figure 4.6 : Correlation Matrix of Top 15 Features.

Following the correlation analysis, an XGBoost Feature Importance analysis conducted to identify most influencial features as a second alternative method to better understand the effect of most important features while the prediction model learns the dataset. The process focuses on the information gain concept that starts with evaluation of relative importance which is the point that a feature reducing the prediction error when used to split to data during the tree construction phase of the XGBoost algorithm. Therefore, the feature importance scores indicate quantifically the contribution to XGboost prediction model performance overall. While conducting statistical analysis during the Pearson correlation heatmap stage, also machine learning model interpretation is tested and analyzed with the XGBoost feature importance analysis with advanced explainability of features and easier optimization for next steps of the prediction study.

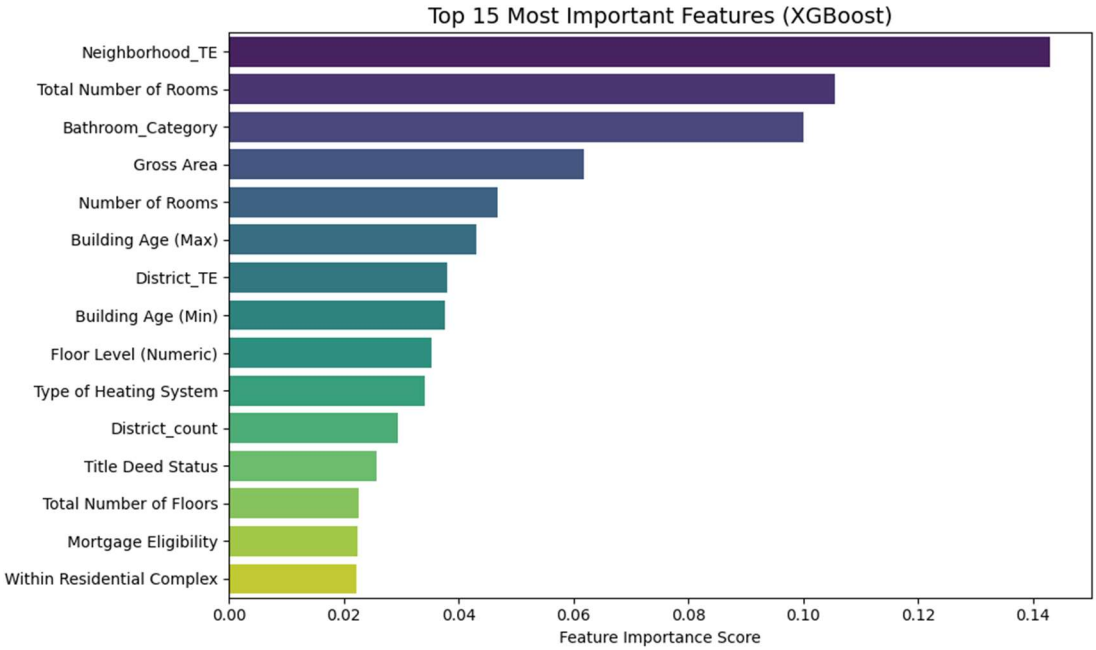

Figure 4.7 : XGBoost Feature Importance Graph.

4.3 Application of the Proposed Models for House Price Prediction

In this study, the models that are XGBoost Regressor, LightGBM Regressor, CatBoost Regressor, Gradient Boosting Regressor and Random Forest Regressor are used. Firstly, the proposed models applied to non-segmented data with default parameters and optimized parameters with Optuna, lastly ensemble model constructed.

Table 4.1 : Performance Analysis of the Base Scenario.

Base	R-Square	MAE	RMSE	SMAPE (%)
XGBoost	0,684	1.355.935	2.220.801	21,4%
CatBoost	0,749	1.196.335	1.979.297	18,4%
LightGBM	0,716	1.285.093	2.103.873	19,7%
GradientBoosting	0,697	1.323.836	2.173.970	20,4%
RandomForest	0,698	1.312.408	2.168.161	20,3%

Base	R-Square	MAE	RMSE	SMAPE (%)
XGBoost-Optuna	0,745	1.215.670	1.993.915	18,8%
CatBoost-Optuna	0,744	1.228.182	1.996.617	19,0%
LightGBM-Optuna	0,754	1.200.584	1.959.589	18,6%
GradientBoosting-Optuna	0,749	1.249.507	1.978.537	19,5%
RandomForest-Optuna	0,683	1.353.136	2.221.721	20,9%

Base	R-Square	MAE	RMSE	SMAPE (%)
Ensemble	0,756	1.189.277	1.948.901	18,4%

Then, automatic segmentation with K-Means algorithm is conducted for 4 segments, then XGBoost Regressor applied to measure performances of the segment predictions by performance metrics selected as Coefficient of Determination (R-Square) , Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Symmetric Mean Absolute Percentage Error (SMAPE) which are all mentioned at the Methodology section of the study. Due to poor results of Segment 0, it is removed from the study. Furthermore, parameter optimization with Optuna algorithm is applied to all models for all segments each to improve singular performances of the prediction models. After optimized results, an ensemble model is developed by the weighted averages of R-Square results of all the models tested for all 3 segments.

This study used holdout method to analyze the performance of prediction models. The data set is divided into 80%-20% training and testing data. The proposed prediction

models and parameter optimization problem are coded using Python language and related packages. All analyzes are performed on a PC with AMD Ryzen 5-1600 processor at 3.2 GHz, 32GB DDR4 RAM and Windows 10 Home. The predicted performance of the proposed pure XGBoost, CatBoost, LightGBM, GradientBoosting and RandomForest models using default parameters for testing data is given below:

Table 4.2 : Performance Analysis of the Proposed Models with Default Parameters.

Segment 1	R-Square	MAE	RMSE	SMAPE (%)
XGBoost	0.672	958,024	1,567,498	19.4%
CatBoost	0.678	923,424	1,554,830	18.6%
LightGBM	0.690	918,843	1,523,965	18.1%
GradientBoosting	0.681	930,989	1,547,445	18.6%
RandomForest	0.609	1,021,383	1,711,081	20.6%

Segment 2	R-Square	MAE	RMSE	SMAPE (%)
XGBoost	0.575	1,618,204	2,605,055	22.2%
CatBoost	0.586	1,599,347	2,569,623	21.3%
LightGBM	0.530	1,739,206	2,739,347	23.2%
GradientBoosting	0.572	1,611,137	2,613,053	21.2%
RandomForest	0.560	1,629,742	2,650,738	21.0%

Segment 3	R-Square	MAE	RMSE	SMAPE (%)
XGBoost	0.728	1,036,028	1,661,432	22.9%
CatBoost	0.839	818,056	1,276,386	18.1%
LightGBM	0.849	846,623	1,235,719	18.0%
GradientBoosting	0.796	958,487	1,437,219	20.7%
RandomForest	0.838	845,611	1,283,349	18.3%

As it can be seen from the results table, the performances of the proposed base models differs between segments. For example, LightGBM model resulted in better results in respect to R-Square error metric for Segment 1 and Segment 3 than other prediction models but in Segment 2 the LightGBM model resulted in lowest prediction performance of the models. Hence, it is critical to understand the dataset specifications and segmentations to prefer prediction model for better price prediction accuracy. However, it is important to note that these results are achieved with default parameters of the given models. It is known that parameter modifications can significantly improve prediction accuracy and overall model performance.

4.3.1 Optimization with optuna

The parameter optimization is critical for better machine learning prediction model performances, and in this study Optuna is used as it is faster than Genetic Algorithm and offers more compatibility for a wide range of prediction models. All models and all segments are throughly optimized with Optuna algorithm giving appropriate value ranges for each of the hyperparameter for each of the prediction models. Parameter optimization details for Segment 1 and all the prediction models are given to provide an example at the Table 4.2.

Table 4.3 : Optuna Hyperparameter Optimization Details for Segment 1.

Segment	Model	Hyperparameter	Value Range	Data Type	Optimum Value
1	CatBoost	depth	[3, 10]	Integer	4.00
1	CatBoost	iterations	[100, 1000]	Integer	597.00
1	CatBoost	l2_leaf_reg	[1, 10]	Real	7.37
1	CatBoost	learning_rate	[0.01, 0.3]	Real (log)	0.11
1	CatBoost	random_strength	[1e-9, 10]	Real (log)	0.01
1	GradientBoosting	learning_rate	[0.01, 0.3]	Real (log)	0.02
1	GradientBoosting	max_depth	[3, 10]	Integer	10.00
1	GradientBoosting	min_samples_leaf	[1, 20]	Integer	15.00
1	GradientBoosting	min_samples_split	[2, 20]	Integer	9.00
1	GradientBoosting	n_estimators	[100, 1000]	Integer	578.00
1	GradientBoosting	subsample	[0.5, 1.0]	Real	0.58
1	LightGBM	colsample_bytree	[0.5, 1.0]	Real	0.52
1	LightGBM	learning_rate	[0.01, 0.3]	Real (log)	0.14
1	LightGBM	max_depth	[3, 10]	Integer	4.00
1	LightGBM	min_child_samples	[5, 100]	Integer	33.00
1	LightGBM	n_estimators	[100, 1000]	Integer	139.00

Table 4.3 (continued) : Optuna Hyperparameter Optimization Details for Segment 1.

Segment	Model	Hyperparameter	Value Range	Data Type	Optimum Value
1	LightGBM	num_leaves	[20, 150]	Integer	100.00
1	LightGBM	reg_alpha	[1e-8, 10]	Real (log)	0.00
1	LightGBM	reg_lambda	[1e-8, 10]	Real (log)	0.00
1	LightGBM	subsample	[0.5, 1.0]	Real	0.66
1	RandomForest	max_depth	[3, 30]	Integer	24.00
1	RandomForest	min_samples_leaf	[1, 20]	Integer	1.00
1	RandomForest	min_samples_split	[2, 20]	Integer	6.00
1	RandomForest	n_estimators	[100, 1000]	Integer	608.00
1	XGBoost	colsample_bytree	[0.5, 1.0]	Real	0.73
1	XGBoost	gamma	[1e-8, 10]	Real (log)	0.00
1	XGBoost	learning_rate	[0.01, 0.3]	Real (log)	0.03
1	XGBoost	max_depth	[3, 10]	Integer	3.00
1	XGBoost	min_child_weight	[1, 10]	Integer	3.00
1	XGBoost	n_estimators	[100, 1000]	Integer	921.00
1	XGBoost	reg_alpha	[1e-8, 10]	Real (log)	0.00

Table 4.4 : Performance Analysis of the Proposed Models with Optimized Parameters.

Segment 1	R-Square	MAE	RMSE	SMAPE (%)
XGBoost-Optuna	0.738	870,074	1,402,819	17.8%
CatBoost-Optuna	0.705	925,406	1,486,904	18.3%
LightGBM-Optuna	0.679	917,104	1,551,930	18.1%
GradientBoosting-Optuna	0.695	912,863	1,510,974	18.0%
RandomForest-Optuna	0.608	1,016,117	1,713,993	20.3%

Segment 2	R-Square	MAE	RMSE	SMAPE (%)
XGBoost-Optuna	0.620	1,522,914	2,461,474	20.8%
CatBoost-Optuna	0.602	1,592,307	2,519,170	21.6%
LightGBM-Optuna	0.615	1,619,667	2,478,925	22.0%
GradientBoosting-Optuna	0.606	1,485,813	2,507,227	20.1%
RandomForest-Optuna	0.558	1,626,216	2,656,045	21.1%

Segment 3	R-Square	MAE	RMSE	SMAPE (%)
XGBoost-Optuna	0.837	801,664	1,287,221	16.7%
CatBoost-Optuna	0.840	804,044	1,273,823	17.7%
LightGBM-Optuna	0.885	742,693	1,082,412	16.5%
GradientBoosting-Optuna	0.841	762,812	1,271,804	16.0%
RandomForest-Optuna	0.826	861,209	1,330,001	18.4%

The results showed significant improvement over the base default parameter results for each of the prediction model with Optuna hyperparamater optimization. On the other hand, the performance differences and rankings of prediction models differ between different segments. Hence, a R-Square weighted ensemble model deployed for a consolidated prediction model for all the segments and resulted in overall robust and steady performance in terms of performance evaluation metrics as R-Square, MAE, RMSE and SMAPE (%).

Table 4.5 : Performance Analysis of the Ensemble Model with Optimized Parameters.

Segment 1	R-Square	MAE	RMSE	SMAPE (%)
Ensemble	0.718	857,251	1,453,567	17.1%

Segment 2	R-Square	MAE	RMSE	SMAPE (%)
Ensemble	0.640	1,434,774	2,397,634	19.2%

Segment 3	R-Square	MAE	RMSE	SMAPE (%)
Ensemble	0.868	685,291	1,156,407	15.1%

Ensemble model with greater average performance in respect to all performance evaluation metrics for all the segments provides accurate and stable prediction performance in this study. Segment 3 consolidated results are given as example at Figure 4.7.

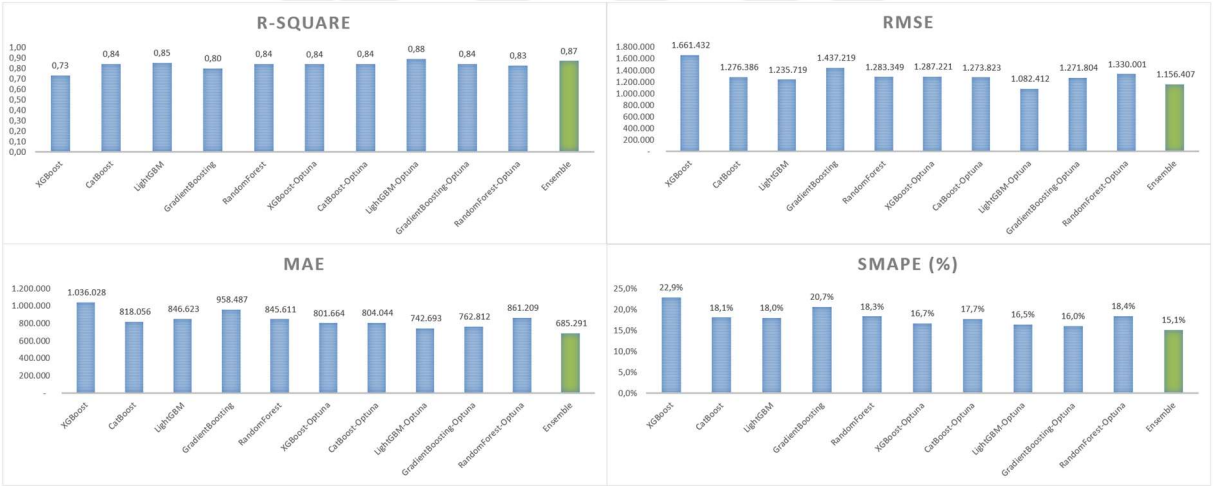

Figure 4.8 : Segment 3 Performance Evaluation Metrics Results.

5.CONCLUSIONS

This research focused on developing a reliable and interpretable house price prediction framework tailored to the Istanbul real estate market. The study combined ensemble learning techniques, Bayesian hyperparameter optimization, and a segment-based modeling strategy to improve predictive performance and interpretability. The dataset captured a wide range of property-level and locational features, including square meter area, number of rooms, building age, floor level, heating system, and neighborhood information.

A detailed data preprocessing and feature engineering pipeline was applied prior to model training. The dataset was refined through cleaning and transformation steps to ensure consistency and eliminate bias. Outliers were handled via the Winsorization approach, which mitigated the effect of extreme values while maintaining distributional integrity. Several engineered variables — such as floor ratio, square meters per room, and bathroom-to-area ratio — were introduced to represent hidden relationships among housing features. The log1p transformation of the target variable stabilized variance, while target encoding for categorical attributes like District and Neighborhood helped prevent data leakage and enhance predictive strength. Correlation-based feature selection was performed to retain the most relevant predictors and reduce redundancy.

Five ensemble-based regression algorithms — Extreme Gradient Boosting (XGBoost), CatBoost, Light Gradient Boosting Machine (LightGBM), Gradient Boosting (GB), and Random Forest (RF) — were initially developed. Each model underwent Bayesian hyperparameter optimization via Optuna, which efficiently explored parameter spaces to identify the optimal configuration for each segment. Model evaluation relied on several

performance metrics, including R^2, MAE, RMSE, and SMAPE, ensuring a balanced view of accuracy and stability.

To better capture heterogeneity in the housing market, the dataset was divided into distinct market segments, allowing models to be trained and evaluated separately for each group. This segmentation revealed meaningful differences in feature importance and predictive behavior across property categories. Based on the results of individual and segment-level analyses, an R^2-weighted ensemble model was constructed by integrating the predictions of optimized base learners. This ensemble achieved the most stable and accurate results, outperforming individual algorithms and reflecting the benefits of model diversity and weighting.

While the results demonstrate strong predictive capability, several directions remain open for future research. Integrating macroeconomic and temporal indicators (such as interest rates, inflation, or construction cost indices) could further enrich predictive insights. Extending segmentation into more granular subgroups — for example, by location clusters or price brackets — may further improve adaptability to market dynamics. Additionally, experimenting with deep learning and spatial econometric frameworks could enhance the model's ability to capture complex spatial and nonlinear relationships.

In summary, this study presents a comprehensive and data-driven methodology that combines advanced feature engineering, Bayesian optimization through Optuna, segment-based modeling, and weighted ensemble learning. The proposed framework offers an effective and adaptable approach for forecasting housing prices in dynamic urban environments like Istanbul, balancing interpretability with high predictive accuracy.

Future works can include variable analysis depending on LLM models and sentiment analysis from the listing descriptions can be conducted to enhance the dataset. Furthermore, due to data being a real time data the limitations about the data quality and missing values for critical inputs can be overcomed with a tailored and purpose focused dataset usage for future studies.

REFERENCES

Abdul-Rahman, S., Ismail, N., & Tan, L. (2021). *Comparative study of LightGBM, XGBoost, and regression models for property valuation in Kuala Lumpur. Asian Journal of Real Estate Research, 9*(2), 113–127.

Adetunji, O., Johnson, D., & Lee, S. (2022). *Random forest regression for housing price prediction: A Boston case study. International Journal of Data Science, 7*(1), 99–112.

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 2623–2631). ACM. https://doi.org/10.1145/3292500.3330701

Anelli, L., Romano, D., & Fabbri, G. (2025). *Evolutionary polynomial and support vector regression for housing price modeling: Evidence from Rome. European Journal of Intelligent Systems, 14*(2), 145–160.

Basysyar, I., & Dwilestari, N. (2022). *Comparative performance of regularized regression models for house price prediction in Ames, USA. Data Analytics Letters, 6*(3), 185–193.

Begum, S., Rahman, M., & Khan, T. (2023). *Machine learning regression models for real estate price prediction: A study from Pennsylvania. Computational Economics Review, 12*(1), 41–52.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

Deaconu, M., Pop, A., & Briciu, V. (2021). *Artificial neural networks and generalized linear models for housing price estimation: A Romanian case study. Romanian Economic Journal, 24*(1), 89–103.

Erkek, M., Aksoy, Y., & Yılmaz, O. (2020). *Predicting real estate prices using support vector machines and neural network approaches: Evidence from major Turkish cities. Journal of Real Estate Studies, 12*(3), 45–58.

Geerts, J., Van der Voort, M., & Langen, T. (2023). *A review of spatial econometric models for housing price estimation. Spatial Analysis and Modeling, 31*(1), 33–49.

Han, J., Kamber, M., & Pei, J. (2011). Data mining: Concepts and techniques (3rd ed.). Morgan Kaufmann.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning: With applications in R. Springer.

Kayakuş, S., Öztürk, F., & Uğur, B. (2022). *Forecasting housing prices in Turkey using tree-based and neural network models. Journal of Economics and Administrative Sciences, 38*(3), 405–420.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., … Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems (Vol. 30). Curran Associates, Inc.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI) (Vol. 2, pp. 1137–1143). Morgan Kaufmann.

Mora-Garcia, S., Gonzalez, R., & Martinez, A. (2022). *Ensemble machine learning approaches for house price prediction in Spain. European Journal of Artificial Intelligence, 5*(2), 202–217.

Mostofi, H., Aydin, R., & Cengiz, O. (2022). *Deep learning approaches for housing price estimation in Trabzon. Turkish Journal of Computer Science, 10*(4), 301–312.

Pastukh, V., & Khomyshyn, Y. (2025). *Advanced ensemble regression methods for real estate valuation in Ukraine. Ukrainian Journal of Data Science, 6*(1), 77–91.*

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. In Advances in

Neural Information Processing Systems (Vol. 31). Curran Associates, Inc.

Satish, A., Kumar, R., & Patel, S. (2019). *Comparative analysis of regression models for house price prediction in King County, USA.* In *Proceedings of the International Conference on Machine Learning Applications* (pp. 210–218). IEEE.

Sharma, P., Singh, R., & Patel, A. (2024). *Comparative evaluation of regression and ensemble models for housing price prediction in Ames City. International Journal of Smart Data Science, 12*(1), 56–74.

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems (Vol. 25). Curran Associates, Inc.

Tandon, R. (2024). *Assessing the performance of ensemble learning models for real estate valuation in Kuala Lumpur. Asian Journal of Computational Economics, 8*(2), 87–99.

Tekin, M., & Sari, D. (2022). *A comparative analysis of regression and boosting models for real estate price prediction: Evidence from Istanbul. Istanbul University Journal of Business, 51*(2), 245–260.

Thamarai, S., & Malarvizhi, R. (2020). *Comparative study of decision tree and multiple regression for housing price prediction in India. International Journal of Advanced Computer Science, 11*(2), 233–240.

Truong, H., Li, X., & Chen, Y. (2020). *Hybrid ensemble regression models for large-scale real estate price prediction in Beijing. Applied Soft Computing, 90*, 106–117.

Wang, J., Lin, C., & Tsai, S. (2021). *Evaluation of regression and boosting models in predicting Taipei housing prices. International Journal of Housing Markets and Analysis, 14*(4), 575–590.

Yağmur, E., Kaya, B., & Gül, F. (2023). *Forecasting Antalya housing prices using ANN, MLR, and SVM models. Journal of Statistical Modeling and Analytics, 17*(2), 199–214.

Zhang, X., Li, J., & Liu, H. (2023). *Integrating machine learning and deep learning for real estate price prediction in Canada. Canadian Journal of AI Research, 9*(3), 120–134.

Zulkifley, A., Rahman, N., & Idris, M. (2020). *Machine learning regression techniques for housing price estimation in Jakarta. Indonesian Journal of Data Science, 8*(1), 54–66.

CURRICULUM VITAE

Name Surname : Artun Pınar

EDUCATION :

- **B.Sc. :** 2022, Istanbul Technical University, Faculty of Management, Management Engineering Department

PROFESSIONAL EXPERIENCE AND REWARDS:

- 2020-2023 Unilever Turkey Finance Analyst
- 2023-2025 Unilever Turkey Assistant Finance Manager
- 2025 Unilever Turkey Finance Manager
- 2025-present GarantiBBVA CIB Strategy, Budget and Performance Supervisor

PUBLICATIONS, PRESENTATIONS AND PATENTS ON THE THESIS:

- **Pınar, A. & Özcan, T. (2025, December 24–25).** *A Novel Approach to House Price Prediction Using Machine Learning Algorithms* [Symposium Presentation]. 3rd International Symposium on Engineering, Design and Innovative Research (ISEDIR 2025), Sinop, Türkiye.