ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

RFMLP BASED CUSTOMER SEGMENTATION AND CUSTOMER CHURN ANALYSIS IN HEAVY EQUIPMENT INDUSTRY USING CUSTOMER TRANSACTIONS DATA

M.Sc. THESIS

Mustafa ÇAMLICA

Department of Industrial Engineering

Industrial Engineering MSc Programme

FEBRUARY 2022

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

RFMLP BASED CUSTOMER SEGMENTATION AND CUSTOMER CHURN ANALYSIS IN HEAVY EQUIPMENT INDUSTRY USING CUSTOMER TRANSACTIONS DATA

M.Sc. THESIS

Mustafa ÇAMLICA
(507181139)

Department of Industrial Engineering

Industrial Engineering MSc Programme

Thesis Advisor: Prof. Dr. Fethi ÇALIŞIR

FEBRUARY 2022

ISTANBUL TEKNİK ÜNİVERSİTESİ ★ LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

İŞ MAKİNESİ SEKTÖRÜNDE MÜŞTERİ İŞLEM VERİLERİNİ KULLANARAK RFMLP TABANLI MÜŞTERİ SEGMENTASYONU VE MÜŞTERİ KAYIP ANALİZİ

YÜKSEK LİSANS TEZİ

Mustafa ÇAMLICA
(507181139)

Endüstri Mühendisliği Bölümü

Endüstri Mühendisliği Yüksek Lisans Programı

Tez Danışmanı: Prof. Dr. Fethi ÇALIŞIR

ŞUBAT 2022

Mustafa ÇAMLICA, an M.Sc. student of ITU Graduate School student ID 507181139, successfully defended the thesis/dissertation entitled “RFMLP BASED CUSTOMER SEGMENTATION AND CUSTOMER CHURN ANALYSIS IN HEAVY EQUIPMENT INDUSTRY USING CUSTOMER TRANSACTION DATA," which he prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor :	**Prof. Dr. Fethi ÇALIŞIR** İstanbul Technical University	.....
Jury Members:	**Assis. Prof. Dr. Ömer Faruk BEYCA** Istanbul Technical University	.....
	Prof. Dr. Selim ZAİM Sabahattin Zaim University	.....

Date of Submission: 14 January 2022
Date of Defense: 04 February 2022

This dissertation is dedicated to my father Ramazan Çamlıca,
my mother Nuray Çamlıca and my wife Melisa Çamlıca

FOREWORD

Firstly, I would like to thank my M.Sc. advisor, Prof. Dr. Fethi Çalışır, for his guidance and support during the preparation of the thesis. I would like to thank Borusan CAT for providing the study's data. Special thanks to the Industrial Engineering Department faculty members that I was fortunate to take courses. I would like to take this opportunity to thank my parents, Nuray and Ramazan Çamlıca, and my wife Melisa Çamlıca, for their belief in me during my M.Sc. studies. I also want to thank my brother and sisters, Ahmet Efe, Asude Rana, Sümeyye, for always thinking highly of me. Lastly I want to thank my grandmother and my grandfather who are passed away, Makbule and Hacı Şerif Çamlıca, for their endless support for whole our family.

January 2022

Mustafa ÇAMLICA
Industrial Engineer

TABLE OF CONTENTS

Page

ABBREVIATIONS

RFMLP : Recency - Frequency - Monetary - Length - Periodicity
RFM : Recency-Frequency-Monetary
ML : Machine Learning
AHP : Analytical Hierarchy Process
DM : Decision Making
MCDM : Multiple Criteria Decision Making
CI : Consistency Index
RI : Random Consistency Index
AI : Artificial Intelligence
KNN : K-Nearest Neighbors
LR : Logistic Regression
SVM : Support Vector Machines
RF : Random Forest
TP : True Positive
TN : True Negative
FP : False Positive
FN : False Negative
AUC : Area Under Curve
ROC : Receiver Operating Characteristics

LIST OF TABLES

Page

LIST OF FIGURES

Page

RFMLP BASED CUSTOMER SEGMENTATION AND CUSTOMER CHURN ANALYSIS IN HEAVY EQUIPMENT INDUSTRY USING CUSTOMER TRANSACTIONS DATA

SUMMARY

With the enhancement in technology and information systems, tons of data have been created and will be created in the future. In today's world, businesses must utilize from data to survive in a challenging environment. For long-term success and to make a better decision, companies are trying to use data. Correct and efficient usage of data is inevitable for companies. Amount of data that is created by companies' actions is getting higher day by day, this situation seems to be good however it has some disadvantages such as the complexity of getting meaningful insights from data.

Understanding and identifying the customer seems key to success in the changing and challenging market places. Companies should know and recognize their customers to become successful. Massive amount of data of customers stored in databases such as customer information data, transaction data, interaction data, behavioral data, attitudinal data etc. Business experts need to find ways to extract meaningful insights from such diversified data. The complexity of extracting valuable information from data gets challenging when amount and dimensionality of data is increased. That's why business experts should use analytical and up-to-date methods. To become successful in the long run, companies should seek ways to increase their financial incomes, enhance their customer service levels and invest in finding ways of improved customer experience. In today's business environment companies realized that their most important asset is their customers, therefore they need to invest to understand their customers and their requirements. Customer segmentation gives opportunities to companies to identify their customers, understand their requirements, and recognize the most important customers. Thus companies should invest to create continuous customer segmentation stream.

Companies should need to find and contract with new customers in order to maintain business. Finding new customers is important, but customer retention is also critical. Creating long-term contracts with profitable customers is inevitable to increase profits. Customers who are in the long run with the company create more profits and they are likely to be less sensitive to chalenging market conditions and they can provide positive referrals. Attracting new customers is very hard and more costly than preventing the current customer from churning. Historical customer information data is very important for companies to combat with customer churn. Companies should detect future churners by using predictive models to survive competitive markets. Prediction of churn can be completed by observing customer behavior or individual behaviors that have signs of attrition. Churn analysis methods allow companies to take some actions against customers at risk. Different customers need different preferences; thus, proactively communicating with them is critical to retain them in the customer list. Churn analysis projects are completed in many different industries such as digital markets, communication industries, banking, insurance, retail stores, online platforms

and automative industries. Usage of mathematical modelling and machine learning techniques is unavoidable during prediction process. In order to have long term success companies need to find ways to prevent their customers from churning.

Machine learning is a sub-field of artificial intelligence that gives learning and improvement opportunities from past experiences to systems and computers. Machine learning generally focuses on developing computer programs that can access data and learn from data without human assistance. By using machine learning one can predict a future outcome or uncover patterns that are hard to spot by humans. Users apply machine learning algorithms to uncover meaningful insights, find relations and predict future results of events. Supervised learning is one of the sub-fields of machine learning that can handle and use labeled datasets which is a dataset that has known values for each record of the target variable. Classification and regression models can be represented as categories of supervised learning. Unsupervised learning is another sub-field of machine learning that is used when it comes to an unlabeled dataset or in other words there are no predefined outcomes in unsupervised learning algorithms. Clustering, anomaly detection, association mining and latent variable models are categories of unsupervised learning.

With the industrialization of construction work, the need for heavy equipment is increased for both efficiency and productivity. The market size of the machinery industry is also growing. There are many sectores in which heavy equipment is used such as infrastructure, mining, agriculture, forestry. Turkey has also an active market in the heavy equipment industry and has very challenging conditions for competition. Therefore in Turkey's heavy equipment market, companies should complete projects in order to understand their customers and prevent them from churning.

The main target of this study is creating customer segments with customer transaction data of one of the leading heavy equipment industry companies, Borusan-CAT operating in Turkey which is a solution partner of Caterpillar Inc., and assigning churn probabilities to each customer. Customer transaction data is collected for the 2018 – 2020 period from complex database of the company. Data pre-processing step is completed in order to use raw data in this study. To decide the importance of the variables and customer segments Analytical Hierarchy Process was used 5 managers of the company respond a questionnare. After deciding the weight importance of the variables customer segmentation was completed with one of the unsupervised machine learning algorithms known as k-means clustering. 4 different customer segments were created. The importance of each customer segment was calculated with the help of weights that is result of Analytical Hierarchy Method. After customer segmentation, customer churn analysis was conducted. Churn analysis was completed with the help of supervised machine learning algorithms such as Logistic Regression, Support Vector Machines, Random Forest, and k-Nearest Neighbors. By comparing performance of each algorithm with the others, Random Forest was found as the most successful algorithm with highest accuracy rate in this study when it comes to predicting the customers who will churn in upcoming periods. There are no violations of the assumptions of each algorithm, therefore each of them can be used in this study. With customer churn analysis, each customer in the dataset labeled as churners or non-churners. Companies can use this information in order to complete such projects to prevent possible churners from churning in the future.

Contributions of this study can be said as applying RFMLP based customer segmentation with a time-effective and efficient machine learning algorithm and applying customer churn analysis with the help of supervised machine learning algorithms to the customer transaction data of one of the biggest heavy equipment companies in Turkey. With this study 4 different customer segments are created and customer churn prediction is completed with high accuracy. Companies in the heavy equipment industry can utilize from this study to identify different customer groups and profile them, they manage their CRM and marketing strategies and allocation of resources can be completed with high effectiveness.

İŞ MAKİNESİ SEKTÖRÜNDE MÜŞTERİ İŞLEM VERİLERİNİ KULLANARAK RFMLP TABANLI MÜŞTERİ SEGMENTASYONU VE MÜŞTERİ KAYIP ANALİZİ

ÖZET

Teknoloji ve bilgi sistemlerindeki gelişmelerle birlikte tonlarca veri oluşturuldu ve gelecekte de oluşturulmaya devam edecek. Günümüz dünyasında işletmeler zorlayıcı iş dünyasında bir ayakta kalabilmek için verilerden yararlanmak zorundadır. Şirketler uzun vadede daha başarılı olmak ve daha iyi kararlar verebilmek adına veriden yararlanmaya çalışmaktadır. Verinin doğru ve verimli kullanılması şirketler için kaçınılmazdır. Şirketlerin aldığı aksiyonlar ve günlük operasyonları sonucu üretilen veri günden güne artmaktadır, bu durum oldukça iyi görünse de bu veriden anlamlı içgörü elde etmenin karmaşıklığı gibi dezavantajları da vardır.

Değişen ve zorlayıcı pazar koşullarında müşteriyi tanımak ve anlamak başarının anahtarı gibi görünüyor. Şirketler başarılı olmak için müşterilerini tanımalı ve onları anlamalıdırlar. Şirketlerin veritabanlarında depolanan oldukça büyük miktarda veri mevcuttur bunlara örnek olarak müşteri bilgi verisi, etkileşim verisi, işlem verisi, davranış verisi ve tutum verisi verilebilir. İş uzmanlarının bu derece çeşitlendirilmiş veriden anlamlı içgörüler elde etmenin yollarını bulması gerekmektedir. Verinin miktarı ve boyutu arttığında, veriden anlamlı bilgileri elde etmenin karmaşıklığı da artmaktadır. Bu nedenle iş uzmanları analitik ve güncel yöntemler kullanmalıdırlar. Uzun vadede başarılı olmak için şirketler finansal gelirlerini artırmanın yollarını aramalı, müşteri hizmet seviyelerini geliştirmeli ve daha iyi müşteri deneyimi sağlamanın yöntemlerini bulmaya yatırım yapmalıdırlar. Günümüz iş ortamında şirketler en önemli varlıklarının müşterileri olduğunu anladılar ve bu nedenle müşterileri ve onların isteklerini anlamak için yatırım yapmalıdırlar. Müşteri segmentasyonu şirketlere müşterilerini tanıma, onların gereksinimlerini anlama ve en önemli olan müşterileri belirleme fırsatı vermektedir. Bu nedenle şirketler müşteri segmentasyonu akışı kurgulamak için çeşitli yatırımlar yapmalıdırlar.

Şirketler, işlerini sürdürmek için yeni müşteriler bulmalı ve onlarla sözleşme yapmalıdır. Yeni müşteriler bulmak önemlidir, ancak mevcut müşterileri elde tutmak da oldukça önemlidir. Kârlı müşterilerle uzun vadeli sözleşmeler yapmak, kârı artırmak için kaçınılmazdır. Şirketle uzun vadeli ilişki içinde olan müşteriler daha fazla kâr sağlar, zorlu piyasa koşullarına karşı daha az duyarlı olur ve şirket için olumlu yönlendirmelerde bulunabilirler. Yeni müşteri bulmak mevcut müşterinin kaybolmasını önlemekten çok daha zorlayıcı ve maliyetlidir. Geçmiş müşteri bilgileri verisi, şirketlerin müşteri kaybıyla mücadele etmesinde oldukça yüksek bir öneme sahiptir. Şirketler, rekabetçi pazarlarda ayakta kalabilmek için tahmine dayalı modeller kullanarak gelecekteki müşteri kayıplarını tespit etmelidir. Müşteri kayıpları müşteri davranışları gözlemlenerek tahminlenebilmektedir. Müşteri kayıp analizi yöntemleri, şirketlerin risk altındaki müşterilere karşı bir takım önlemler almasına olanak tanıyabilmektedir. Farklı müşterilerin, farklı tercih ve öncelikleri vardır, bu nedenle onlarla proaktif bir yaklaşımla iletişim kurmak, onları elde tutmak için oldukça önemlidir. Dijital pazarlar, iletişim sektörleri, perakende satış ve otomotiv endüstrisi gibi birçok farklı sektörde müşteri kayıp analizi projeleri yapılabilmektedir. Kayıp analizinde makine öğrenmesi ve matematiksel modelleme teknikleri kullanılmaktadır.

Makine öğrenmesi yapay zekanın bir alt dalıdır ayrıca sistemlere ve bilgisayarlara geçmiş deneyimlerden öğrenme ve geliştirme fırsatı vermektedir. Makine öğrenmesi kendi kendine veriye erişebilen ve insan yardımı olmadan veriden öğrenebilen bilgisayar programları geliştirmeye odaklanmaktadır. Makine öğrenmesi ile gelecekteki bir olayın sonucu tahmin edilebilir veya insanlar tarafından tespit edilmesi zor olan kalıpları ortaya çıkarılabilir. Kullanıcılar, anlamlı içgörü elde etmek, olaylar arası ilişkileri bulmak ve olayların gelecekteki sonuçlarını tahmin etmek için makine öğrenmesinden faydalanmaktadırlar. Gözetimli öğrenme, makine öğrenmesinin bir alt dalıdır ve her bir kaydı için hedef değişkeninde bilinen bir değer olan verisetini işleyebilmektedir. Sınıflandırma ve regresyon modelleri gözetimli öğrenmeye örnek olarak verilebilir. Gözetimsiz öğrenme ise hedef değişkeni tanımlı olmayan verisetlerinin kullanıldığı makine öğrenmesinin başka bir alt dalıdır. Gözetimsiz öğrenmeden önceden tanımlanmış bir çıktı bulunmamaktadır. Kümeleme, anomali tespiti, birliktelik analizi ve gizli değişken modelleri gözetimsiz öğrenme modellerine örnek olarak verilebilir.

İnşaat işlerinin sanayileşmesi ile birlikte hem verimlilik hem de üretkenlik açısından iş makinelerine olan ihtiyaç artmaktadır. İş makinesi sektörünün pazar büyüklüğü günden güne artmaktadır. Altyapı, madencilik, ormancılık ve tarım gibi iş makinelerinin kullanıldığı bir çok sektör bulunmaktadır. Türkiye, iş makineleri sektöründe aktif bir pazara ve rekabet açısından oldukça zorlu koşullara sahiptir. Bu nedenle Türkiye iş makineleri pazarında bulunan firmalar, müşterilerini anlamak ve müşteri kayıplarını azaltmak adına projeler tamamlamalıdır.

Bu çalışmanın ana hedefi, Türkiye'de faaliyet gösteren iş makinesi sektörünün önde gelen firmalarından Borusan-CAT'in müşteri işlem verisini kullanarak müşteri segmentleri oluşturmak ve her bir müşteriye kayıp olasılıkları atamak olarak özetlenebilir. Bu bağlamda şirketin veritabanından 2018 – 2020 dönemi için müşteri işlem verisi çekildi ardından ham verinin çalışmada kullanılabilmesi adına veri ön işleme aşaması tamamlanmıştır. Değişkenlerin ve müşteri segmentlerinin önemine karar verebilmek adına şirketin 5 yöneticisinin katılımıyla Analitik Hiyerarşi Süreci'nden yararlanılmıştır. Değişkenlerin önemine karar verildikten sonra, k-ortalama kümeleme olarak bilinen gözetimsiz makine öğrenmesi algoritması ile müşteri segmentasyonu tamamlanmıştır. Bu kapsam 4 farklı müşteri segmenti oluşturulmuştur. Analitik Hiyerarşi Yöntemi ile elde edilen ağırlıkların kullanılması ile her bir müşteri segmentinin önemi hesaplanmıştır. Müşteri segmentasyonun ardından müşteri kayıp analizi tamamlanmıştır. Müşteri kayıp analizi lojistik regresyon, destek vektör makineleri, k en yakın komşu ve rastgele orman olarak bilinen gözetimli öğrenme teknikleriyle tamamlanmıştır. Her bşr algoritmanın performansı diğerleri ile karşılaştırıldığında, bu çalışmada gelecekteki müşteri kayıplarını tahmin etmek konusunda Rastgele Orman algoritması en başarılı ve en yüksek doğruluk oranına sahip olan algoritma olarak belirlenmiştir. Algoritmaların varsayımları kontrol edildiğine herhangi bir ihlal bulunmamaktadır, böylece her bir algoritma bu çalışmada kullanılabilir. Müşteri kayıp analizi ile veri setinde bulunan her bir müşteri gelecekte kaybedilecek veya kaybedilmeyecek olacak şekilde etiketlenmiştir. Şirketler, bu bilgileri gelecekteki olası müşteri kayıplarını engellemek adına çeşitli projelerde kullanabilirler.

Bu çalışmanın katkıları, Türkiye'de faaliyetlerine devam eden bir şirketin müşteri işlem verisini kullanarak, zaman-etkin ve verimli bir makine öğrenmesi algoritması ile RFMLP tabanlı müşteri segmentasyonunun uygulanması ve müşteri kayıp analizinin gözetimli makine öğrenmesi algoritmaları ile uygulanması olarak özetlenebilir. Bu çalışma ile 4 farklı müşteri segmenti oluşturulmuş ve müşteri kayıp analizi yüksek doğruluk oranıyla hesaplanmıştır. İş makinesi sektöründeki firmalar bu çalışmadan farklı müşteri gruplarını belirlemek, müşteri profilleri oluşturmak , müşteri ilişkileri yönetimi ve pazarlama stratejilerini yönetebilmek ve kaynak tahsisini yüksek verimlilikle tamamlamak adına faydalanabilir.

1. INTRODUCTION

Data has been used for decades. With the help of technological advancements, billions of data have been created and will be created in upcoming years. However, correct usage of data is inevitable for companies for succeed. However, for long-term achievement and to make better decisions, companies are trying to use the insights of data day by day. According to International Data Corporation, the total amount of data in 2018 were 33 zettabytes, and by the end of 2025, it will be about 175 zettabytes [1]. This situation seems to be good. However, it also has some disadvantages, such as the complexity of getting meaningful insights from such an amount of data.

In the rapidly changing marketplaces, understanding and identifying the customer is the key to the business's success. Thus, companies and business experts should know and recognize the customer for long-term achievement. There can be several ways to identify customers; however, in today's technological development, a massive amount of data of customers are stored in databases. For example, business experts need to know how to extract valuable information from customer data such as basic data, transaction data, interaction data, behavioral data, attitudinal data, etc. The complexity of extracting valuable information from different kinds of data increases when the amount of data increases. That's why business professionals and consultants need to use analytical methods.

To increase financial incomes, the level of investment is growing exponentially when it comes to customer experience investment [2]. Besides, as reported by IDC, customer experience investments will reach $641 billion in 2022 [3]. Thus, companies should invest and find effective ways to increase customer experience and understanding. In the competitive and complicated business environment creating different segments of customers enhances customer loyalty [4].

With the technological advancements and needs of the era, companies should seek new ways to become more innovative. This can also be explained as a transformation of both companies and sectors. Due to globalization and dynamism, the heavy equipment industry is one of the most affected industries with the transformation. Companies in

this industry should find new ways of doing business and identify their customers. One of the most important ways of knowing each customer is simply segmenting them. Customer segmentations have a positive effect on companies' profit [5]. In today's challenging business environment, the customer is the most crucial asset for companies, and segmenting customers is a strategic tool to understand customers' requirements. It is also important to recognize the most valuable customers [6]. Thus, regardless of industry, segmentation is crucial for companies, and companies in the heavy equipment industry should invest in customer segmentation for long-term success.

Although finding new customers is important, customer retention has higher importance [5]. Companies need to have a long-term relationship with their profitable customers to have a successful business and increase profits. In addition, historical customer information is a powerful asset for companies to combat customer churn [7]. It is very challenging to attract new customers and nearly six times harder than prevent current customers from churning; detecting the customers who will be churners in the future is also very difficult; companies need to use and trust predictive churn models to be able to survive in the competitive market [8].

Aside from conventional methods, machine learning techniques have been successfully used in many industries; machine learning algorithms target to optimize performance metrics that evaluate task performance in a given task [9]. There are many ways to explain the classification of ML algorithms; they can be classified as supervised and unsupervised. Supervised learning algorithms use labeled data to regress or classify; unsupervised learning algorithms are generally used to demonstrate unknown patterns [10]. While trying to solve customer churn analysis, Supervised ML algorithms can be used because of historical churn data.

1.1 Purpose of Thesis

This study aims to create customer segments and assign churn probabilities to each customer. Customer transaction data has been collected to use in this study. Customer segmentation has been completed by expert opinion previously. However, it is very challenging when a company has too many customers. With the help of this study, customers will be segmented with the analytical and time-effective methods. With the historical transactional data of the company, customers are segmented by using a

supervised machine learning algorithm. After segmentation, each customers' churn probability is assigned. Contributions of this study to literature can be shown as below:

- Applying RFMLP based analytical methods to create customer segments.
- Using time-effective machine learning models to create customer segments.
- Applying churn analysis methods with the help of Supervised ML algorithms to the customer transaction data of one of the biggest heavy equipment companies in Turkey.

2. LITERATURE REVIEW

2.1 Heavy Equipment Industry

With the industrialization of construction work, the need for heavy equipment is increased for both efficiency and productivity; in the last decades, heavy equipment has been used in horizontal construction such as excavation and vertical construction such as positioning of materials [11]. As a result, the market size of machinery is huge and highly correlated with the construction sector. Also, the volume of imported and used machines is highly correlated with the size of the construction industry [12]. In Figure 2.1, the global market size of heavy industry was 139.85 billion $ and is predicted as 175.57 billion $ in 2025 [13].

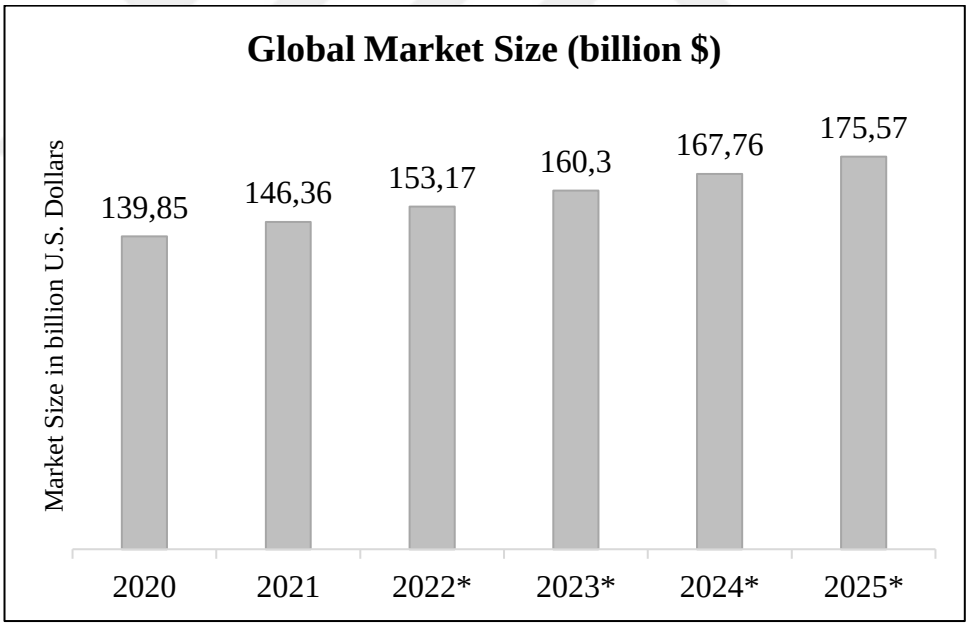

Figure 2.1: Yearly global market size of heavy equipment and estimations [13].

There are many sectors in which heavy equipment is used, such as infrastructure, mining, agriculture, forestry, etc. Cranes, bulldozers, backhoe loaders, excavators are examples of heavy equipment used in sectors [14]. In table 2.1, machinery types and descriptions can be seen.

Table 2.1: Machinery types and descriptions [12].

Machinery Types	Description
Crawler Excavator	Excavation and loading
Wheeled Excavator	Similar to crawler excavator but lighter
Wheeled Backhoe Excavator	Excavation and loading
Crawler Loader	Excavation and loading
Wheeled Loader	Similar to crawler loader but lighter
Dozer	Pushing the material
Grader	Finishing and shaping
Compactor	Compacting the material

Global leading producers of heavy equipment and their sales can be seen as of April 2021 in Figure 2.2, the leading company is Hitachi (Japan), and Caterpillar (U.S.) is the second company according to their sales [14].

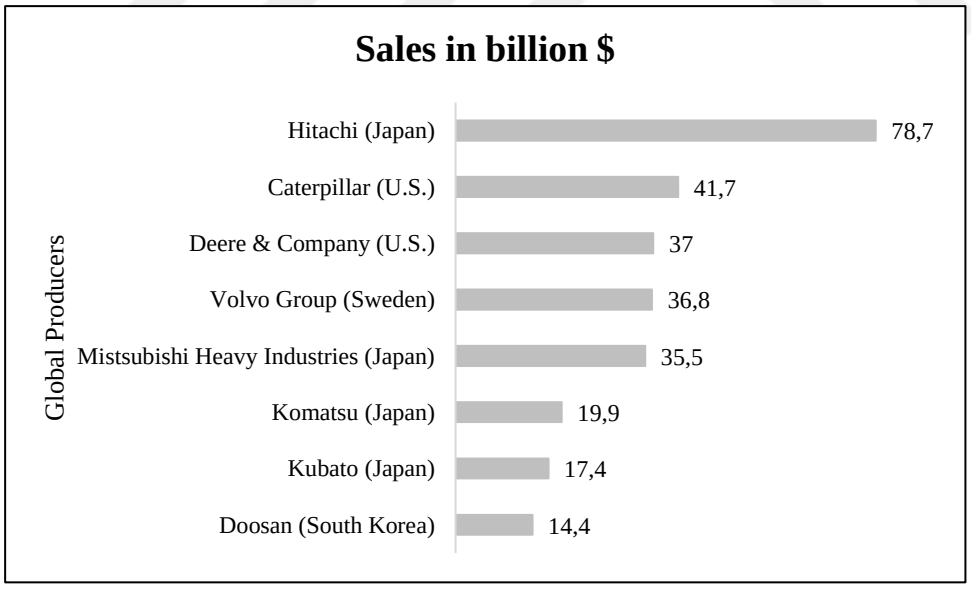

Figure 2.2: Global producers and sales as of April 2021 [14].

Turkey is also an active market for the heavy equipment industry. There are 131.000 employees in the industry. The total market size was also about 6,5 billion $ in Turkey, and 1,1 billion $ worth of import actions were completed. According to sales size,

Turkey is the 4th country in Europe. Borusan, Çukurova, Hidromek, Sanko, and Sabancı are Turkey's biggest active heavy equipment firms [15].

2.2 Customer Segmentation

Customer segmentation can be defined as simply dividing customers into subsets according to their characteristics; with the segmentation, customers in the same subset will have similarities, and customers in a different subset will have dissimilarities [16]. The segmentation concept was introduced in 1956, proving that the whole market is diversified with internal similarities. To achieve higher customer requirements, customer segmentation is necessary because customers are in different segments have different preferences [17].

Companies in diversified industries realized that the most crucial asset of them is their customers, how to learn and understand the needs of customers is key for long term success and is a key for surviving in a competitive market, answering the needs of the customer in the correct time is also another key for creating long term interconnection between firms and customers. Therefore companies are paying extra attention to attracting and retaining valuable customers by identifying and segmenting customers [18]. To increase profit, enterprises should select their valuable customers rather than working with all customers equally; in the long term, some customers can be less profitable and have big problems to work with, segmenting customers into different customer groups can help firms to better understand customer needs, with this approach various marketing tactics can be applied to varied customer groups [18].

There are many techniques and variables proposed for segmenting customers; clustering, classification, self-organizing maps, and neural networks are some techniques that are widely used in literature. For example, Recency-frequency-monetary (RFM) is a widely used variable among the techniques and variables, and clustering is the most used technique [19].

Firms can utilize customer segmentation to optimize resource allocation, create and manage marketing campaigns, and give correct service to customers who are in specific customer segment; previous research have used diverse techniques and methods that can be seen in Table 2.2, RFM and expanded types of RFM is a widely used technique in the literature [20].

Table 2.1: Used segmentation techniques in the literature [20].

Applied Area(s)	Description: proposed method
Customer-centric industries [21]	Constrained Laplacian Rank (CLR) method and extensions
Robust and sparse fuzzy k-means clustering [22]	Sparse fuzzy K-Means clustering algorithm.
Clustering and projected clustering with adaptive neighbors [23]	CAN clustering algorithm with adaptive neighbors
Forecasting supply chain demand [24]	K-means
A public sector hospital in Tehran [25]	K-means
Japanese grocery stores [26]	K-medoids
Customer pain points [27]	Biclustering-based method
Online customized fashion business [28]	K-Medoids and CN2-SD algorithm
Mobile service customers g [29]	Extended fuzzy C-means
Load estimation for microgrid planning [30]	Self-Organizing Map (SOM)
Customer grouping for better resources allocation using GA based clustering technique [31]	Robust genetic algorithm (GA) based k-means
Segmenting customers in online stores [32]	SOM and K-means
Identifying next relevant variables for segmentation [33]	SOM and K-means
Textile manufacturing business [34]	WARD + K-means
Cinema consumers segmentation based on values and lifestyle [35]	WARD + K-means
An application of a metaheuristic algorithm-based clustering ensemble method to APP customer segmentation [36]	Real-coded genetic algorithm-based K-means clustering ensembles (GKCE), particle swarm optimization-based K-means clustering ensembles (PSOKCE), and artificial bee colony-based K-means clustering ensembles (ABCKCE).
Techniques for profiling profitable hotel customers [37]	SOM, K-means, and RFM
Children's dental clinic [38]	SOM and LRFM
Hairdressing industry [39]	K-means and RFM
Classifying the segmentation of customer value via RFM model and RS theory [40]	K-means and LRFM
An intelligent market segmentation system using k-means and particle swarm optimization [41]	K-means,SOM + K-means and Particle Swarm Optimization (PSO) +K-means

While segmenting customers, many methods have been used in the literature, as represented in Table 2.1. Among the methods, the most preferred one can be said as K-means clustering. Because of the efficiency and applicable easiness K-means algorithm was selected to be used in the customer segmentation part of this study.

2.3 Analytical Hierarchy Process (AHP)

Saaty developed the analytical hierarchy process. It has been utilized when it comes to complex decision making (DM), and it can be useful when a group of people working together to make a decision; it can be described as one of the multiple criteria decision making (MCDM) methods that can use qualitative and quantitative factors together in the decision-making process [42]. With the help of AHP, the optimal answer to an MCDM problem can be provided; AHP uses a hierarchy model where goals are at the top, followed by criteria and alternative decisions, respectively [43]. In Figure 2.3, the hierarchical structure of AHP can be seen. At the top-level, the overall goal of the problem is described. In the second level, there are criteria that contribute to the overall goal, and at the bottom, there are diversified alternatives that will be assessed according to criteria [44]. In AHP, the problem is decomposed into sub-problems, each problem can be evaluated one by one, AHP is one of the well-known and utilized in DM mechanisms, because of its simplicity and high precision rates AHP is one of the most used MCDM mechanisms, and AHP has six fundamental steps [45].

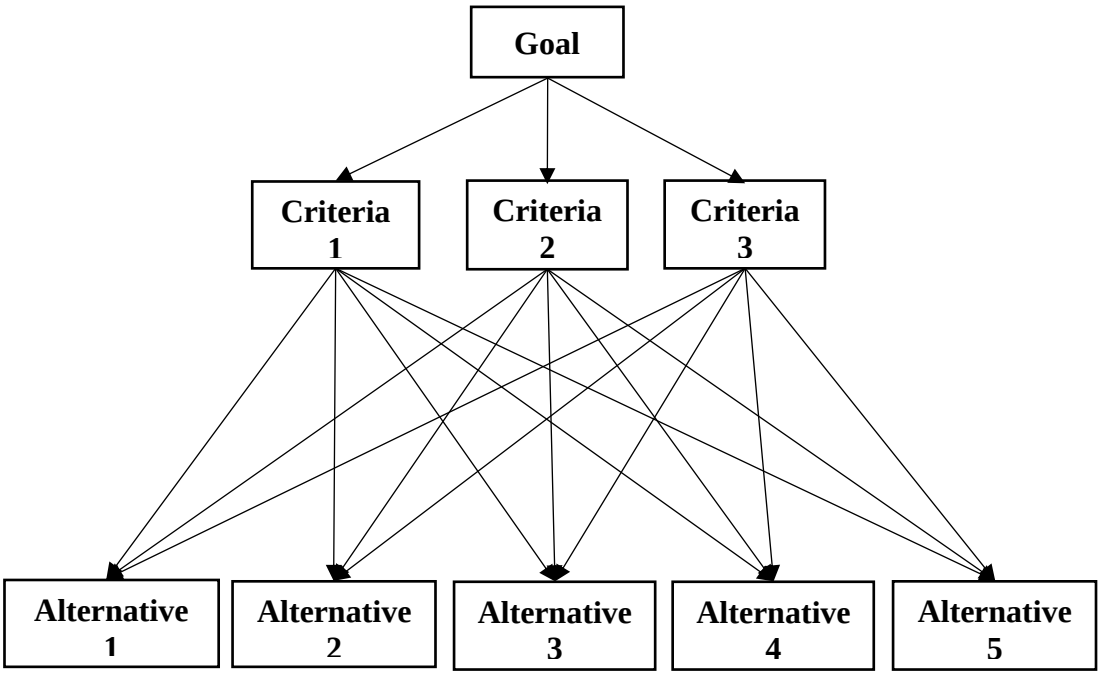

Figure 2.3: Hierarchical Structure of AHP [44].

Six fundamental steps are explained in [45]:

- Step1, in AHP, the problem is decomposed into a sub-problem hierarchy so that each problem can be solved one by one.
- Step 2, a decision matrix is created in this step with respect to Table 2.3. Priority scores are determined by the basic 1-9 scale described in [44].
- Step 3, comparison matrix A is created with respect to the basic 1-9 scale by Saaty was used [44], and W is created as an eigenvector. Weight calculation of each item is also more accurate since it is logical to pairwise comparison rather than making an overall assignment. A sample comparison matrix can be seen in Table 2.4.

Table 2.3: Intensity of importance table [44].

Intensity of Importance	Definition	Explanation
1	Equal importance	Two activities contribute equally to the objective
3	Moderate importance of one over another	Experience and judgment strongly favor one activity over another
5	Essential or strong importance	Experience and judgement strongly favor one activity over another
7	Very strong importance	An activity is strongly favored and its dominance demonstrated in practice
9	Extreme importance	The evidence favoring one activity over another is of tile highest possible order of affirmation
2,4,6,8	Intermediate values between the two adjacent judgments	When compromise is needed

Table 2.4: Comparison Matrix.

Criteria	Criteria (C1)	Criteria (C2)	Criteria (C3)
Criteria (C1)	-	C1 - C2 Comparison	C1 - C3 Comparison
Criteria (C2)	C2 - C1 Comparison	-	C2 - C3 Comparison
Criteria (C3)	C3 - C1 Comparison	C3 - C2 Comparison	-

- Step 4, consistency ratio is also measured to evaluate the decision performance; a lower inconsistency index means higher consistency. The related index must be lower than 10% to conclude a consistent output.

- Step 5, standardization of the matrix needs to be done before doing further calculations. Each column must be divided by the entered number of the respective column. After this step normalized matrix has been created.
- Step 6, measuring each matrix value that gives the criteria weights, is crucial. The related weights should be tested with the values in Table 2.5.

Table 2.5: Random consistency index table.

n	2	3	4	5	6	7	8	9	10	11
Random Consistency Index	0	0,58	0,9	1,12	1,24	1,32	1,41	1,45	1,49	1,51

The pairwise comparison matrix, denoted as **A,** is an m×m matrix where m is the number of criteria. Entry a_{jk} represents the relative importance of the *j*th criterion to the *k*th criterion. If $a_{jk} > 1$, *j*th criteria is more important than *k*th criteria. If $a_{jk} < 1$, the *k*th criterion is less important than the *j*th criterion. If $a_{jk} = 1$, then both criteria have the same importance level. Entries of pairwise comparison matrix A should satisfy equation 2.1 [46].

$$a_{jk} * a_{kj} = 1 \tag{2.1}$$

After creating pairwise comparison matrix A, normalized pairwise comparison matrix A_{norm} can be derived from A. Entries of A_{norm} are calculated with the help of equation 2.2 [46].

$$\bar{a}_{jk} = \frac{a_{jk}}{\sum_{i=1}^{m} a_{jk}} \tag{2.2}$$

Criteria weight vector $\boldsymbol{w}$ can be created by taking the average of the entries of every row of A_{norm}; vector elements of $\boldsymbol{w}$ can be calculated using equation 2.3 [46].

$$w_j = \frac{\sum_{l=1}^{m} \bar{a}_{jl}}{m} \tag{2.3}$$

Matrix of option scores S is an n×m matrix, s_{ij} represents the score of option i with respect to criterion j. To get the scores, a pairwise comparison matrix $\mathbf{B}^{(j)}$ is built for

each m criterion. AHP applies the same two-step process as in A, and $s^{(j)}$ is obtained, $j = 1,\ldots,m$. Vector $s^{(j)}$ contains the score of the options to the criteria j. Score matrix is created as in equation 2.4 [46].

$$S = [s^{(1)} \ldots s^{(m)}] \tag{2.4}$$

After obtaining **w** and score matrix **S**, a global score vector **v** is created by multiplying S and v as in equation 2.5. The *i*th entry of **v** shows the calculated score for the *i*th option. The final step is completed by simply ranking the global scores in decreasing order [46].

$$v = S \cdot w \tag{2.5}$$

In AHP, there is also an effective and useful consistency checking step. The technique includes calculating *Consistency Index (CI)* as in equation 2.6. In equation 2.6, x is a scalar computed as the average of the vector elements whose *j*th element is the ratio of **A****w* to the corresponding element of ***w*** [46].

$$CI = \frac{x - m}{m - 1} \tag{2.6}$$

In a perfect DM process, the expectation is having *CI = 0*. However, a low value of *CI* is also accepted. To complete the AHP steps, equation 2.7 needs to be satisfied. If equation 2.7 is satisfied, then the decision-maker can conclude that inputs of A are random [46]. In equation 2.7, *RI* denotes *Random Consistency Index*, described in Table 2.5.

$$\frac{CI}{RI} < 0{,}1 \tag{2.7}$$

2.4 RFM Method and Extensions

Recency-Frequency-Monetary (RFM) analysis is a segmentation technique that uses a company's customer database that consists of customers' past behavior; the technique originates from examples of direct marketing in sales companies in the 1960s [47]. RFM is a widely used, well-known behavioral-based method, which completes the

creation of customer profiles with respect to criterion. Besides with RFM method, the complexity of the analysis is also reduced [48]. Customer segmentation is completed with respect to 3 variables that can be named as Recency (R), Frequency (F), and Monetary (M). Recency (R) demonstrates the length of time in months or days from the latest purchase. Frequency (F) is the number of actions a customer takes in a specified period. Monetary (M) reflects the total amount of money spent by a customer in the analysis period [48]. Small R values indicate higher customer visits, higher F values indicate higher customer visits, and higher M values indicate more customer-generated profit [4]. RFM Analysis technique has some pros and cons. The main advantages can be summarized as a crucial method for calculating customer lifetime value, a basis for a continuing stream for customer segmentation, and increasing company profits. The main disadvantages can be summarized: using only 3 variables can result in inefficient analysis, frequency and monetary attributes can be highly correlated, underutilization from non-profit customers, depending on the industry importance of RFM attribute can be different [48].

Table 2.6: RFM Scores [49].

Score	Characteristic
5	Potential
4	Promising
3	Can't lose them
2	At-risk
1	Lost

RFM models have evolved with new research. In the traditional technique, all dimensions of customer data are divided into five categories with codes. This method is known as the “customer quintile method” after coding each customer compared to others with respect to dimensions. The coding step can be completed by using scores as in Table 2.6. 20% of the customers are coded in 5 in a dimension. The following 20% are coded in 4, and so on. All the customers are coded as 555, 554, 553, …, 111. Customers with the highest RFM scores are known as the company’s most profitable customer group [48].

In “behavior quintile scoring,” monetary score creates almost equal sales in each quintile. In this method, customers who have similar behavior can be grouped together. Weighted RFM method, depending on the industry, weights to R, F, M parameters can also be used with the help of AHP [48]. During the decades, improvement of the RFM method is not only continued with different types of techniques but also evolved with the addition of new parameters. A recent study added customer relation length as a new parameter L [50]. An extension of the RFM method is known as RFMTC. In this method, two variables are added as time since first purchase (T) and churn probability (C) [51]. Another extension of the traditional RFM model is known as the RFMLP model; in RFMLP, two parameters are added as Length (L) and Periodicity (P) [52]. Length can be described as the difference of customer’s first and last visit in days; it can be explained as a measure of loyalty; higher length demonstrates more loyal customers [52]. Periodicity checks whether a customer visits the store regularly or not. It can be defined as the standard deviation of inter-visit times (IVT) as in equation 2.8, and calculation of IVT can be seen in equation 2.9 [52].

$$Periodicity = stdev(IVT_1, IVT_2, \dots, IVT_n) \quad \textbf{(2.8)}$$

$$IVT_i = date\_dif(t_{i+1}, t_i) \quad \textbf{(2.9)}$$

2.5 Churn Analysis

In most emerging markets, companies are willing to detect customers who plan to cease the transaction and switch to one of the competitors. Besides, captivating new customers is also very challenging and nearly six times more costly than avoiding current customers from churning [53]. To maintain the success of a business, retaining existing customers is crucial because it is hard and costly to find new customers. Customer churn can be defined as the aptitude to cease the contract with the company or switch to one of the competitors in the future [54]. Most customer churn models target customers more likely to churn during a retention campaign, which affects the highly efficient usage of limited resources. Customer churn analysis projects can be profitable for companies, as explained in [55]:

- Finding new customers is nearly six times costly than customer retention.

- Customers who are in the long run with the company create more profits and likely to be less sensitive to challenging market conditions and can provide positive referrals.
- Opportunity cost is increased due to lost sales.

Predicting the churn can be completed by observing the consumer behavior or individual behaviors that have signs of attrition. Using modeling and ML techniques is unavoidable during the prediction process [56]. Churn analysis projects are completed in many different industries such as digital markets, communication industries, banking, insurance, retail stores, online platforms, and automotive industries. Also, many other AI & ML techniques are used in churn analysis, such as clustering, decision trees, random forests, logistic regression, support vector machines [56]. There are many models in churn analysis in literature, as summarized in Table 2.7 [55].

Table 2.7: Techniques used in literature for Churn analysis [55].

Researcher(s)	Techniques
Eiben et al.	Logistic regression, genetic programming, rough data analysis
Mozer et al.	Logistic regression, decision tree, neural network
Hwang et al.	Logistic regression, decision tree, neural network
Buckinx et al.	Logistic regression, neural network, random forests
Larivière et al.	Logistic regression, linear regression, random forests
Burez et al.	Logistic regression, gradient boosting, (weighted) random forests
Coussement et al.	Logistic regression, support vector machine, random forests
Kumar et al.	Logistic regression, support vector machine, random forests

2.6 Machine Learning

Machine Learning (ML) is a sub-field of Artificial Intelligence (AI) that automatically gives learning and improvement from experiences abilities to systems. ML generally focuses on developing computer programs that can access data and learn from data; the main target of ML is allowing computers to learn without human assistance [57]. By using ML, one can predict an outcome or uncover some patterns that are hard to spot by humans. In other words, with the help of ML, users apply algorithms to uncover meaningful insights, find relations and predict the future results of events [58]. In most industries like healthcare, insurance, energy, marketing, manufacturing, financial technology, ML has important effects. When applied in a meaningful way, ML allows businesses to reach an optimal solution to some practical daily problems,

increasing business value[58]. Most statistical analysis methods depend on rule-based decision-making. ML handles tasks that are very problematic when it comes to defining rules. ML can be applied to uncountable business cases in which an outcome depends on numerous factors that are impossible to monitor by humans [58]. Companies that effectively utilize ML and AI technologies will gain a huge competitive advantage, and AI technologies will create nearly $50 trillion of value in 2025 [58].

2.6.1 Supervised Machine Learning

In supervised machine learning, algorithms reveal relationships, insights, and patterns from a labeled dataset, which is a dataset that has known values for each record of the target variable. During the training phase of the solution, the correct answer is provided to an algorithm that leads the algorithm to learn the relationship between features and target variable, enabling it to reveal insight and make predictions of future outcomes [59]. The computer learns from past data in supervised learning and uses this information to predict future events. For improved model accuracy, the input data is labeled [60].

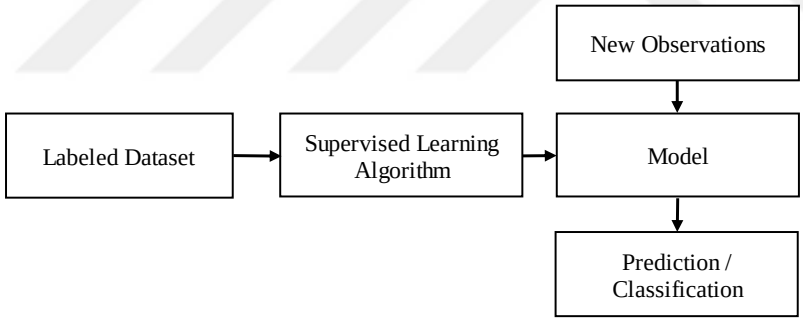

Figure 2.4: General overview of Supervised Machine Learning Models [60].

Supervised machine learning algorithms can be categorized as either classification or regression models.

- Classification models: Classification models are used when the output is categorized as “Yes” or “No,” “Pass,” or “Fail,” etc. This model is used to predict the outcome category [60].
- Regression models: Regression models are used when the output is a value, such as salary, weight, or pressure. These models are used to predict future numerical values based on observed data values [60].

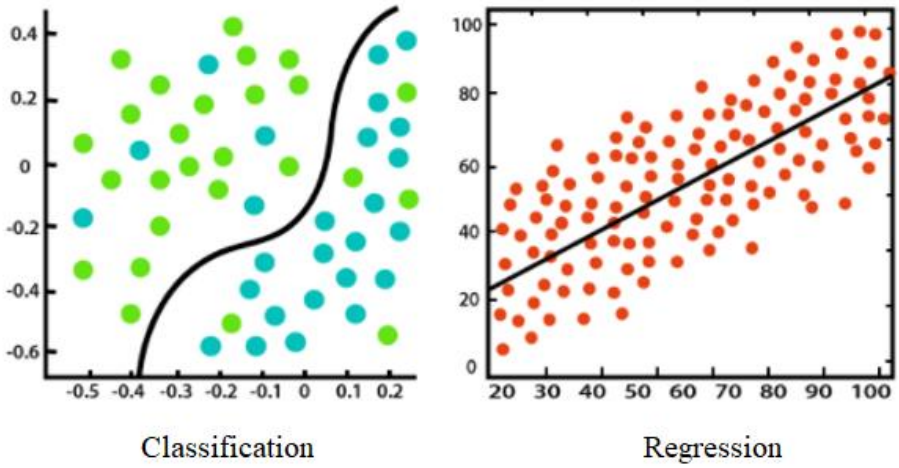

Figure 2.5: Example illustration of classification and regression models [60]

Supervised machine learning models are used in several industries to solve real-life problems such as [59]:

- Reducing customer churn
- Determining customer lifetime
- Recommendations systems
- Sales forecasting
- Supply and Demand forecast
- Fraud detection
- Resource allocation
- Equipment maintenance prediction
- Face detection
- Spam detection
- Weather forecasting

2.6.2 Unsupervised Machine Learning

In unsupervised machine learning, neither classified nor labeled data is trained. The machine must classify data without any information [60]. The main idea behind unsupervised machine learning is providing unknown insights and uncovering hidden

patterns in the data. Another feature of unsupervised machine learning is there are no predefined outcomes in the given dataset [60].

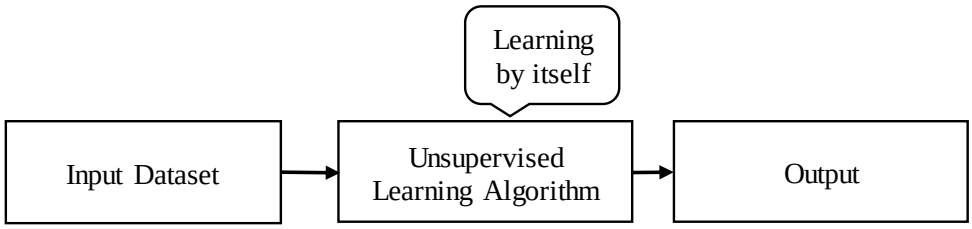

Figure 2.6: An overview of unsupervised machine learning models [60].

Unsupervised machine learning algorithms can be summarized in 4 different categories: clustering, anomaly detection, association mining, and latent variable models.

- Clustering: Data is split into groups based on similarity measures in clustering algorithms. The method involves grouping unlabeled data into clusters [61].
- Anomaly detection: With the help of anomaly detection algorithms, unusual data points in the dataset discovered can be used to find fraudulent transactions or identify a data entry error [61].
- Association mining: Based on the frequency of occurring together in the dataset association mining method identifies sets of items. It can be used in the retail industry for basket analysis because it enables the discovery of purchased goods simultaneously [61].
- Latent variable models: Latent variable models are generally used for data preprocessing, such as dimensionality reduction or decomposition of the dataset into multiple components [61].

Unsupervised machine learning models are used in several industries to solve real-life problems such as [59]:

- Fraud detection
- Data entry errors
- Market basket analysis

2.6.3 K-Nearest Neighbors Algorithm

The K-nearest neighbors (KNN) algorithm is a supervised machine learning technique and classification algorithm used at the beginning of the 1970s as a statistical estimation and pattern recognition technique [62]. In KNN, all available cases are stored, and new cases are classified according to a similarity function. Similarity

functions can be defined as Equation 2.10 Euclidean distance function, Equation 2.11 Manhattan distance function, or Equation 2.12 Minkowski distance function [62].

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \quad \textbf{(2.10)}$$

$$\sum_{i=1}^{k}|x_i - y_i| \quad \textbf{(2.11)}$$

$$\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q} \quad \textbf{(2.12)}$$

A new case is classified according to the majority votes of neighbors. The case is being assigned to the class most common it's K nearest neighbors, determined by one of the distance functions. Determination of the optimal value of K is done by inspecting the data. In general, a large value of k is good because it reduces the noise in the data. However, it is not guaranteed. Optimal K has been between 3-10 because it produces better results than 1-NN [62].

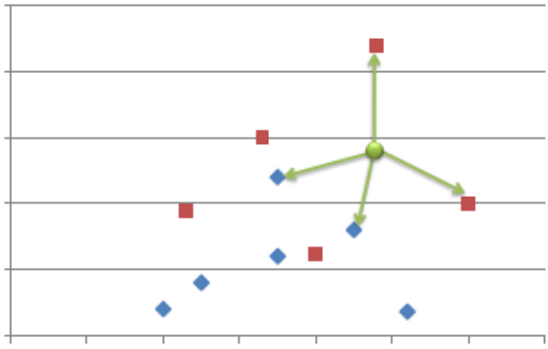

Figure 2.7: An example of the K-NN algorithm structure.

One major problem of the KNN algorithm is the variables that have different scales; for instance, if one variable is yearly income and the other variable is age in years, then the income value will have a higher effect on calculated distance, to solve that kind of problem, variables can be standardized with the help of Equation 2.13 [62].

$$X_i = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (2.13)$$

Calculation steps of the KNN algorithm can be summarized as [63]:

1. Load the training and test data
2. Choose the value of K
3. For each point in test data:
   - Find Euclidean distance to all training data points
   - Store the Euclidean distance in a list and sort it
   - Choose the first k points
   - Assign a class to the test point based on classes of present chosen points
4. End

2.6.4 Logistic Regression

Logistic regression is another supervised machine learning method used in the case of classification. When the outcome has only two values, the logistic regression method can be used predictors can be both numerical or categorical. It creates a logistic curve that is limited to values between 0 and 1 [64]. The logistic curve, which can be represented in Figure 2.8, is constructed with the natural logarithm of the odds of the target variable. Besides, predictors do not have to be normally distributed or do not necessarily have equal variance in each group [64].

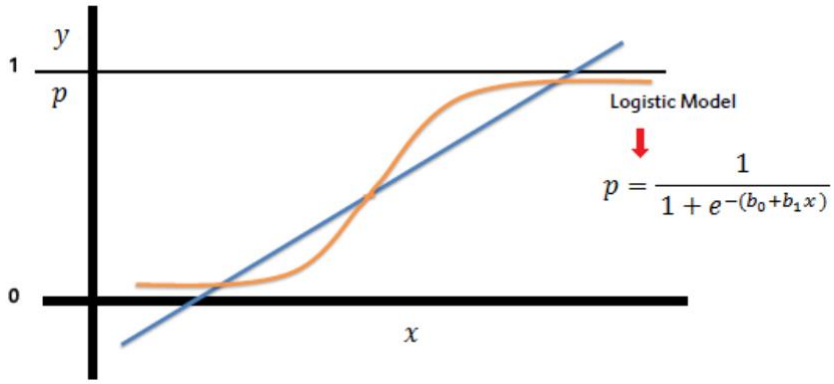

Figure 2.8: Logistic Curve and Model.

Constant b_0 moves the curve horizontally, and slope b_1 affects the steepness. By simple math, the logistic regression model can be transferred in terms of an odds ratio as in Equation 2.14 [64].

$$\frac{p}{1-p} = e^{(b_0 + b_1 x)} \quad \textbf{(2.14)}$$

Taking the natural logarithm of both sides of Equation 2.14, the equation becomes a linear function of predictors, defined as log-odds (logit) as in Equation 2.15 [64].

$$\ln(\frac{p}{1-p}) = b_0 + b_1 x \quad \textbf{(2.15)}$$

2.6.5 Support Vector Machines

Support Vector Machine (SVM) is a classification algorithm that performs classification by maximizing the margin between two classes with the found hyperplane. A hyperplane is defined by the vectors called Support Vectors [65].

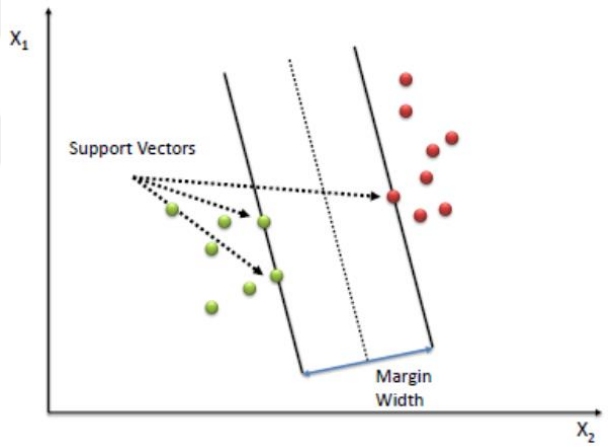

Figure 2.9: Illustration of Support Vectors [65].

To define an optimal hyperplane width of the margin *(w)* needs to be maximized, and *b* needs to be found using Quadratic Programming to solve the system represented in Equation 2.16 and Equation 2.17 [65].

$$\min \frac{1}{2} \|w\|^2 \quad \textbf{(2.16)}$$

$$s.t. \quad y_i(w \,.\, x_i + b) \geq 1, \forall x_i \quad \textbf{(2.17)}$$

In the case of linearly separable data, algorithm results in global minimum, ideal SVM should produce hyperplane that separates the vectors into classes. However, sometimes this cannot be achieved by SVM. In this case, SVM tries to find a hyperplane that maximizes the margin and minimizes misclassification [65]. To solve this situation, slack variable s_i is added to the model, which allows some instances to fall in the margin. Still, it also penalizes them, as shown in Equation 2.18 and Equation 2.19, also represented in Figure 2.10 [65].

$$\min \frac{1}{2}\|w\|^2 + C\sum_i s_i \quad \textbf{(2.18)}$$

$$s.t. \quad y_i(w\,.x_i + b) \geq 1 - s_i, \forall x_i \quad \textbf{(2.19)}$$

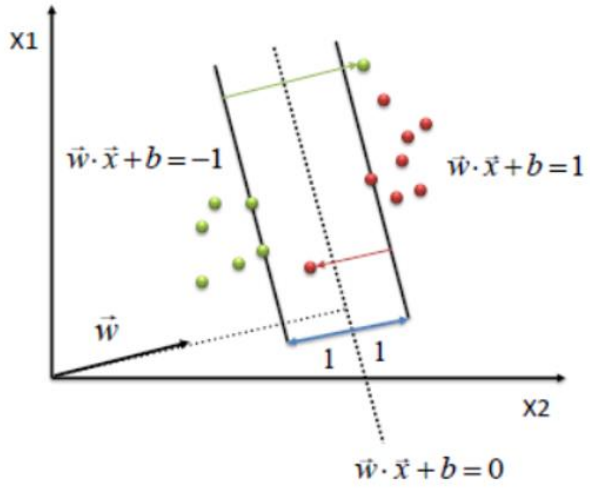

Figure 2.10: SVM model with slack variable [65]

Separation of groups is completed by a straight line, flat plane, or N-dimensional hyperplane, however in some cases, a non-linear region can separate the groups. In that case, SVM uses kernel function, which is a non-linear function to map the data in a different space, which means that the system learns non-linear function in a high-dimensional feature space, which is called kernel trick [65].

2.6.6 Random Forest

Random Forest (RF) is a supervised machine learning algorithm used to solve both classification and regression types of problems. It utilizes ensemble learning which can be described as a combination of many classifiers [66]. In RF, many decision trees are constructed, forest generated by the RF is trained through bootstrap aggregating, which is a method that improves the accuracy of the model [66]. Decision Trees are building blocks of Random Forest. Every decision tree has decision nodes, leaf nodes, and a root node. A leaf node can be described as the output of the decision tree. After having outputs by the majority voting system, the RF algorithm can create the final output, as shown in Figure 2.10 [66].

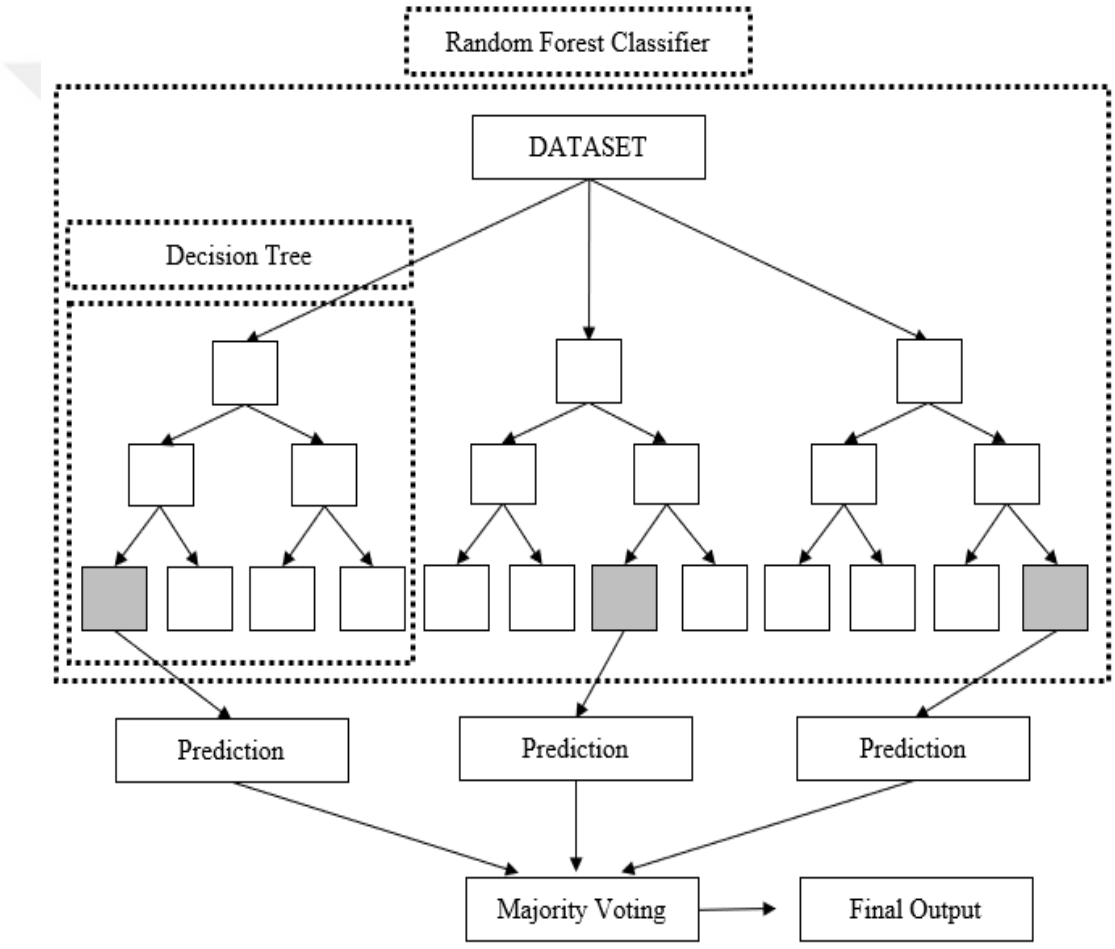

Figure 2.10: Overview of Random Forest Algorithm.

RF algorithm can handle large datasets, results with higher accuracy than decision tree algorithms; however, it requires more resources and consumes more time, and has a more complex structure [66].

2.6.7 Performance of Classification

After performing the classification step via algorithms, the performance of the classifier should also be tested to understand the goodness of the model. A confusion matrix can be used to demonstrate the performance of the classification model on the test data by comparing the actual and predicted results [67]. To build a confusion matrix True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN) are used, and the accuracy score of the model is calculated [67].

Table 2.8: Confusion Matrix.

		Actual Values	
		Positive	Negative
Predicted Values	Positive	TP	FP
	Negative	FN	TN

When the prediction is positive and true, it is called TP. If the value is predicted as negative and true, it is called TN. FP is also called a Type 1 Error, and it occurs if the prediction is positive and false. FN is known as Type 2 Error, and it emerges if the prediction is negative and false [67]. Using the values calculated in the confusion matrix, the accuracy score of the model, which can be described as the ratio of correct predictions, can also be calculated as in Equation 2.20.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad \textbf{(}2.20\textbf{)}$$

Precision and recall are used to demonstrate the performance of the models. Precision can be described as the ratio of relevant samples among all samples, as presented in Equation 2.21. On the other hand, recall is also known as sensitivity, the ratio of retrieved samples among all samples, as shown in Equation 2.22. A perfect classifier should have precision and recall values 1 [68]. Precision and Recall values are generally given together. When a single number is required F-score is calculated from the Precision and Recall values as in Equation 2.23.

$$Precision = \frac{TP}{TP + FP} \quad (2.21)$$

$$Recall = \frac{TP}{TP + FN} \quad (2.22)$$

$$F\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (2.23)$$

When visualization of the performance of the model is needed, Area Under Curve (AUC) Receiver Operating Characteristics (ROC) curve can be used as shown in Figure 2.11, ROC can be described as likelihood bend, and AUC shows the performance of recognizing the classes, if the value of AUC is higher than the model is distinguished better. [68]

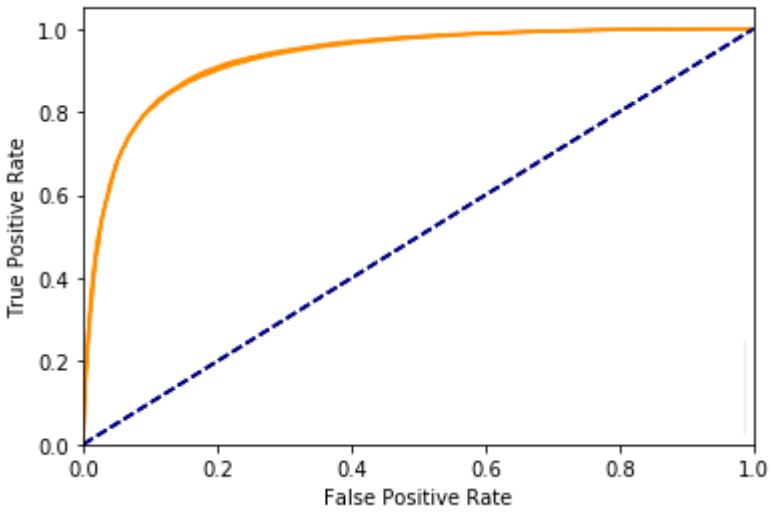

Figure 2.11: Sample ROC curve [68].

2.6.8 K-Means Clustering

K-means clustering algorithm is an unsupervised machine learning algorithm that tries to partition n objects into k clusters that every object is in the cluster to the nearest mean. Clusters are created with the greatest possible distinction [69]. The object of the k-means clustering method can be explained as minimizing squared error function as shown in Equation 2.24 [69].

$$J = \mathbf{1}\sum_{j=1}^{k}\sum_{i=1}^{n}\left\|x_i^{(j)} - c_j\right\|^2 \quad \textbf{(2.24)}$$

Calculation steps of the K-means Algorithm can be summarized as [69]:

1. By using predefined *k*, cluster the data into k groups.
2. Select k cluster centers randomly.
3. Assign every instance into their closest cluster center using the Euclidean distance function.
4. Calculate the mean of all instances in each cluster.
5. Repeat 2nd, 3rd, and 4th steps until no change in clusters until consecutive iteration.

It can be said that K-Means is an efficient method. However, the number of clusters should be predefined because it can affect the results as terminating in a local optimum [69]. One of the methods that are used in literature to decide the number of clusters is known as the Elbow Method, which provides us a decision about what a good *k* should be based on the sum of squared errors between instances and their clusters' centroids, *k* value selected when the sum of squared errors is flattened out [70].

3. MATERIAL AND METHOD

3.1 Problem

The heavy equipment industry is one of the most dynamic and competitive industries due to globalization and the increased amount of construction sites globally. Global market size is growing, and that's why competition is getting higher day by day. Turkish heavy equipment market size is also increasing. To survive in this competitive global market, companies should seek ways to know their customers and reduce customer churn.

Companies realized that their most important assets are their customers in the last decades. To achieve success in the long run, companies need to identify customers and their needs. Segmentation allows companies to meet their customers deeply and to understand their requirements. To manage marketing campaigns, companies can utilize customer segmentation.

It is a known fact that finding new customers is costly and time-consuming than avoiding current customers from churning. Because of its advantageous situation, retaining existing customers is very important for companies. Therefore, to maintain the success of a business, companies need to find ways of reducing the churn of customers.

The main idea of this study is to create customer segments with the RFMLP method and predict customers that will churn in the future. Customer transaction data is collected. The customer segmentation process has been completed by expert opinion previously. However, this process can be time-consuming and insufficient when a company has too many customers. With the help of this study, analytical and time-effective methods will be used in customer segmentation. A supervised ML algorithm is used to create customer segments. After segmentation, customers that will be churners in the future are predicted with the help of supervised ML algorithms such as LR, RF, SVM, and K-NN.

3.2 Dataset Preprocessing

To implement proposed techniques and algorithms, it is vital to understand and preprocess the data. The dataset is taken from the complex data warehouse from one of the biggest companies in Turkey's heavy equipment industry. Dataset consists of customer transaction data of the company, which consists of spare parts sales data in years between 2018 – 2020. There are 777053 rows in the raw data, 137735 unique invoices, 54321 different material numbers, and 4568 different customers in total.
The details of raw data can be seen in Table 3.1.

Table 3.1: The details of raw data**.**

Attribute Name	Description	Data Type
VBELN	Invoice ID	Numerical
MATNR	Material Number	Text
FKDAT	Invoice Date	Date
ZFKIMG	Amount of Material Invoiced	Numerical
KUNRG	Customer ID	Numerical
DMBE2	Invoice Amount in EUR	Numerical

To implement RFMLP based customer segmentation technique, raw data needs to be preprocessed. A limited view of transformed data can be seen in Table 3.2.

Table 3.2: Transformed raw data to conduct RFMLP analysis.

Customer Number	Recency	Frequency	Monetary	Length	Period
0004006849	31	32	44337,4	990	59,8
0004006877	133	3	783,2	503	24,7
0004006928	195	3	280,4	823	200,1
0004007044	2	221	179357,5	1088	10,0
0004007075	65	10	4103,5	506	44,8
0004007093	190	13	8081,0	881	100,0

Because of having different scales of the transformed data min-max scaler method of Python 3.8 was used with default parameters to standardize the attributes in order not to have weighting problems while conducting RFMLP analysis. Limited view of standardized data can be seen in Table 3.3. Churn information is also added according to the customer number to be used in the churn analysis part.

Table 3.3: Standardized raw data to conduct RFMLP analysis.

Customer Number	Recency	Frequency	Monetary	Length	Period
0004006849	0,00378	0,00281	0,90494	0,02834	0,05976
0004006877	0,00024	0,00005	0,45978	0,12157	0,02475
0004006928	0,00024	0,00002	0,75229	0,17824	0,20011
0004007044	0,00024	0,00006	0,52377	0,34186	0,26092
0004007075	0,02686	0,01135	0,99452	0,00183	0,01001
0004007093	0,00110	0,00026	0,46252	0,05941	0,04482

3.3 Proposed Method

After completing the required data preprocessing steps, 2018 and 2019 data are used as training data, and 2020 data is used as test data. AHP method was used to calculate the importance weights of each RFMLP attributes. After that, the KNN algorithm was used to create customer segments. At the churn analysis part, KNN, LR, RF, and SVM algorithms were used to predict customers that will be churned in the future. Detailed information will be given in the thesis's 4th part (Findings). A schematic view of the proposed method can be seen in Figure 3.1. All the coding part is completed in Spyder 5.1.5 IDE and Python 3.8.

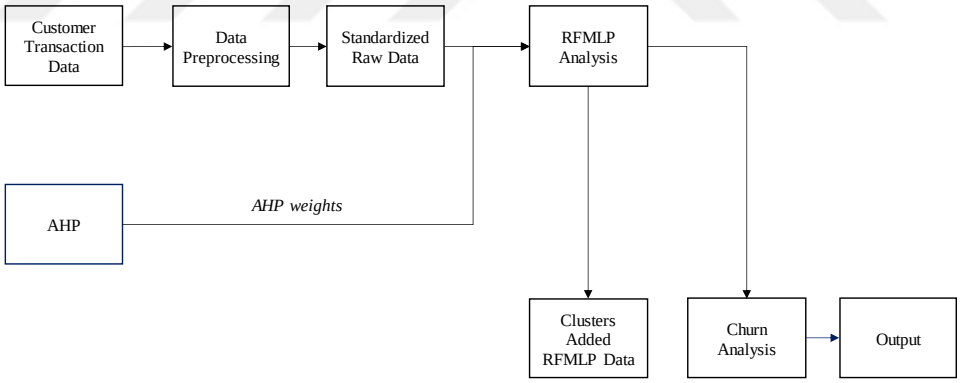

Figure 12: Proposed method RFMLP based Segmentation and Churn Analysis.

4. FINDINGS

4.1 Analytical Hierarchy Process (AHP)

After data preprocessing steps, to handle the relative importance of Recency, Frequency, Monetary, Length, and Periodicity variables, Analytical Hierarchy Process was conducted with five different sales managers from distinct regions of the company. The definitions of the variables, which can be seen in Table 4.1, are explained to the respondents, and they were asked to fill an excel table which can be seen in Appendix A, to complete AHP.

Table 4.1: Definition of R, F, M, L, and P variables.

Variable	Definition
Recency	The length of time from the latest purchase
Frequency	Number of purchases of a customer in a specific period
Monetary	The total amount of money spent by customers in the analysis period
Length	Difference between customer's first and last purchase date
Periodicity	The standard deviation of customer's inter-visit times

All the responses were analyzed and checked according to CI and CR values. The aggregated Pairwise Comparison Matrix was created using Geometric Means of responses and shown in Table 4.2.

Table 4.2: Pairwise Comparison Matrix.

Criteria	Recency	Frequency	Monetary	Length	Periodicity
Recency	1	2,141	3,380	1,176	0,500
Frequency		1	1,607	0,644	0,292
Monetary			1	0,822	0,201
Length				1	0,290
Periodicity					1

A normalized matrix was created with the help of Table 4.2 and can be seen in Table 4.3.

Table 4.3: Normalized matrix.

Criteria	Recency	Frequency	Monetary	Length	Periodicity
Recency	0,122	0,116	0,101	0,173	0,135
Frequency	0,261	0,249	0,212	0,316	0,231
Monetary	0,412	0,401	0,340	0,248	0,335
Length	0,143	0,161	0,279	0,204	0,232
Periodicity	0,061	0,073	0,068	0,059	0,067

Weights of each criterion can be seen in Table 4.4, Monetary was found as the most important criterion by sales managers, and Periodicity was found as the least important criterion. Regarding calculation steps of AHP, inconsistency was found acceptable because the Consistency Ratio is less than 0,1. Calculated CI, RI, and CR values are represented in Table 4.5.

Table 4.4: Calculated weights as a result of AHP.

Criteria	Weights
Recency	0,129
Frequency	0,254
Monetary	0,347
Length	0,204
Periodicity	0,066

Table 4.5: CI, RI, and CR values.

Index	Weights
CI	0,018
RI	1,120
CR	0,016

4.2 Customer Segmentation

The K-means algorithm was used in the customer segmentation part of the study. Before going into the details of the K-means algorithm, k must be decided. With the help of the elbow method, k is determined as 4 and 4 different customer segments were created. The output of the Elbow Method can be seen in Figure 4.1.

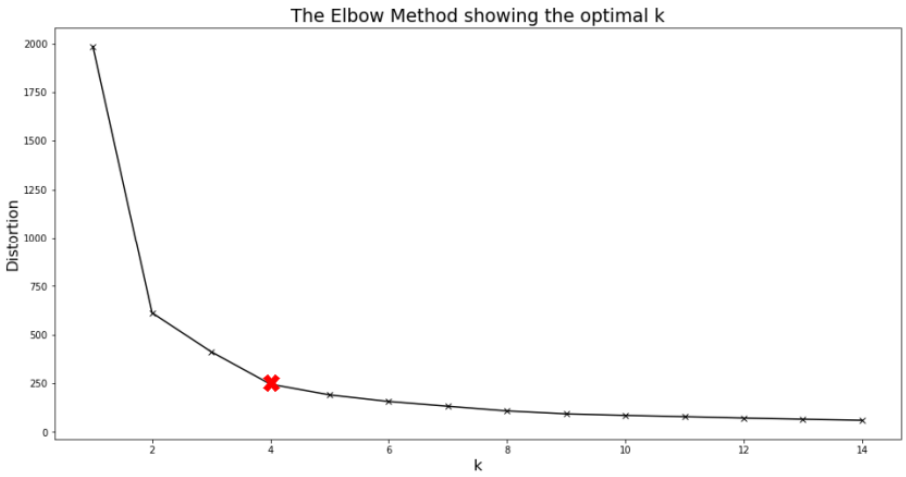

Figure 4.1: Elbow method output.

To decide the importance order of clusters, an extended version of Equation 4.1, as in Equation 4.2, has been used with the weights that come from as a result of AHP.

$$C^j = W_R C_R^j + W_F C_F^j + W_M C_M^j \quad \textbf{(4.1)}$$

$$C^j = W_R C_R^j + W_F C_F^j + W_M C_M^j + W_L C_L^j - W_P C_P^j \quad \textbf{(4.2)}$$

The created clusters, the number of customers in each cluster, and the centroids for each cluster can be seen in Table 4.6 and the summarized results in Table 4.7.

Table 4.6: Summarized results of clustering.

Number of Customers	R_Center	F_Center	M_Center	L_Center	P_Center	Clusters	C(j)
1422	0,07110	0,01065	0,00766	0,80993	0,10802	A	0,1726
1105	0,77902	0,00004	0,00010	0,02281	0,90736	B	0,0459
1092	0,23560	0,00007	0,00021	0,11393	0,87321	C	-0,0036
949	0,38278	0,00106	0,00119	0,30782	0,08258	D	0,1076

Table 4.7: Summarized results of customer segmentation.

Clusters	Recency	Frequency	Monetary	Length	Period
A	78,00492	88,14698	172.023.072,43	885,92	108,1
B	852,24796	1,34027	1.757.177,32	24,95	124,2
C	257,75092	1,57051	3.595.205,93	124,64	136,5
D	418,79874	9,68809	17.877.076,32	336,38	82,4

4.3 Churn Analysis

Before implementing churn analysis, the data containing Customer Code, R, F, M, L, and P values are merged with Churn information data given. In churn information data, there are customer codes and a binary variable that provides 1 if customer churned and 0 if the customer is still working with the company. In churn analysis, 4 different algorithms were used: KNN, LR, RF, and SVM. Before implementing algorithms, each assumption of algorithms was checked and resulted as data can be used because of no violated assumption. The correlation matrix of input variables is represented in Table 4.8.

Table 4.8: Correlation matrix of input variables.

	Recency	Frequency	Monetary	Length	Periodicity
Recency	1	-0,12	-0,11	-0,06	0,18
Frequency	-0,12	1	0,14	0,15	-0,13
Monetary	-0,11	0,14	1	0,14	-0,13
Length	-0,06	0,15	0,14	1	-0,07
Periodicity	0,18	-0,13	-0,13	-0,07	1

Data is split as 75% training data and 25% test data, with the Python scikit learn train test split method.

In the first model, the KNN algorithm is used, and created confusion matrix of the KNN algorithm can be seen in Table 4.9.

Table 4.9: KNN Algorithm confusion matrix.

	0	1
0	714	76
1	125	227

In the second model, the logistic regression algorithm is used, and created confusion matrix of the LR algorithm can be seen in Table 4.10.

Table 4.10: LR algorithm confusion matrix.

	0	1
0	734	56
1	152	200

In the third model, the random forest algorithm is used, and created confusion matrix of the RF algorithm can be seen in Table 4.11

Table 4.11: RF algorithm confusion matrix.

	0	1
0	721	69
1	28	324

In the last model, the support vector machines algorithm is used, and created confusion matrix of the SVM algorithm can be seen in Table 4.12.

Table 4.12: SVM algorithm confusion matrix.

	0	1
0	711	79
1	134	218

Overall classification performances of algorithms, precision, recall, F1 – score, and AUROC can be found in Table 4.13. From Table 4.13, it can be said that RF outperforms all the other algorithms, and it has the best precision value. From Figure 4.2, it can also be seen that RF has the best AUROC value and outperforms other algorithms.

Table 4.13: General classification performance of all algorithms.

	Precision	Recall	F1 - Score	AUROC
KNN	0,80	0,77	0,78	0,8782
LR	0,80	0,75	0,77	0,8276
RF*	0,89	0,92	0,90	0,9689
SVM	0,79	0,76	0,77	0,8128

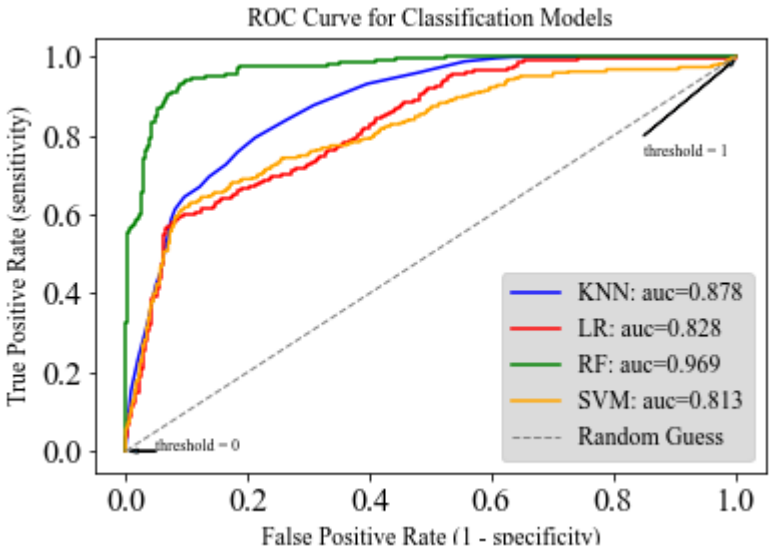

Figure 4.2: ROC curve graph.

When it comes to assumptions, in a KNN model, the data is needed to be in a feature space. This assumption is satisfied since the data in this study is in multidimensional vectors.

In logistic regression, observations should be independent of each other, and independent variables have no multicollinearity. In this study, no observations are repeated, and from the correlation matrix of inputs, there is no multicollinearity in the data. Thus assumptions are satisfied. The requirement of having a large sample size is also satisfied due to the size of the data.

There is no formal distributional assumption in the RF case since RF is non-parametric. That's why it can handle skewed and multi-modal data. In this study, the assumption of the RF algorithm is also satisfied.

In SVM, the only requirement is to have independent and identically distributed data. In this study, data is independent and identically distributed. The assumption of the SVM algorithm is also satisfied.

5. CONCLUSIONS AND RECOMMENDATIONS

In this study, transactional data-based customer segmentation study and characteristics are examined. As mentioned in the previous sections, various methods exist to create customer segments. Apart from the conventional methods, new methods have been invented and used in the last decades. In this study, RFMLP based customer segmentation with K-Means algorithm is used, and the importance of customer segments is created with the help of an analytical method. Besides, a customer churn analysis method based on RFMLP attributes was also conducted to predict which customers will churn in the future. Churn analysis is performed with four different supervised machine learning algorithms: k-nearest neighbors, logistic regression, random forest, and support vector machines.

It is crucial, especially for companies that are operating in the intensely competitive heavy equipment industry sector, to create and adjust strategies and create a long-run relationship with their customers. To solve these problems, it is essential to develop customer segments and understand the characteristics of each specific segment. In this context, this study recommends a new method called RFMLP which can be described as an augmented version of the classical RFM method to gain deeper and meaningful insights about customers. Customers are grouped into four different segments, and the relative importance score of each segment is calculated.

As explained in former sections of the study, finding new customers is comlex and more costly, thus attaining current customers to work with the company is beneficial for business. In this manner, it is vital to know which customers will end up their contracts to work with the company in the future. To handle this issue, customer churn analysis models were created with RFMLP variables. The random forest model outperforms others with the highest accuracy score and AUROC value.

Thanks to machine learning and analytical methods, all the results, and accuracy of models were examined with assumptions and provided meaningful insights. This study can provide former literature by proposing RFMLP based customer segmentation combined with churn analysis and providing meaningful insights into various customer types in the Turkish heavy equipment industry. Companies in the heavy equipment industry can utilize this study to identify different groups and profile them. With these efforts, a company can effectively manage and adjust its CRM and marketing strategies, and allocation of resources can be completed with higher

effectiveness. Besides, the quality of service level can be increased, and customer satisfaction can also be improved. Thus customers become highly satisfied, their loyalty is increased, and profit of the company is increased, and competitive advantage can be maintained. Customer churn can also be decreased because of correct strategies predicted as churners. Results of the churn analysis part make it possible for business professionals to decide and suitable ways to prevent churn of customers.

Despite several benefits of the study, it has some limitations. Firstly, customer transaction data of a company in Turkey was used. However, companies operating in different countries and customer transaction data of other countries can also be used. Besides, customer transaction data can also be separated according to geographical areas of Turkey in order to gain deeper insights at the regional level. Secondly, only one high-level brand company data was used. However, to validate results and ascertain findings, another brand data can also be used in the future. Mid-level and low-level brand companies' data can also be added to study in the future to see the effects of brand level on the results. Additionally, for future research, the model can be enhanced by adding new variables in the context of customer's behavior, such as the number of unique types of items purchased, types of items bought according to usage, customer's geographical location. Defined strategies and marketing programs can also be added to the end of the model to create an end-to-end approach.

REFERENCES

[1] Reinsel, D., Gantz, J., & Rydning, J. (2018). The Digitization of the World From Edge to Core. November. https://www.seagate.com/files/wwwcontent/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

[2] Cambra-Fierro, J., Gao, L. X., Melero-Polo, I., & Trifu, A. (2021). How do firms handle variability in customer experience? A dynamic approach to better understanding customer retention. *Journal of Retailing and Consumer Services*, *61*, 102578.

[3] Url-1 <*https://www.idc.com/getdoc.jsp?containerId=prUS45422819*>, date retrieved 17.02.2021

[4] Anitha, P., & Patil, M. M. (2019). RFM model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University-Computer and Information Sciences*.

[5] Christy, A. J., Umamakeswari, A., Priyatharsini, L., & Neyaa, A. (2018). RFM ranking–An effective approach to customer segmentation. *Journal of King Saud University-Computer and Information Sciences*.

[6] Mosaddegh, A., Albadvi, A., Sepehri, M. M., & Teimourpour, B. (2021). Dynamics of customer segments: A predictor of customer lifetime value. *Expert Systems With Applications*, *172*, 114606.

[7] De Caigny, A., Coussement, K., & De Bock, K. W. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *European Journal of Operational Research*, *269*(2), 760-772.

[8] Stripling, E., vanden Broucke, S., Antonio, K., Baesens, B., & Snoeck, M. (2018). Profit maximizing logistic model for customer churn prediction using genetic algorithms. *Swarm and Evolutionary Computation*, *40*, 116-130.

[9] Mannila, H. (1996, June). Data mining: machine learning, statistics, and databases. In *Proceedings of 8th International Conference on Scientific and Statistical Data Base Management* (pp. 2-9). IEEE.

[10] Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160(1), 3-24.

[11] Zhang, F., Ju, Y., Gonzalez, E. D. S., & Wang, A. (2020). SNA-based multi-criteria evaluation of multiple construction equipment: A case study of loaders selection. *Advanced Engineering Informatics*, 44, 101056.

[12] Petroutsatou, K., & Giannoulis, P. (2020). Analysis of construction machinery market: the case of Greece. *International Journal of Construction Management*, 1-8.

[13] Url-2 <*https://www.statista.com/statistics/280344/size-of-the-global-construction-machinery-market/*>, date retrieved 11.03.2021.

[14] Url-3 <*https://www.statista.com/statistics/257347/leading-machinery-manufacturers-worldwide/*>, date retrieved 14.05.2021.

[15] Url-4 <*https://imder.org.tr/tr/hakkimizda.html*>

[16] Sun, Z. H., Zuo, T. Y., Liang, D., Ming, X., Chen, Z., & Qiu, S. (2021). GPHC: A heuristic clustering method to customer segmentation. *Applied Soft Computing*, 111, 107677.

[17] Chen, H., Zhang, L., Chu, X., & Yan, B. (2019). Smartphone customer segmentation based on the usage pattern. *Advanced Engineering Informatics*, 42, 101000.

[18] Tsai, C. F., Hu, Y. H., & Lu, Y. H. (2015). Customer segmentation issues and strategies for an automobile dealership with two clustering techniques. *Expert Systems*, 32(1), 65-76.

[19] Güçdemir, H., & Selim, H. (2015). Integrating multi-criteria decision making and clustering for business customer segmentation. *Industrial Management & Data Systems*.

[20] Khalili-Damghani, K., Abdi, F., & Abolmakarem, S. (2018). Hybrid soft computing approach based on clustering, rule mining, and decision tree analysis for customer segmentation problem: Real case of customer-centric industries. *Applied Soft Computing*, 73, 816-828.

[21] Nie, F., Wang, X., Jordan, M., & Huang, H. (2016, March). The constrained laplacian rank algorithm for graph-based clustering. *In Proceedings of the AAAI conference on artificial intelligence* (Vol. 30, No. 1).

[22] Xu, J., Han, J., Xiong, K., & Nie, F. (2016, July). Robust and sparse fuzzy k-means clustering. *In IJCAI* (pp. 2224-2230).

[23] Nie, F., Wang, X., & Huang, H. (2014, August). Clustering and projected clustering with adaptive neighbors. *In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 977-986).

[24] Murray, P. W., Agard, B., & Barajas, M. A. (2015). Forecasting supply chain demand by clustering customers. *IFAC-PapersOnLine*, 48(3), 1834-1839.

[25] Mohammadzadeh, M., Hoseini, Z. Z., & Derafshi, H. (2017). A data mining approach for modeling churn behavior via RFM model in specialized clinics Case study: A public sector hospital in Tehran. *Procedia computer science*, 120, 23-30.

[26] Sano, N., Tsutsui, R., Yada, K., & Suzuki, T. (2016). Clustering of customer shopping paths in Japanese grocery stores. *Procedia Computer Science*, 96, 1314-1322.

[27] Wang, B., Miao, Y., Zhao, H., Jin, J., & Chen, Y. (2016). A biclustering-based method for market segmentation using customer pain points. *Engineering Applications of Artificial Intelligence*, 47, 101-109.

[28] Brito, P. Q., Soares, C., Almeida, S., Monte, A., & Byvoet, M. (2015). Customer segmentation in a large database of an online customized fashion business. *Robotics and Computer-Integrated Manufacturing*, 36, 93-100.

[29] Bose, I., & Chen, X. (2015). Detecting the migration of mobile service customers using fuzzy clustering. *Information & Management*, 52(2), 227-238.

[30] Llanos, J., Morales, R., Núñez, A., Sáez, D., Lacalle, M., Marín, L. G., ... & Lanas, F. (2017). Load estimation for microgrid planning based on a self-organizing map methodology. *Applied Soft Computing*, 53, 323-335.

[31] Ho, G. T., Ip, W. H., Lee, C. K., & Mou, W. L. (2012). Customer grouping for better resources allocation using GA based clustering technique. *Expert Systems with Applications*, 39(2), 1979-1987.

[32] Hong, T., & Kim, E. (2012). Segmenting customers in online stores based on factors that affect the customer's intention to purchase. *Expert Systems with Applications*, 39(2), 2127-2131.

[33] Seret, A., Maldonado, S., & Baesens, B. (2015). Identifying next relevant variables for segmentation by using feature selection approaches. *Expert systems with applications*, 42(15-16), 6255-6266.

[34] Li, D. C., Dai, W. L., & Tseng, W. T. (2011). A two-stage clustering method to analyze customer characteristics to build discriminative customer management: A case of textile manufacturing business. *Expert Systems with Applications*, 38(6), 7186-7191.

[35] Díaz, A., Gómez, M., Molina, A., & Santos, J. (2018). A segmentation study of cinema consumers based on values and lifestyle. *Journal of Retailing and Consumer Services*, 41, 79-89.

[36] Kuo, R. J., Mei, C. H., Zulvia, F. E., & Tsai, C. Y. (2016). An application of a metaheuristic algorithm-based clustering ensemble method to APP customer segmentation. *Neurocomputing*, 205, 116-129.

[37] Dursun, A., & Caber, M. (2016). Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis. *Tourism management perspectives*, 18, 153-160.

[38] Wei, J. T., Lin, S. Y., Weng, C. C., & Wu, H. H. (2012). A case study of applying LRFM model in market segmentation of a children's dental clinic. *Expert Systems with Applications*, 39(5), 5529-5533.

[39] Wei, J. T., Lee, M. C., Chen, H. K., & Wu, H. H. (2013). Customer relationship management in the hairdressing industry: An application of data mining techniques. *Expert Systems with Applications*, 40(18), 7513-7518.

[40] Cheng, C. H., & Chen, Y. S. (2009). Classifying the segmentation of customer value via RFM model and RS theory. *Expert systems with applications*, 36(3), 4176-4184.

[41] Chiu, C. Y., Chen, Y. F., Kuo, I. T., & Ku, H. C. (2009). An intelligent market segmentation system using k-means and particle swarm optimization. *Expert systems with applications*, 36(3), 4558-4565.

[42] Alsamaray, H. S. (2017). AHP as multi-criteria decision making technique, empirical study in cooperative learning at Gulf University. *European Scientific Journal*, ESJ, 13(13), 272-289.

[43] Bhadra, D., Dhar, N. R., & Salam, M. A. (2021). Sensitivity analysis of the integrated AHP-TOPSIS and CRITIC-TOPSIS method for selection of the natural fiber. *Materials Today: Proceedings*.

[44] Saaty, T. L. (1990). How to make a decision: the analytic hierarchy process. *European journal of operational research*, 48(1), 9-26.

[45] Rajput, V., Sahu, N. K., & Agrawal, A. (2021). Optimization of waste kota stone dust filled epoxy composite by analytical hierarchy process (AHP) approach. *Materials Today: Proceedings*.

[46] Url-5 *<https://www3.diism.unisi.it/~mocenni/Note_AHP.pdf>*, date retrieved 04.05.2021.

[47] Coussement, K., Van den Bossche, F. A., & De Bock, K. W. (2014). Data accuracy's impact on segmentation performance: Benchmarking RFM analysis, logistic regression, and decision trees. *Journal of Business Research*, 67(1), 2751-2758.

[48] Dursun, A., & Caber, M. (2016). Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis. *Tourism management perspectives*, 18, 153-160.

[49] Christy, A. J., Umamakeswari, A., Priyatharsini, L., & Neyaa, A. (2018). RFM ranking–An effective approach to customer segmentation. *Journal of King Saud University-Computer and Information Sciences*.

[50] Chang, H. H., & Tsay, S. F. (2004). Integrating of SOM and K-mean in data mining clustering: An empirical study of CRM and profitability evaluation.

[51] Yeh, I. C., Yang, K. J., & Ting, T. M. (2009). Knowledge discovery on RFM model using Bernoulli sequence. *Expert Systems with Applications*, 36(3), 5866-5871.

[52] Peker, S., Kocyigit, A., & Eren, P. E. (2017). LRFMP model for customer segmentation in the grocery retail industry: a case study. *Marketing Intelligence & Planning*.

[53] Stripling, E., vanden Broucke, S., Antonio, K., Baesens, B., & Snoeck, M. (2018). Profit maximizing logistic model for customer churn prediction using genetic algorithms. *Swarm and Evolutionary Computation*, 40, 116-130.

[54] Sun, Y. (2021). Case based models of the relationship between consumer resistance to innovation and customer churn. *Journal of Retailing and Consumer Services*, 61, 102530.

[55] Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert systems with applications*, 38(3), 2354-2364.

[56] Kozak, J., Kania, K., Juszczuk, P., & Mitręga, M. (2021). Swarm intelligence goal-oriented approach to data-driven innovation in customer churn management. *International Journal of Information Management*, 102357.

[57] Url-5 *https://www.expert.ai/blog/machine-learning-definition/*, date retrieved 07.15.2021.

[58] Url-6 <*https://www.datarobot.com/wiki/machine-learning/*>, date retrieved 07.5.2021.

[59] Url-7 <*https://www.datarobot.com/wiki/supervised-machine-learning/*>, date retrieved 12.07.2021

[60] Url-8 <*https://www.analyticsvidhya.com/blog/2020/04/supervised-learning-unsupervised-learning/*>, date retrieved 12.07.2021.

[61] Url-9 <*https://www.datarobot.com/wiki/unsupervised-machine-learning/*>, date retrieved 13.07.2021.

[62] Url-10 <*https://www.saedsayad.com/k_nearest_neighbors.htm*>, date retrieved 13.07.2021.

[63] Url-11 <*https://towardsdatascience.com/k-nearest-neighbours-introduction-to-machine-learning-algorithms-18e7ce3d802a*>, date retrieved 13.07.2021.

[64] Url-12 <*https://www.saedsayad.com/logistic_regression.htm*>, date retrieved 13.07.2021

[65] Url-13 <*https://www.saedsayad.com/support_vector_machine.htm*>, date retrieved 14.07.2021.

[66] Url-14 <*https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/*>, date retrieved 15.07.2021.

[67] Url-15 <*https://www.guru99.com/confusion-matrix-machine-learning-example.html*>, date retrieved 16.07.2021.

[68] Url-16 <*https://deepai.org/machine-learning-glossary-and-terms/precision-and-recall*>, date retrieved 16.07.2021.

[69] Url-17 <*https://www.saedsayad.com/clustering_kmeans.htm*>, date retrieved 17.07.2021.

[70] Url-18 <*https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a*>, date retrieved 18.07.2021.

APPENDICES

Appendix – A

RFMLP - Importance Selection

Comparison Matrix (A)

Criteria	Recency	Frequency	Monetary	Length	Periodicity
Recency	1,00	4,00	2,00	5,00	3,00
Frequency	0,25	1,00	3,00	3,00	9,00
Monetary	0,50	0,33	1,00	4,00	7,00
Length	0,20	0,33	0,25	1,00	9,00
Periodicity	0,33	0,11	0,14	0,11	1,00
SUM	2,28	5,78	6,39	13,11	29,00

CURRICULUM VITAE

Name Surname : Mustafa ÇAMLICA

EDUCATION :

- **B.Sc.** **:** 2018, Ozyegin University, Department of Indsutrial Engineering

PROFESSIONAL EXPERIENCE AND REWARDS:

- 2018 – 2021, Borusan-CAT, Data Analytics Specialist
- 2021 – Present, SHV Energy – Ipragaz A.Ş., Sales Analytics Executive

OTHER PUBLICATIONS, PRESENTATIONS AND PATENTS:

- **Ötken, Ç. N., Organ, Z. B., Yıldırım, E. C., Çamlıca, M., Cantürk, V. S., Duman, E., ... & Kayış, E.** (2019). An extension to the classical mean–variance portfolio optimization model. *The Engineering Economist*, *64*(3), 310-321.