KADİR HAS UNIVERSITY

SCHOOL OF GRADUATE STUDIES

PROGRAM OF COMPUTER ENGINEERING

LOCATION-ALLOCATION THROUGH MACHINE LEARNING FOR E-COMMERCE LOGISTIC SERVICES

TAYYİP TOPUZ

MASTER'S THESIS

İSTANBUL, FEBRUARY, 2022

LOCATION-ALLOCATION THROUGH MACHINE LEARNING FOR E-COMMERCE LOGISTIC SERVICES

TAYYİP TOPUZ

Assoc. Prof. Dr. Tamer DAĞ

MASTER'S THESIS

Submitted to the School of Graduate Studies of Kadir Has University in partial fulfillment of the requirements for the degrees of Master's in the Program of Computer Engineering

ISTANBUL, FEBRUARY, 2022

KADİR HAS UNIVERSITY

SCHOOL OF GRADUATE STUDIES

ACCEPTANCE AND APPROVAL

This work entitled **LOCATION-ALLOCATION THROUGH MACHINE LEARNING FOR E-COMMERCE LOGISTIC SERVICES** prepared by TAYYİP TOPUZ has been judged to be successful at the defense exam held on **17/02/2022** and accepted by our jury as **MASTER'S THESIS.**

APPROVED BY:

Assoc. Prof. Dr. Tamer DAĞ (Advisor) Kadir Has University _____

Prof. Dr. Funda SAMANLIOĞLU Kadir Has University _____

Assoc. Prof. Dr. Tansal GÜÇLÜOĞLU Yıldız Technical University_____

I certify that the above signatures belong to the faculty members named above

_____

Prof. Dr. Mehmet Timur AYDEMİR

Dean of School of Graduate Studies

DATE OF APPROVAL:17/02/2022

DECLARATION OF RESEARCH ETHICS AND PUBLISHING METHODS

I, TAYYİP TOPUZ, hereby declare that;

- this Master's Thesis is my own original work, and that due references have been appropriately provided on all supporting literature and resources;
- this Master's Thesis contains no material that has been submitted or accepted for a degree or diploma in any other educational institution;
- I have followed "Kadir Has University Academic Ethics Principles" prepared in accordance with the "The Council of Higher Education's Ethical Conduct Principles."

In addition, I understand that any false claim in respect of this work will result in disciplinary action in accordance with University regulations.

Furthermore, both printed and electronic copies of my work will be kept in Kadir Has Information Center under the following condition as indicated below:

☐ The full content of my thesis/project will be accessible from everywhere by all means.

TAYYİP TOPUZ

_____

17/02/2022

Dedicated to people who value and support education throughout their lives

ACKNOWLEDGEMENT

I would like to express my appreciation to all people who contributed to my thesis. First and foremost, I would like to thank Assoc. Prof. Dr. Tamer Dağ for his support and guidance throughout the thesis process.

LOCATION-ALLOCATION THROUGH MACHINE LEARNING FOR E-COMMERCE LOGISTIC SERVICES

ABSTRACT

Companies desire to expand their businesses in such a way that there will not be any loss in their revenues. An e-commerce logistics company functions as the distribution and delivery of goods to buyers. To expand the business, opening new branches is a critical decision since determining the location of a branch correctly will not only help an e-commerce logistics company to increase its revenue but also improve customer satisfaction. The logistic network, which is based on locations, is the most vital input for their business. For such decisions, data science is becoming an essential tool in recent years. Research shows that demographic information has a considerable impact on consumer behavior in e-commerce. In this thesis, the demand potential is studied by using demographic data and current demand for an e-commerce logistics company. The outcome of this work can be used to determine the location of new branches. Machine learning techniques are being used to decide the location of a new branch with the help of delivery demand potential prediction.

Keywords: Machine Learning, Location-allocation, E-commerce, Logistics, Geodemographic Analysis

E-TİCARET LOJİSTİK HİZMETLERİ İÇİN MAKİNE ÖĞRENİMİ YOLUYLA KONUM TAHSİSİ

ÖZET

Şirketler, gelirlerinde herhangi bir kayıp olmayacak şekilde işlerini büyütmek isterler. Bir e-ticaret lojistik şirketi, malların alıcılara dağıtımı ve teslimatı olarak işlev görür. İş ağlarını büyütmek için yeni şubeler açmak kritik bir karardır, çünkü bir şubenin yerini doğru belirlemek, bir e-ticaret lojistik şirketinin sadece gelirini artırmasına değil, aynı zamanda müşteri memnuniyetini de artırmasına yardımcı olacaktır. Lokasyonlara dayalı lojistik ağ, işletmeleri için en hayati girdidir. Bu tür kararlar için veri bilimi son yıllarda önemli bir araç haline geliyor. Araştırmalar, demografik bilgilerin e-ticarette tüketici davranışları üzerinde önemli bir etkiye sahip olduğunu göstermektedir. Bu çalışmada bir e-ticaret lojistik firması için demografik veriler ve mevcut talep kullanılarak talep potansiyeli incelenecektir. Bu çalışmanın sonucu yeni şubelerin yerini belirlemek için kullanılabilir. Teslimat talebi potansiyel tahminin yardımıyla yeni bir şubenin konumuna karar vermek için makine öğrenme teknikleri kullanılacaktır.

Anahtar Sözcükler:Makine Öğrenimi, Konum Tahsisi , E-ticaret, Lojistik, Demografik Analiz

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

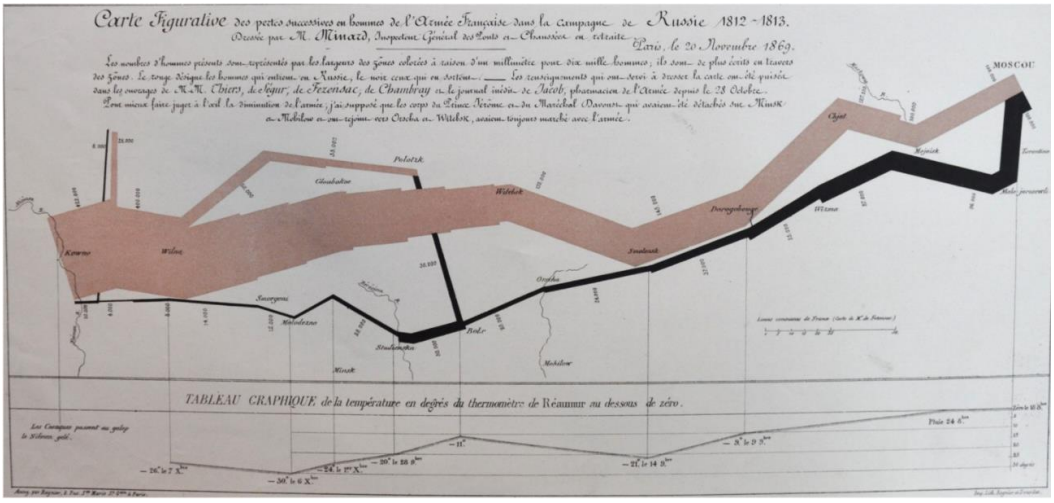
LIST OF ABBREVIATIONS

AHP	Analytic hierarchy process
BDDK	Banking Regulation and Supervision Agency
BTK	Information and Communication Technologies
COVID-19	Coronavirus disease pandemic
DBSCAN	Density-based Spatial Clustering of Applications with Noise
EV	Explained variance
FHWA	Federal Highway Administration
GIS	Geographic Information System
KNN	K-nearest neighbor
KNNR	K-nearest neighbor regression
LARS	Least Angle Regression
LR	Linear Regression
MAE	Mean absolute error
ML	Machine Learning
MSE	Mean squared error
RMSE	Root-mean-squared-error
RR	Ridge regression
RTD	Round Trip Distance
SEGE	Socio-Economic Development Rank
SGK	Social Security Institution
SVM	Support Vector Machine
SVR	Support Vector Regression
TSP	Traveling Salesperson Problem
TL	Turkish Lira
TURKSTAT	Turkish Statistical Institute
VRP	Vehicle Routing Problem

1. INTRODUCTION

Throughout history, logistics that is distributing materials from one location to another has been a challenging problem for civilizations. Logistic skills have shown their importance first in the military sector [1]. The most important skill was their infrastructure that caused civilizations to rise and fall. Soldiers must be well-fed if you want to continue their campaign at least. Thus, armies require their rations which logistic skills can supply. The foundation of human civilization was laid with the Roman Empire. One of the reasons why the Roman Empire rose and fell was infrastructure which affected the logistic skills directly. They used their skills to build roads and canals to make their movement more easy. Their logistic skills made it easy to manage their rulership. Towards the last years of the Roman Empire, they could not allocate a budget for the infrastructure, causing problems in logistics and speeding up the collapse of the empire. Another example is Napoleon's failed campaign to attempt to conquer Russia in 1812. There are many reasons why the campaign failed, but the logistics are the most crucial reason. When the army marched in the territory of Russia, they faced poor roads, which led to inadequate logistics skills of food and supplies. Minard's map was created by cartographer Charles Joseph Minard to demonstrate the numerical data for Napoleon's campaign of Russia shown in Figure 1.1 [2]. This map can be counted as one of the best ways to visualize data.

Figure 1.1: Minard's Map [2]



The increasing number of urban areas adds a new logistics challenge: speed and satisfaction. Customers want to obtain their demands as fast as possible. They will be satisfied if they get their orders within their desired time limit. In other words, customers satisfaction depends on the companies' service. Companies should boost their customer loyalties. The best way to increase customer loyalty in the logistic business is the speed of the delivery. Thus, how good the companies' logistic skills will determine their business success.

E-commerce business is rapidly increasing because products are available through the Internet in such a way customers do not need to leave their homes. The statistics show that retail e-commerce sales worldwide almost tripled their amounts in the last five years [3]. The availability of e-commerce increases the competition with other stakeholders. The most important and challenging is logistics in the e-commerce business. The research shows that the e-commerce service quality framework consists of two dimensions electronic and logistic service [4]. Y. Lin, J. Luo, S. Cai, S. Ma, and K. Rong, mentioned that there is no direct impact on customer loyalty because of customer satisfaction. However, If customers are satisfied with the logistics, the possibility of using the same e-commerce service will be increased.

Logistics is dependent on organizing the products to transport the end-users. The most important topic is the networks of branches. The products must be transported from warehouses to branches then to end-users. To efficiently distribute products to users, locating the branches is the most crucial decision on e-commerce logistics. Determining the location of the branch will affect both customer satisfaction and the revenue of the company. If a branch is located near customers, delivery time will be short. The amount of time spent on the road will be less, affecting the delivery time and fuel usage amount. Delivery time and company revenue have a negative correlation in between. When the average delivery time is increased, the income of the company will decrease.

Machine learning techniques can be used to make an important decision on the location-allocation problem. Machine learning techniques can make sophisticated decisions to solve problems. Data that people produce are rapidly increasing is the most critical

variable to have. Machine learning techniques are made possible to make decisions instead of human interference.

The problem is finding a new best location for a branch that will help increase the revenue and customer loyalty of the e-commerce logistic company. The aim of this thesis is to determine a location for a branch to be opened by using machine learning techniques with demographic data. The location-allocation problem is tried to be solved in two stages. In the first stage delivery demand potential of e-commerce, logistics tried to be predicted using several machine learning techniques. In the second stage, a location can be proposed to be opened by using another machine learning algorithm with the help of predicted delivery amounts. Thus, two problems, predicting the delivery demand potentials and finding the best location for a new branch, are solved by using geodemographic data.

The organization of the thesis is as follows. Chapter 2 provides information on what are the essential data which affects consumer behavior and how other researchers solve the problem of location-allocation. Chapter 3 gives information on Machine Learning and its different techniques. Chapter 4 gives information on the data set. It explains which features are used and how it is acquired and transformed to be used for Machine Learning techniques. Chapter 5 gives information on methods which are the steps of the Machine Learning techniques. Chapter 6 provides data analysis on the location-allocation problem. Results of Machine Learning techniques are compared in this chapter. Also, it proposes a new location for a branch. Chapter 7 gives final thoughts and a summarization of the thesis.

2. THE LOCATION ALLOCATION PROBLEM IN THE LOGISTICS SECTOR

This chapter focused on explaining two main topics, the characteristics of consumers when they purchased a good from e-commerce and the location-allocation problem in the logistics and some other sectors. There are some key characteristics features that affect the consumer's way of buying products. The location-allocation problem is a studied research topic that researchers proposed and compared many ways to overcome that problem.

Consumer behavior is how an identifiable consumer group makes buying decisions [5]. The most influential factor is the consumer behavior to purchase anything. Because of that, companies' primary focus is dividing their costumer into groups to make them buy their product. Consumer behavior is affected by five psychological, social, cultural, personal, and economic factors. According to these factors, demographics, data relating to the population and different groups are critical factors for consumer behavior. Many studies show that demographic information affects shopping online.

M. Figliozzi and A. Unnikrishnan explore the impact of socio-demographic characteristics, health concerns, and product type on home delivery rates and expenditures during a strict coronavirus disease pandemic (COVID-19) lockdown period in [6]. The study finds that demographic information is essential for home delivery rates, such as age, gender, income, disability, and household size. Age has a positive relationship with home delivery rates. Adults use home delivery services more than the young population because of their health concerns. Gender is another factor that affects home delivery services. Study shows that male has fewer concerns for their health. Because of that male gender has a negative relationship with the rates.

According to a recent brief in the U.S. Department of Transportation Federal Highway Administration, demographic information affects online shopping behavior such as household, location, and income type [7]. Federal Highway Administration (FHWA) survey to identify the online shopping trends. Household types are one or more adults

living with children, no children, the youngest child 0-5, 6-15, 16-21 years old, and one or more adults living with retired. FHWA reported that one or more adults living with children 0-5 years have slightly had more percentage who made at least one online purchase in the last 30 days. Thus, age is one of the influential factors for online consumer behavior. Location types which are urban and rural, have effects on online shopping. Most of the purchases are made by living in urban locations. The growth in online shopping from 2009 to 2017 was more remarkable for respondents living in urban areas than their rural counterparts. Another finding is that respondents in households above the poverty line were almost twice as likely to make online purchases compared to respondents in households below the poverty level.

J. Hou, and K Elliott study a research on mobile shopping intensity by using consumer demographics and motivations in [8]. Like other researchers, they try to find some demographic information that affects online consumer behavior. This demographic data contains gender, age, education, and income information. They observe that gender and age have no significant effect on the frequency of purchases. However, younger people tend to use mobile shopping more than older people.

J. Chacón-García suggests using the analytic hierarchy process (AHP) to determine the location of retail companies in regulated markets, which is a pharmacy [9]. The research uses real-world data from the city of Seville (Spain). The AHP method is applied to multicriteria decision problems to determine the advantages of the criteria. The data consist of population density, greater than 64 years old's population, and median household income. Past researches indicate that most drug consumers' age is over 64 in Spain. The AHP method supported previous studies and showed that the most important factor is over the age of 64.

S. S. Noorian proposes to create a new decision support system for sales territory planning which proposes to use a genetic algorithm to find the best solution as an efficient and generic method in [10]. The case study of the thesis is applied to an agricultural machinery manufacturer who is selling a specific product such as a tractor in Germany. Even thesis' aim focus is not location-allocation, they suggest to use for the location-allocation problem also. Genetic algorithms mimic the process of natural selection to solve

optimization problems. The proposed method uses the dataset, which is consists of the number of farms as the available potential inside each municipality. They suggest using more inputs: percentage of agricultural land, number of farmers, purchasing power, etc., to make more complicated the calculation of the available potential. The thesis claims that the new decision support system is sufficient to find the best locations for sales associates.

Y. Yanga, J. Tangb, H. Luoc, and R. Lawd, evaluates the hotel locations using machine learning tools [11]. They evaluate five different machine learning algorithms: linear regression, projection pursuit regression, artificial neural network, support vector regression, and boosted regression. The dataset consists of two different variables, including location attributes and individual hotel characteristics. Location attributes are the number of restaurants in a radius of 800 meters, road density in a two-kilometer radius, and distance to the nearest subway station. Individual hotel characteristics include star rating, the total number of beds, and the number of service years. Their research concluded that the most effective method was the simple linear regression model to predict occupancy rate.

In [12], M. A. Salazar-Aguilar, J. L. González-Velarde, and R. Z. Ríos-Mercado proposed using a divide-and-conquer approach to commercial territory design. Model is applied in a beverage distribution firm in Monterrey, Mexico. The model's approach is the simple divide-and-conquer approach, which divides the territories into sub territories. They focus on dividing territories to balanced, which means similar area sizes and numbers of the customer.

In [13], S. Noorian, A. Psyllidis, and A. Bozzon approach the problem with the location-allocation analysis is making an improvement on the p-median problem with a time variable. The P-median problem is determining the demand points that will receive service from the number of p facilities on the network consisting of n nodes with the minimum cost [14]. A time-varying p model suggests that the model should account for fluctuations in travel cost distance at different time intervals. Model is used on facilities that supply goods for restaurants. Model collects data from Google Traffic and Foursquare. It retrieves traffic information data from Google Traffic at different times of

the day and on various days of the week. Foursquare data is collected to estimate the demand fır service in an area. The model suggests adding some parameters to the classical p-median model *k,* which is potential departure time slots for a facility to serve a demand area and travel time from demand area *i* to facility *j* at departure time *k* where *i* is demand areas, *j* is candidate facility sites. The results show that the model outperforms the classical p-median problem formulation.

E. Olivares-Benitez, M. B. Bernábe-Loranca, S Caballero-Morales, R. Granillo-Macias suggests determining balanced sales territories with the tabu search method in [15]. The model is not selecting the locations of a facility directly, but it defines the sale territories, which can be considered the facilities service area. The model tries to balance sales, workload, and geography. Collected data is acquired from a company that sells hand tools and building materials to hardware stores in Mexico. They propose a mathematical model with three objective functions for designing balanced sale territories. The non-linearity of the model motivated the design of a metaheuristic algorithm based on Tabu Search to solve the problem. The results show that all objectives that balance the sales, workload, and geography had improvements between 40% and 50% against the company's current state.

H. Hsieh, F. Lin, C. Li, I. E. Yen, and H. Chen proposed the location-allocation problem is making temporal popularity prediction of locations for the geographical placement of retail stores in [16]. They propose an affinity-based popularity inference model using Foursquare check-in information data. The model is developed in three steps. The first step is determining the popularity distributions of unlabeled nodes from the popularity distributions of labeled nodes. In the second step, nodes with similar features are assumed to have similar popularity distributions. In the last step, the popularity distribution of an unlabeled node can be deducted in terms geographically and temporally. The proposed model is compared with several machine learning techniques. The results show that the proposed model outperforms all of the compared models.

Instead of deciding the facility's exact location, territories can be designed to find the service regions. S. Moreno, J. Pereira, and W. Yushimito, suggested designing commercial territory to improve the K-means algorithm with the integer programming

method in [17]. The proposed models are applied in a case study in meat distribution. The research suggests two modified versions of the K-means algorithm. Both models follow the method of Çavdar's continuous approximation [18]. The first version increases the number of K of clusters until all territories are feasible. The second one pushes aside any feasible territory and repeats the clustering process until all clients have been assigned to the feasible territory. Then both models are combined with an integer programming model that minimizes the number of the territories required to cover all of the clients. The results show that the modified version of the K-means algorithm with the inter programming methods are outperformed other proposed models. The best model is the second version of the K-means algorithm with the integer programming method.

3. MACHINE LEARNING AND OPTIMIZATION

3.1. Introduction To Machine Learning Techniques

Machine Learning (ML) which uses a vast amount of data to improve predictions, is a branch of artificial intelligence studies. ML is a form of artificial intelligence that enables computer systems to automatically learn and refine without being precisely programmed to do so [19]. It can learn from data, identify patterns and make a decision with minimum human intervention. ML techniques teach models from past experiences, which is a set of data. The main idea of developing a model is to divide the data into two parts for training and testing. First, the model is trained by an algorithm with the training data set. After training the model, an evaluation must be made to control the validity of the model. Thus, it can be decided the usability of the model. ML algorithms can be classified into three main branches, which are supervised, unsupervised, reinforcement.

The mathematical background of machine learning dates back to the 1760s with Bayes' theorem. Naive Bayes classifier, which is based on Bayes' theorem, is the most commonly used algorithm to solve classification problems by following the probabilistic approach. In the 1800s, Adrien-Marie Legendre and Carl Friedrich Gauss used least squares, the method of regression, to calculate planetary orbits. Regression methods are used in various disciplines. It tries to find the relationship between one dependent variable with independent variables, which are sometimes called features. In the 1920s, the modern-day computers' concept was introduced by Alan Turing. In the 1940s, Warren McCulloch and Walter Pitts proposed the idea of neural networks, which cover neurons in a network. Neurons have three primary purposes, which are receiving inputs, processing inputs, and generating outputs. In the 1960s, applications of evolutionary programming were applied first in the company called Decision Science Inc, formed by Lawrence J. Fogel. In the 1970s, advanced database management allowed storage and query terabytes and petabytes of data. In the 1990s, the term data mining was started to be used in data science communities. Data is the most valuable asset to solve any problems. Today machine learning is used in diverse business sectors.

ML methods are used in various areas such as banking, advertising, security, health, logistic, etc. Even though we did not finish converting our industrial revolution from 3.0 to 4.0, we produced too much data. After the Covid-19 pandemic, we generate data much faster than the previous years. For example, six million people shopped online every minute in 2021 [20]. Statistics show us data science will be the primary issue of our future because everything we do is recorded as a digital fingerprint. So that companies try to use those digital fingerprints to their advantage. Their first aim is to make their business more efficient. It is impossible to compute those data in the human brain. ML methods solve these problems without human intervention. Thus it is possible to make their business more efficient with the help of the ML.

3.2. Machine Learning Techniques

Before explaining the techniques, two important terms, overfitting and underfitting, must be defined. “Overfitting occurs when an algorithm reduces error through memorization of training examples with noisy or irrelevant features rather than learning the true general relationship between input and output. Underfitting occurs when an algorithm lacks sufficient model capacity or sufficient training to fully learn the true relationship, whether through memorization or not” [21]. Both cases, which will affect the ML models' reliability, are important problems in Machine Learning Techniques.

3.2.1. Supervised learning

“Supervised learning occurs when the training input has been explicitly labeled with the classes to be learned” [22]. Supervised learning uses the labeled data to feed the model to make predictions or classifications. The data includes inputs and outputs, which allow the model to learn. Supervised learning is divided into two categories which are classification and regression. Classification is the process of categorizing, recognizing, and understanding into defined classes. Regression allows the prediction of continuous output. In other words, the classification algorithms use qualitative data, and the regression algorithms use quantitative data.

Regression

"A tool for numerical data analysis that summarizes the relationship among the variables in a data set as an equation or the dependent variable is expressed as a function of one or several explanatory variables" [23]. There are several regression analysis algorithms to overcome the problem.

Linear regression

Linear regression (LR) predicts the dependent variable, which is sometimes called output or target, with the help of a linear equation. In simple LR, a bivariate model is built to predict output (y) from an explanatory variable (x) [24]. In multiple LR, the model is extended to include more than one explanatory variable ($x_0, x_1, \dots x_n$), producing a multivariate model. Correlation between variables is the essential factor in this model. Because of that, independent variables, which are sometimes called features, must have a relationship with the dependent variable to make a meaningful prediction. Here is the equation of multiple linear regression:

$$y = b_0 x_0 + b_1 x_1 + \cdots + b_n x_n = \sum_{i=0}^{n} b_i x_i \tag{2.1}$$

Where;

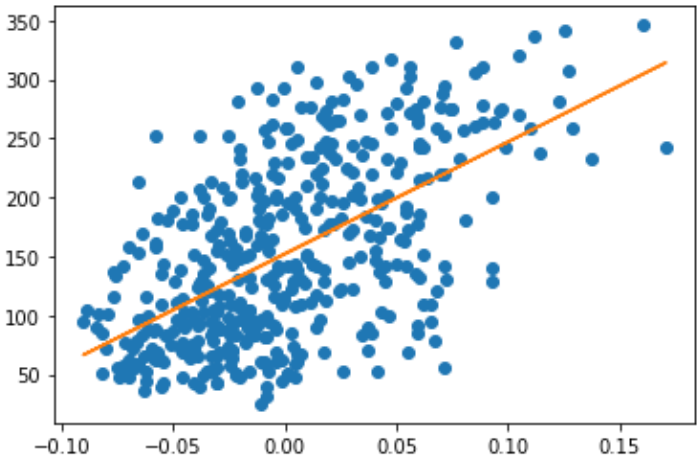
x_n : features of the data set

b_n : coefficients of the features

y : target label

A simple linear regression can be seen in Figure 3.2.1.1. A straight line is drawn with coefficients of the equation by given input and output. The primary purpose of the LR algorithm is the find the best coefficients for the equation to get closer to target label y.

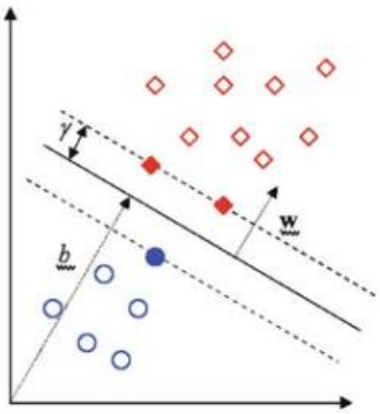
Figure 3.2.1.1: Simple Linear Regression



Support vector regression

Support Vector Machine (SVM), which solves binary classification problems developed by Vladimir Vapnik in 1988, transformed into a continuous variable output for regression analysis [25]. Regression is much difficult than classification because the problem's output has infinite possibilities. The model tries to estimate hyperplane, the decision boundary that helps classify the data point. SVR uses ε input to calculate a tube around the function region for optimizing the solution. The main focus of the SVR is finding the flattest tube that contains most of the training data. Outside the boundary of this tube, the hyperplane represented in terms of support vectors exists to make predictions on labeled output. A simple SVM is demonstrated in Figure 3.2.1.2 [26]. The figure line of the hyperplane is shown as straight, but also it can be a curved line.

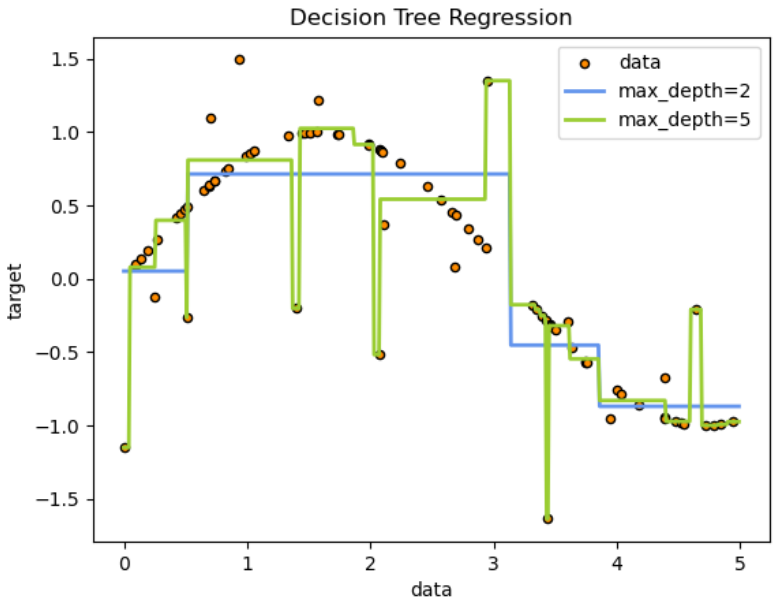
Figure 3.2.1.2: Support Vector Machine [26]



Decision tree regression

The decision tree regression (DTR) algorithm tries to create a tree to classify the data by the partitioning method. “DTR is based on a multistage or hierarchical decision scheme or a tree-like structure” [27]. The tree is built by decision nodes and leaf nodes. The decision node defines the data which is partitioned. Leaf node represents the decision on the numerical data. The main issue of the DTR algorithm is deciding the depth of the tree, which is the maximal length of a path from the root node to the leaf node. If a tree is too short, that means your model can be having problems predicting the target. If a tree is too long, that means your model can be overfitting which is your model is memorizing every single data point. The effects of the maximum depth can be seen in Figure 3.2.1.3 [28]. The most critical input for the DTR algorithm is the maximum depth of the tree.

Figure 3.2.1.3: Decision Tree Regression [28]

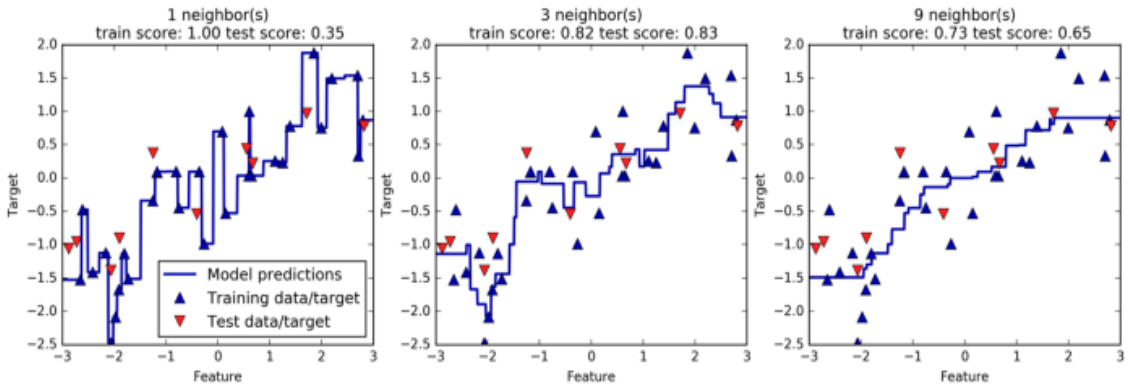


K-nearest neighbors regression

“K-nearest neighbors (KNN) algorithm is a method for classifying objects based on the closest training examples in the feature space” [29]. When the data labels are continuous rather than discrete, neighbor-based regression can be used to make predictions. Thus,

the same method can be used for regression by simply assigning the property value for the object to be the average of the values of its K nearest neighbors. Labels predicted based on neighbors' likelihood. Nearest neighbors regression has two approaches to train the models. It can learn based on k-nearest neighbors or fixed radius r. In the K-nearest neighbor model, k is the number of nearest neighbors. It learns from the nearest k neighbors. In the fixed radius r regression, the neighbors within the fixed radius are included for training the model. The model's complexity increases when the number of independent features increases. An example of the KNN model can be seen in Figure 3.2.1.4. In this example, the approach of KNN is that it learns from k-nearest neighbors instead of fixed-radius. The number of neighbors is crucial to decide to choose because it affects the model's reliability. Incorrectly assigned number of k-nearest neighbors can be cause overfitting or underfitting problem.

Figure 3.2.1.4: K-nearest neighbors [30]



Least-angle regression

“Least Angle Regression (LARS) relates to the classic model-selection method, Forward Selection, or forward stepwise regression” [31]. Least-angle regression is used for high dimensional data in linear regression models. The main approach of the LARS model is determining correlations between features and targeted label to find the largest absolute correlation. The process is repeated in each step to find the most correlated features with the targeted label and perform simple linear regression. Each selected predictor is projected to other predictors orthogonally in each selection process. After the selection process, predictors results are used for constructing multiple linear regression. When the

algorithm encounters the same correlated features, the least-angle regression model averages the attributes and proceeds in a direction at the same angle to the attributes. It is mainly used when the number of features is greater than the number of instances.

Ridge regression

Hoerl and Kennard propose ridge regression to overcome the ordinary least squares estimator problem for achieving better predictions [32, 33] Bühlmann and Yu suggest boosting the ridge regression algorithm in the context of linear regression with the focus on L2 loss [34]. Ridge regression methods are used when the data suffers from multicollinearity, which means that the features are highly correlated with each other. It estimates the coefficients of multiple regression models. It uses the L2 regularization to penalize the model. The model is penalized when the model has overfitted in the training stage. L2 regularization is added L2 penalty, equal to the square of the magnitude of coefficients.

3.2.2. Unsupervised learning

"Unsupervised learning is the opposite of supervised learning. In unsupervised learning, the machine simply receives inputs but obtains neither target outputs nor rewards from its environment" [35]. Instead, it analyzes data to discover hidden patterns or data groupings. Therefore, unsupervised learning algorithms try to convert the inputs into labeled outputs.

Clustering

"Clustering groups data instances into subsets in such a manner that similar instances are grouped together while different instances belong to different groups" [36]. There are several clustering methods to group similar data points, such as Connectivity-based, centroid-based, distribution-based, density-based, and grid-based clustering.

K-means clustering

"K-means clustering finds a collection of data points aggregated together because of certain similarities. K-means is a centroid-based clustering algorithm. This algorithm partitions the data into k clusters, represented by their centers or means" [36]. It calculates the number of k targeted clusters' centroids. Centroid is the mean value of the clustered data. Inputs of the algorithms, which are starting point of the centroids, the number of clusters, and iterations, determine the model's efficiency.

$$J = \sum_{j=1}^{K} \sum_{n \in S_j} |x_n - \mu_j|^2 \tag{2.2}$$

Where;

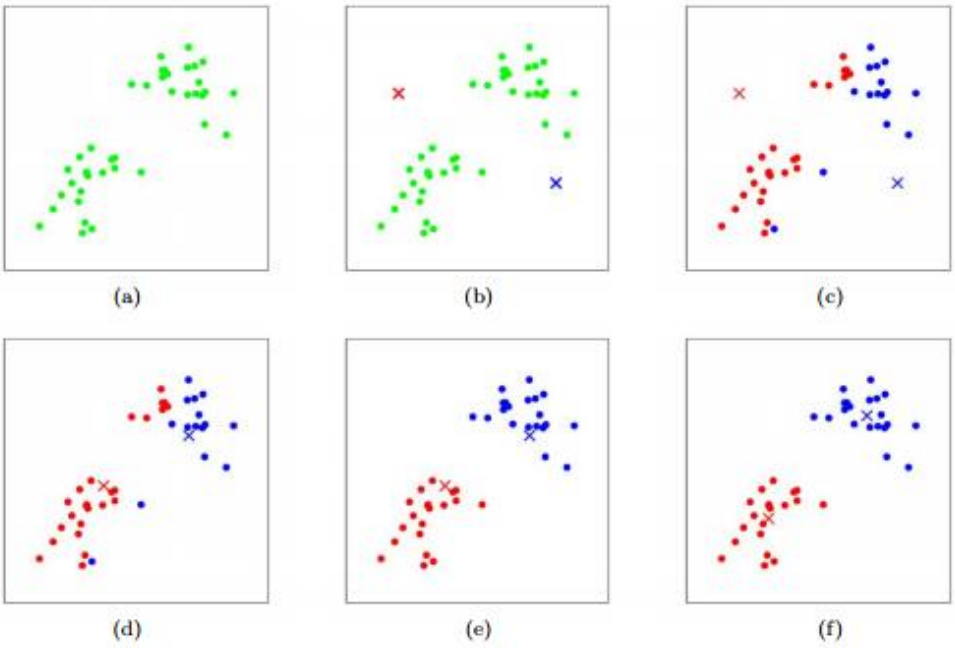
x_n : vector representing the nth data point

μ_j : geometric centroid of the data points in S_j

The first step of the algorithm is choosing the k number of centroids. There are methods to select the location of the centroids. The second step is assigning the data points to the closest centroids. The third step calculates the new centroids' location by assigned data points. Data points will be assigned to new centroid locations. The steps will continue until the stopping criteria are met. There are different stopping criteria, such as choosing a finite number of iterations or iterating until there is no change in the centroid assignment data points.

An example of the K-Means model can be seen in Figure 3.2.2.1 [37]. (a) shows the data points from the data set. (b) the randomly selected initial location of the cluster centroids. (c-f) illustration of running two iterations of k-means. Each data point is assigned to the new closest cluster centroids in each iteration. The number of maximum iteration can be defined because cluster centroids may not converge.

Figure 3.2.2.1: K-Means [37]

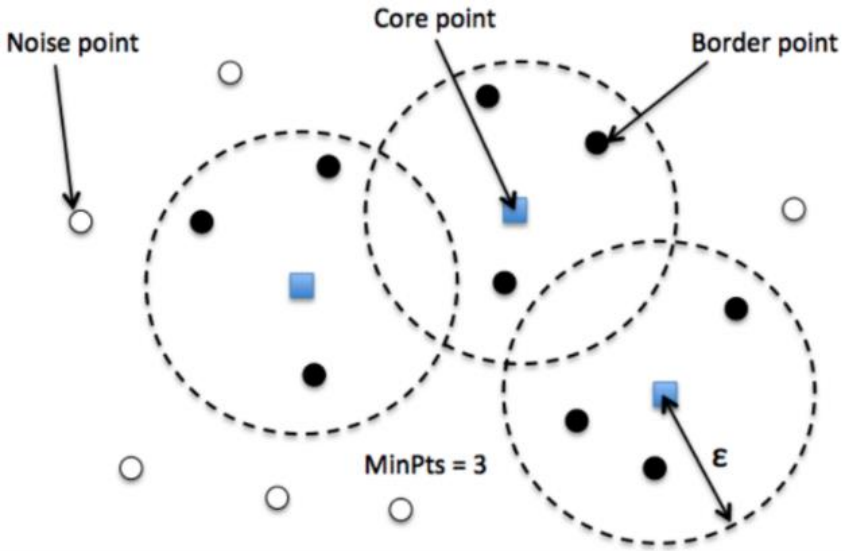


DBSCAN

DBSCAN (Density-based Spatial Clustering of Applications with Noise) algorithm which is designed to discover the clusters and the noise in a spatial database, estimates the data point's clusters by separating between the high-density and low-density areas [38]. As the name suggests, DBSCAN is a density-based clustering algorithm. Unlike K-means algorithms which define the shapes as convex, the shape of the clusters can be any shape. Algorithm work based on core points distance. Core points are the data points closest to their minimum number of neighbors within the radius. Radius is defined as epsilon, the maximum distance between the neighbor points. DBSCAN uses the distance between the individuals' data points. The default distance is the euclidean distance. The number of clusters is not defined before teaching the model, unlike K-means, which is defined before teaching the model. The number of clusters is defined by calculating the distance between the core points. Thus, DBSCAN is sensitive to the minimum points and the epsilon inputs. It affects the estimated number of clusters. For example, it can build more clusters or a single cluster because of the given minimum points and epsilon. So, choosing the best inputs is the most critical point in the DBSCAN algorithm. Minimum points are determined by the domain experts. There are several ways to assess the epsilon

variable. KNN distance plot can be used to determine the epsilon. Another way of using extensions of the DBSCAN algorithm, such as OPTICS, does not to use the epsilon input to estimate the clusters. Examples of core point, border point, nose point, and epsilon can be seen in Figure 3.2.2.2 [39].

Figure 3.2.2.2: DBSCAN [39]



OPTICS

OPTICS (Ordering Points To Identify the Clustering Structure) extends the DBSCAN algorithm [40]. As the DBSCAN, OPTICS is a density-based clustering algorithm. It changes some properties of the DBSCAN algorithm. Instead of using epsilon requirement, which defines the distance between the data points, it uses a value range to estimate clusters. The only difference between DBSCAN and OPTICS is that OPTICS does not assign cluster memberships. Instead, it stores the order in which the objects are processed and the information which an extended DBSCAN algorithm would use to assign cluster memberships. Core points are chosen by the minimum number of points in the radius range. It provides a new term, reachability distance, between a point and a core point. The points are not included if the reachability distance is too long.

3.3. Normalization

Data normalization, “where the data is scaled to uniformity, is needed to study the best features of the data” [41]. Normalization is one of the crucial processes of machine learning techniques. It has overcome the problem of inconsistencies.

3.3.1. Min-Max

Min-Max normalization is a simple method that can precisely fit the data in a pre-defined boundary [42] .

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} * (D - C) + C \tag{2.3}$$

Where;

x : data point

D: Maximum boundary

C: Minumum boundary

x' : normalized data point

3.3.2. Z score

“Z transformation is a transformation technique where data with different levels and spread is adjusted to a standard level and spread” [41]. Z normalization transforms the data in such a way the mean of the data will be zero, and the standard deviation will be one.

$$x' = \frac{(x - \mu)}{\sigma} \tag{2.4}$$

Where;

x : data point

μ: mean value

σ: standard deviation

x' : normalized data point

3.3.3. Logarithmic

"Log transformations that may restore the normality of the data are used when data is of varied and wide ranges" [41].

$$x' = \log x \tag{2.5}$$

Where;

x : data point

x' : normalized data point

3.4. Cross-Validation

Cross-validation is a statistical method of estimating the reliability of machine learning models. This method estimates the errors by applying the test sample from the joint distribution to output [43]. Evaluating machine learning models are another crucial part of the machine learning tasks. The reliability of machine learning models can be determined with several methods. Cross-validation is the most commonly used evaluation method for measuring reliability. There are two main types of cross-validation methods: non-exhaustive and exhaustive cross-validation.

3.4.1. Non-exhaustive cross-validation

Non-exhaustive cross-validation methods do not compute all ways of splitting the original sample.

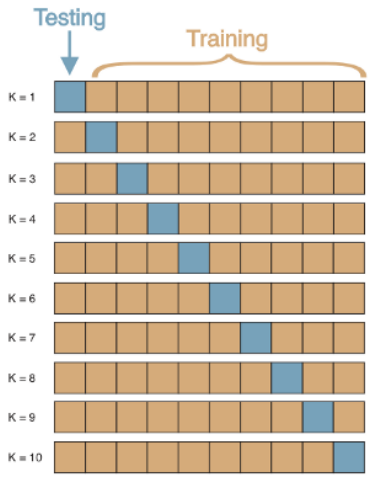
Holdout

The dataset is divided into two data sets in the holdout method: training and testing. It is a simple approach to test the model's reliability. The ratio of these data sets is essential in this method. The most used ratio is 70% of the data set is for training, remaining 30% of the data set is for testing. Another used ratio is 80% of the data set is for training, remaining 20% of the data set is for testing. If the testing data set is partitioned to a small percentage, it can cause an overfitting problem when the model trains with details and noise.

K-fold cross-validation

K-fold cross-validation is an improved method of the holdout. The k-fold cross-validation method partitions data sets into k equal size subsamples. The models' training and testing data sets iterate over those k equal size partitioned subsamples. In each iteration, one partitioned subsample is used as testing data set, others for training data set. An illustration of the k-fold cross-validation can be seen in Figure 3.4.1.1 [44].

Figure 3.4.1.1: K-fold Cross-Validation [44]



3.4.2. Exhaustive cross-validation

Exhaustive cross-validation methods test all possible ways to divide the original sample into training and validation sets. Leave-one-out and leave-p-out are examples of exhaustive cross-validation methods.

Leave-one-out

Leave-one-out is the special case for K-fold cross-validation, where the number of folds equals the number of instances in the data set. One of the instances leaves the data set for validation in each iteration in the testing data set. This is the reason why it is an exhaustive cross-validation method.

Leave-p-out

Leave-p-out cross-validation used p observations as the validation data set, and the remaining data set will be the training data set. Iteration is done until all instances are used in the p observations once.

3.5. Model Evaluation

Model evaluation metrics are used to decide the model's reliability. After training and testing the model, evaluation metrics must be calculated to observe the model's efficiency. There are several evaluation metrics to decide that model can be used on the data set.

3.5.1. Explained variance

Explained variance (EV) compares the variance within the expected outcomes and compares that to the variance in the model's error.

$$EV = 1 - \frac{Var(Y - \hat{Y})}{y_{true}} \tag{2.6}$$

Where;

Y: actual output

$\hat{Y}$: predicted output

3.5.2. Mean squared error

Mean squared error (MSE) calculates the average of squared differences between the predicted output and the actual output. “If the model eventually outputs a single very bad prediction, the squaring part of the function magnifies the error” [45]. In other words, MSE tells how close a fitted line is to actual output.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \tag{2.7}$$

Where;

Y: actual output

$\hat{Y}$: predicted output

n: number of samples

3.5.3. Root-mean-squared error

Root-mean-squared error (RMSE) is the error rate by square root of the MSE. “An ordering of regression models based on MSE will be identical to an ordering of models based on RMSE” [45]. RMSE has the same units as the actual output.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2} \tag{2.8}$$

Where;

Y: actual output

$\hat{Y}$: predicted output

n: number of samples

3.5.4. Mean absolute error

"The mean absolute error (MAE) of a model with respect to a test set is the mean of the absolute values of the individual prediction errors over all instances in the test set. Each prediction error is the difference between the true and predicted values" [45]. MAE is not penalizing too much the training outliers; because of that, it provides a generic performance measure for the model.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{Y}_i - Y_i\right| \tag{2.9}$$

Where;

Y: actual output

$\hat{Y}$: predicted output

n: number of samples

3.5.5. R-squared

R-squared is a statistical measure of fit that indicates how much variation of a dependent variable is explained by the independent variable(s) in a regression model. There is a particular case when R-squared is less than zero. "This case is only possible with linear regression when either the intercept or the slope are constrained so that the best-fit line fits worse than a horizontal line, for instance, if the regression line does not follow the data" [45].

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \quad (2.10)$$

Where;

Y: actual output

$\hat{Y}$: predicted output

$\bar{Y}$: mean of the actual data points

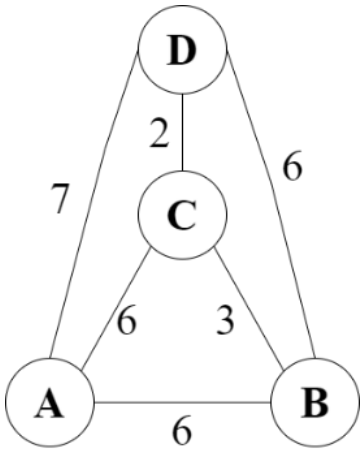
n: number of samples

3.6. Traveling Salesperson Problem Optimization

Vehicle Routing Problem (VRP), which finds the best routes for a fleet of vehicles visiting a set of locations, is one of the essential optimization applications. The best route can be described as the least total distance or cost. Traveling Salesperson Problem (TSP) tries to find the shortest route for a salesperson who needs to visit customers at different locations and return to the starting point [46]. TSP is a special case for VRP, which is that there is a single vehicle to travel all the customer's location points. An example of TSP can be seen in Figure 3.6.1. For starting point D, the best route is DABCD, with a total distance of 18. The amount of routes increases factorially with the number of points to visit, not including starting point. In the figure, the map includes six routes. The number of routes is 24 for 5 and 362880 for 10 points to cover. There are several solution methods for TSP.

There are two main concepts: exact and heuristic methods to solve a TSP. Exact solutions are guaranteed to find the optimal solution for the problem. On the contrary, heuristic solutions do not guarantee to find the optimal solution all the time. Still, one of the reasons to use heuristic solutions is the running time of the algorithms, which is reduced with the heuristic solutions.

Figure 3.6.1: TSP Example



3.6.1. Exact solutions

Cutting planes

The main approach of the cutting plane algorithm is to cut off parts of the feasible region of the linear programming relaxation. Optimal integer solution becomes an extreme point and therefore can be found by the simplex method. The main challenges of cutting plane algorithms are to find valid cuts, to find cuts that will quickly lead to the optimal integer solution, and to find a method for generating cuts that are guaranteed to terminate.

Branch-and-bound

The branch-and-bound algorithms use a branch-and-bound tree to optimize TSP routes. Every node of the tree contains five features, a node number, a label that represents the decision made at that node either to take or not to take a specific link from one city to another, a bound which gives a lower limit on the possible lengths of circuits below that node in the tree, an incoming matrix, and an opportunity matrix. Nodes are added until the unexplored node is left.

Branch-and-cut

The branch-and-cut algorithm is the combination of branch-and-bound and cutting planes algorithms. E. V. Dijck explains the branch-and-cut in [47].

> The first step of branch-and-cut algorithm is to initialize a linear programming relaxation of the original problem. Until no more valid inequalities can be found anymore, cutting plane procedure is used. In every node, the new linear programming relaxation of a problem is solved. If the solution of the relaxed problem is higher than the best solution found for the original problem, this means there is no improvement in this branch of the tree and the node is cut off. Otherwise, the procedure of branching, solving and looking for valid inequalities is repeated.

Brute force

The brute force method is one of the easiest ones in exact solutions. It calculates every possible permutation of the routes then chooses the shortest one from those permutations. Because of examining every possibility, the algorithms run slower than other algorithms.

3.6.2. Heuristic solutions

The most significant difference between the methods is their complexities. The complexity is demonstrated with big O notation, which is used to classify algorithms according to how their run time or space requirements grow.

Nearest insertion

The starting node is selected for the first step of the algorithm. Then, it finds a node with a minimal route and inserts the node between starting and ending nodes; in this case, starting and ending nodes are the same. Then connects a new node with the shortest path in any node from the sub-tour. These steps are repeated until all nodes are inserted into the tour. The complexity of the nearest insertion algorithm is $O(n^2)$ in big O notation.

Cheapest insertion

The starting node is selected for the first step of the algorithm. Then, it finds a node with a minimal route and inserts the node between starting and ending nodes; in this case, starting and ending nodes are the same. Then, it finds the next minimal route between starting and the next node without the previously inserted node. The constraint of the minimal route is that the route starting, previous, and the next node must be minimal from all the other nodes. These steps are repeated until all nodes are inserted into the tour. The complexity of the nearest insertion algorithm is $O(n^2 \log_2 n)$ in big O notation.

Arbitrary insertion

The starting node is selected for the first step of the algorithm. Then, it finds a node with a minimal route and inserts the node between starting and ending nodes; in this case, starting and ending nodes are the same. Then, it selects a new node arbitrarily to insert the node in a sub-tour with the minimal route. These steps are repeated until all nodes are inserted into the tour. The complexity of the nearest insertion algorithm is $O(n^2)$ in big O notation.

Farthest insertion

The starting node is selected for the first step of the algorithm. Then, it finds a node with a maximal route and inserts the node between starting and ending nodes; in this case, starting and ending nodes are the same. Then, it finds the next maximal route between starting and the next node with any of the nodes in sub-tour. Then, it inserts the next node, which minimizes the tour. These steps are repeated until all nodes are inserted into the tour. The complexity of the nearest insertion algorithm is $O(n^2)$ in big O notation.

Convex Hull

The algorithm forms a convex hull of a set of nodes and makes it an initial sub-tour. For each node, finds the cheapest route, then select the least cost node to insert. These steps

are repeated until all nodes are inserted into the tour. The complexity of the nearest insertion algorithm is $O(n^2 \log_2 n)$ in big O notation.

Lin-Kernighan

The Lin-Kernighan algorithm is based on exchanges. The algorithm exchanges the edges to reduce the current tour of the TSP until there is no change in the tour. “The Lin-Kernighan algorithm performs *k*-opt move on tours. A k-opt move changes a tour by replacing k edges from the tour by k edges in such a way that a shorter tour is achieved” [48].

4. DESCRIPTION OF USED DATA

In this chapter, data is described by its features. Every feature of the data is explained in such a way where the data is acquired, why this feature is used, which method of normalization is used.

4.1. Data Acquisition

The data which is used in this thesis is acquired from different sources. The target variable is acquired from an e-commerce logistic company with confidential information and will not be shared. There are 44 branch information for the year 2021 in the data set. Other data set features are acquired from the Turkish Statistical Institute (TURKSTAT). Republic of Turkey Ministry of Industry and Technology, Information and Communication Technologies Authority (BTK), Social Security Institution (SGK), Banking Regulation and Supervision Agency (BDDK), and Iskur.

4.2. Features Of The Data

The description of the features is shown in Table 4.2.1. Sources of the features are shown in Table 4.2.2.

Finding information based on the district is the most challenging task of this data set. It is much easy to obtain province-level data than lower-scale data as districts [49]. SEGE (Socio-Economic Development Rank) report published every 10 to 15 years. "agricultural_density", "sege_score", "0-29", "30-59", "60-90+", "ed_higher", "ed_uneducated", "ed_lower", "population", and “delivery_count” features are collected from districts' demographic information. "credit_card", "cash_loans", "deposit_total", "overdraft", "other_loans", "artisan", "workplaces", "registered_unemployed", "registered_labor_force", "compulsory_insured", "net_subs" features are collected from cities’ demographic information. To summarize, only the data acquired from TURKSTAT and e-commerce logistic company is districs’s demographic information.

Other features are based on cities' demographic information. The level collection of the features show in Table 4.2.3.

Table 4.2.1: Features of the Data Set

No	Feature	Description
1	agricultural_density	Amount of arable areas
2	sege_score	Score of socio-economic development ranking
3	0-29	Number of the population aged between 0-29
4	30-59	Number of the population aged between 30-59
5	60-90+	Number of the population aged greater than 60
6	ed_higher	Number of the population whose education is higher than a college degree
7	ed_uneducated	Number of the population which did not have any education in their life.
8	ed_lower	Number of the population whose education is lower than a college degree
9	credit_card	Amount of credit cards usage
10	cash_loans	Amount of cash loans
11	deposit_total	Amount of cash in banks' deposit
12	overdraft	Amount of overdraft bank accounts
13	other_loans	Amount of other consumer loans
14	artisan	Number of artisans who has their workplaces
15	workplaces	Number of workplaces
16	registered_unemployed	Number of registered unemployed people
17	registered_labor_force	Number of registered employed people
18	compulsory_insured	Number of compulsory insured people
19	net_subs	Number of Internet subscribers
20	population	Number of people
21	delivery_count	Number of delivery amount

Table 4.2.2: Sources of the Features

No	Feature	Source
1	agricultural_density	TURKSTAT
2	sege_score	Republic of Turkey Ministry of Industry and Technology General Directorate of Development Agency
3	0-29	TURKSTAT
4	30-59	TURKSTAT
5	60-90+	TURKSTAT
6	ed_higher	TURKSTAT
7	ed_uneducated	TURKSTAT
8	ed_lower	TURKSTAT
9	credit_card	BDDK

10	cash_loans	BDDK
11	deposit_total	BDDK
12	overdraft	BDDK
13	other_loans	BDDK
14	artisan	Republic of Turkey Ministry of Trade
15	workplaces	SGK
16	registered_unemployed	Iskur
17	registered_labor_force	SGK
18	compulsory_insured	SGK
19	net_subs	BTK
20	population	TURKSTAT
21	delivery_count	E-commerce logistic company

Table 4.2.3: Level Collection of the Features

No	Feature	Level
1	agricultural_density	District
2	sege_score	District
3	0-29	District
4	30-59	District
5	60-90+	District
6	ed_higher	District
7	ed_uneducated	District
8	ed_lower	District
9	credit_card	City
10	cash_loans	City
11	deposit_total	City
12	overdraft	City
13	other_loans	City
14	artisan	City
15	workplaces	City
16	registered_unemployed	City
17	registered_labor_force	City
18	compulsory_insured	City
19	net_subs	City
20	population	District
21	delivery_count	District

4.3. Data Preprocessing

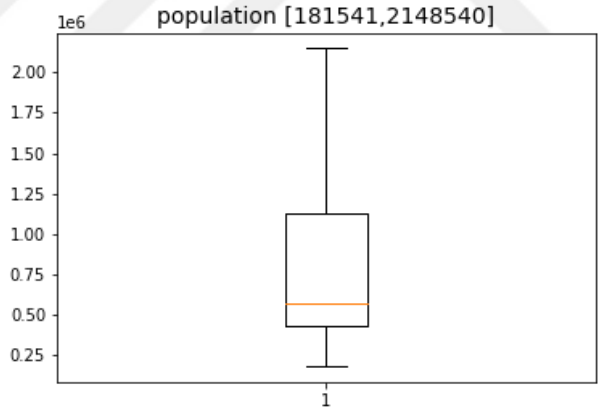
Most of the previous literature to decide on the location of a branch uses three basic data, which are the population of gender groups, education levels, and age groups. In this thesis, the minimum number of data attributes is tried to keep as possible because having too

many features in the data set can cause an overfitting problem. Overfitting, which can be encountered easily when using machine learning techniques, is a statistical error when the model learns the details and noise in the training data. Because of that, similar correlation attributes are combined in a single feature.

4.3.1. Gender groups

Gender is one of the most used data to distinguish when an analysis is made on e-commerce services. In this thesis, the population of gender groups is not considered because the distribution of gender is almost equal to each other in branches and Turkey. Instead of using a population of genders, the total population is included in this data set. The distribution of the total population can be seen in Figure 4.3.1.1. The minimum and maximum available population in branches are between approximately 180 thousand and 2.1 million people. The figure shows that most of the branches services low population.

Figure 4.3.1.1:Distribution of the total population



4.3.2. Education level

As with the literature, education levels positively correlate with the delivery amounts. Education levels were collected more than seven attributes from TURKSTAT. As mentioned before, data sets contain a minimum number of features because of the overfitting problem. From previous research across different places around the globe, most online shopping users' education level is higher education, whose education is more

significant than a college degree. In the dataset, a similar observation is made with this dataset. Thus, seven attributes are reduced into three levels: higher, lower, and uneducated. Higher means that the level of graduation is college or higher. Lower means that the level of graduation is between primary school and high school. Uneducated means that the person never has education in their life. The distribution of education levels can be seen in Figure 4.3.2.1, Figure 4.3.2.2, and Figure 4.3.2.3. The minimum and maximum available population who did not get any education in branches are between approximately 4 thousand and 24 thousand people. The minimum and maximum available population with education between primary school and high school in branches are between approximately 56 thousand and 800 thousand people. The minimum and maximum available population with at least college education in branches are between approximately 80 thousand and 1 million people. Figure 4.3.2.1 shows that uneducated people living inside of branch service are in the minority. There are only two branches that serve uneducated population. Figure 4.3.2.2 indicates that the number of people with at least elemantary degrees is in minorty. Figure 4.3.2.3 shows that well-educated people distributed almost evenly across the branches' available service areas.

Figure 4.3.2.1: Distribution of uneducated population

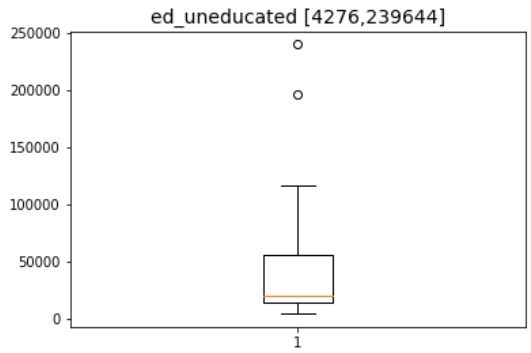


Figure 4.3.2.2: Distribution of population whose education level is lower than a college degree

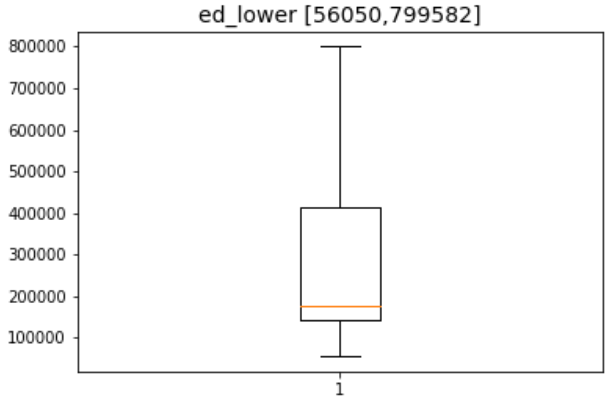
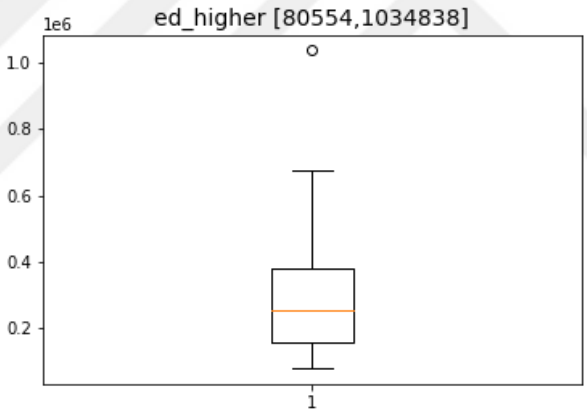


Figure 4.3.2.3: Distribution of population whose education level is higher than a college degree



4.3.3. Age groups

Another data that is used to distinguish into groups is age. Usage of e-commerce services is differentiated with the age groups. There are two main reasons lies behind it. The first one is the technology usage age. Accordingly to a survey, Internet usage age dramatically decreases after age 55 [50]. Another reason is the age of youths' getting economic freedom from their parents. Accordingly, research, the median age of leaving the parental home is increased to 24.4 [51]. This research shows that the age of leaving the parental home increases over the years. This information is important because they are getting more freedom in terms of economically after leaving their parental home. Thus, age groups are

separated into three ages between 0-29, 30-59, 60-90 plus. Most of the online shopping users in this dataset are distributed between 30-59. The distribution of age groups can be seen in Figure 4.3.3.1, Figure 4.3.3.2, and Figure 4.3.3.3. Those figures indicate that people living inside branches' available service are between 0-29 age. The minimum and maximum available population aged between 0-29 in branches are approximately 63 thousand and 1.4 million people. The minimum and maximum available population aged between 30-59 in branches are approximately 85 thousand and 938 thousand people. The minimum and maximum available population aged greater than 60 in branches are between approximately 21 thousand and 311 thousand people.

Figure 4.3.3.1: Distribution of population whose age between 0-29 years old

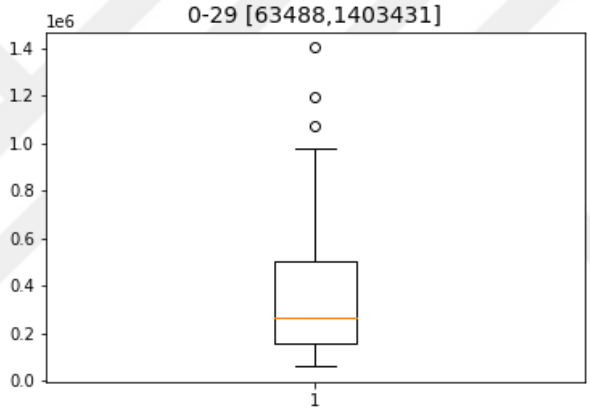


Figure 4.3.3.2: Distribution of population whose age between 30-59 years old

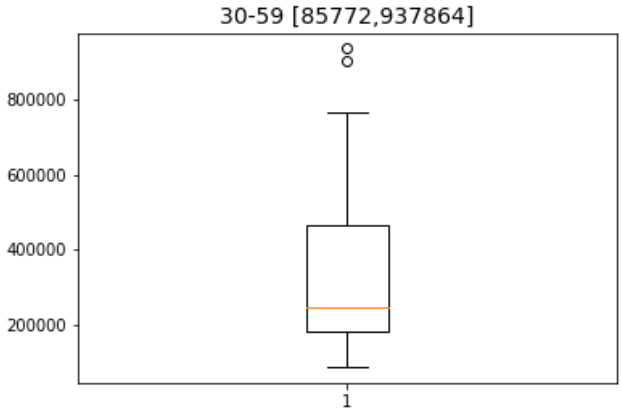
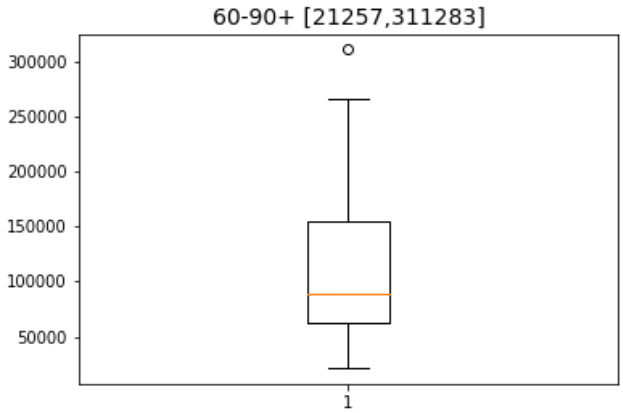


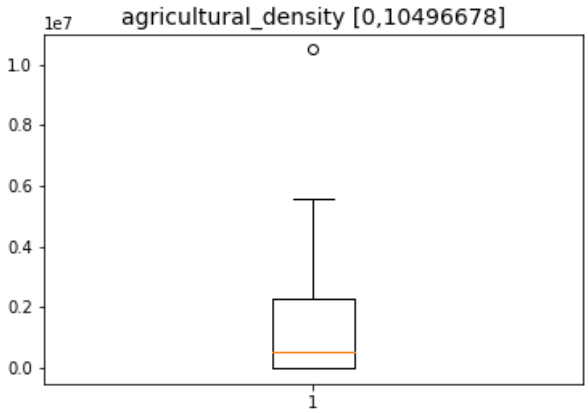
Figure 4.3.3.3: Distribution of population whose age greater than 60



4.3.4. Agricultural area

According to the Federal Highway Administration in the United States of America, most online shopping users are located in urban places. Agricultural density is acquired from TURKSTAT to determine the urban areas. Because Turkey's agricultural area amount is higher than most countries, if the agricultural density is low, the area is urbanized. The distribution of agricultural area density can be seen in Figure 4.3.4.1. The figures indicate most of the branches used in this thesis are available in urban areas. The minimum available area is zero, which means that area has no available agricultural area in the branches' location. The maximum available area is approximately 10.5 million decares.

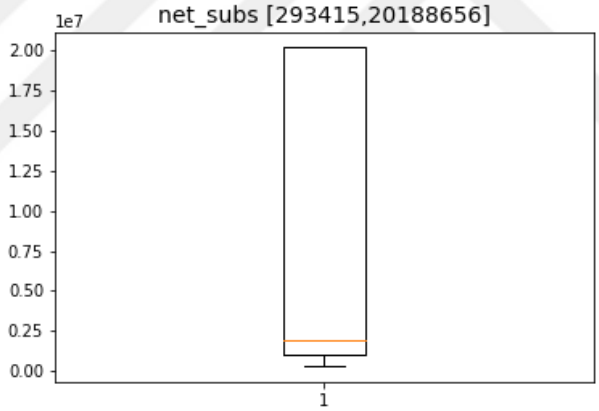
Figure 4.3.4.1: Distribution of agricultural area density



4.3.5. Internet subscription

Online shopping can be used both with mobile and fixed internet users. To be able to use e-commerce services, users must have Internet access. Data which is acquired from BTK had two attributes which are fixed and mobile internet subscribers. In this thesis number of subscribers is combined because online shopping is available from both channels. The distribution of the population who have a subscription to Internet service can be seen in Figure 4.3.5.1. The figures show that too many people do not have any access to the Internet in the branches' available service area. The minimum and maximum available population who have a subscription to Internet service in branches are between approximately 293 thousand and 20 million people.

Figure 4.3.5.1: Distribution of population who have a subscription to Internet service



4.3.6. Banking information

Accordingly, online shopping experts, most purchases are made with credit cards. So, among the credit card usage information, other information is considered helpful. BDDK provides all the banking information publicly. The distribution of banking information can be seen in Figure 4.3.6.1, Figure 4.3.6.2, Figure 4.3.6.3, Figure 4.3.6.4, and Figure 4.3.6.5. All of the figures indicate that banking information data set features to have similar distributions.

Figure 4.3.6.1: Distribution of other loans amount in thousand TL

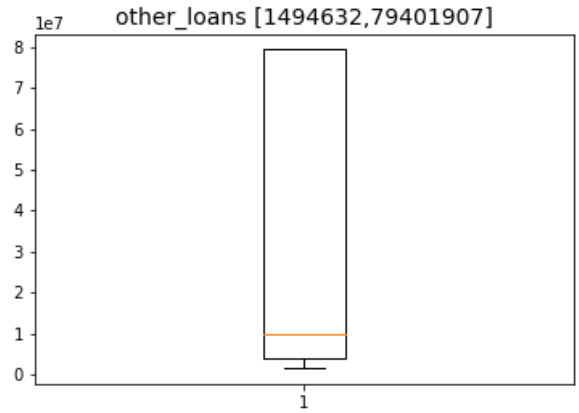


Figure 4.3.6.2: Distribution of total deposit amount in thousand TL

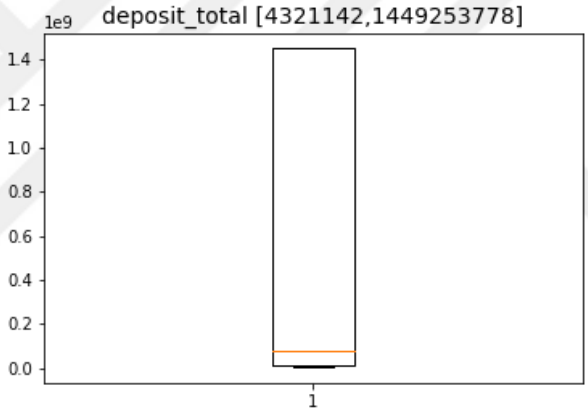


Figure 4.3.6.3: Distribution of credit card usage amount in thousand TL

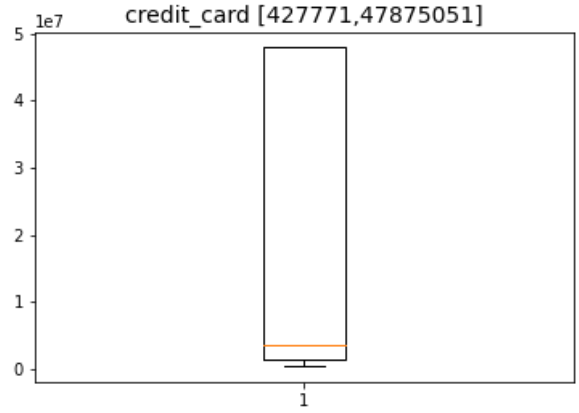


Figure 4.3.6.4: Distribution of cash loans amount in thousand TL

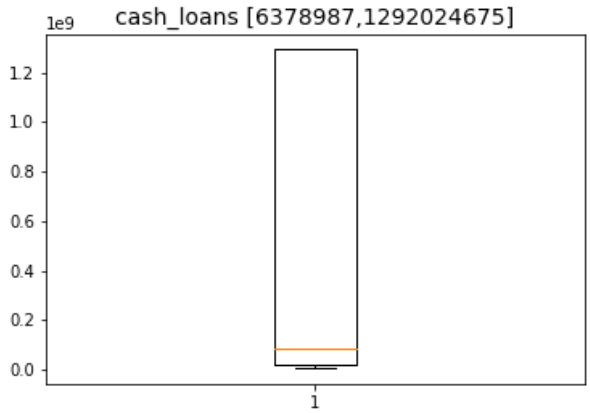
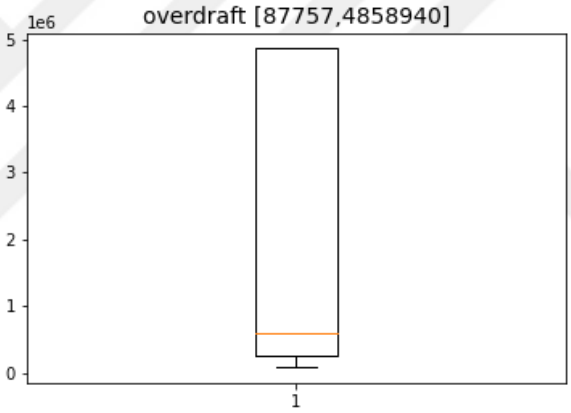


Figure 4.3.6.5: Distribution of overdraft account amount in thousand TL



4.3.7. Financial information

If literature is studied, another essential data is income groups. In Turkey, collecting district-based data is the most challenging thing. The even more difficult problem is that some data is not publicly available such as income information. However, the problem is overcome by using the number of artisans, workplaces, the registered labor force, unemployment, and compulsory insured people. The distribution of the financial information can be seen in Figure 4.3.7.1, Figure 4.3.7.2, Figure 4.3.7.3, Figure 4.3.7.4, and Figure 4.3.7.5. All of the figures indicate that financial information data set features to have similar distributions.

Figure 4.3.7.1:Distribution of the registered unemployed population

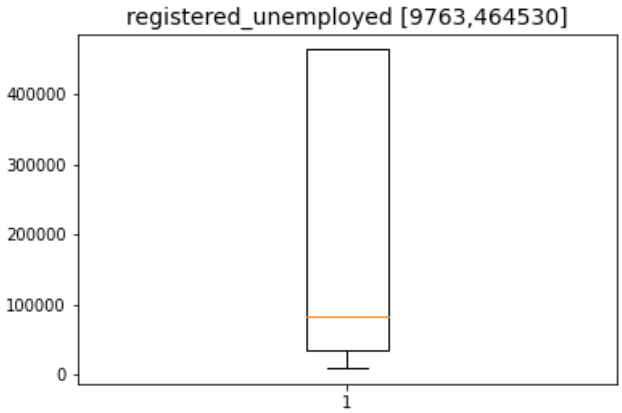


Figure 4.3.7.2: Distribution of registered labor force population

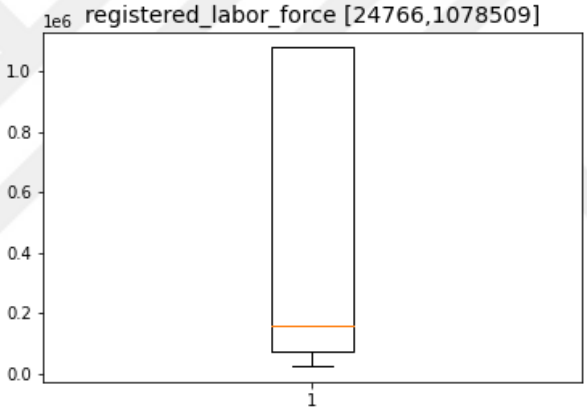


Figure 4.3.7.3: Distribution of compulsory insured population

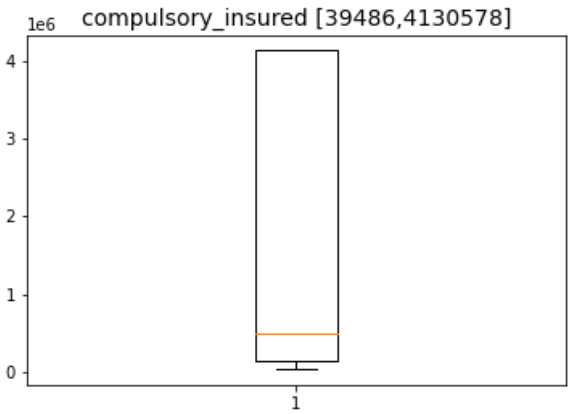


Figure 4.3.7.4: Distribution of the number of artisans

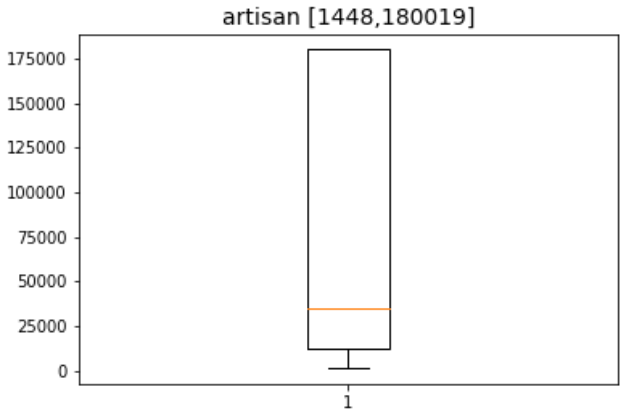


Figure 4.3.7.5: Distribution of the number of workplaces

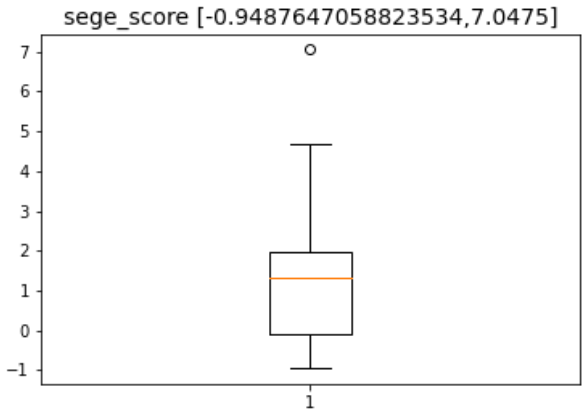
4.3.8. Socio-economic rank

Data publicly available between 10 to 15 years is called SEGE [49]. SEGE score includes 32 attributes that affect the socio-economic rank of the districts. Attributes are divided into seven main sections, which are demographic, employment, education, health, competitiveness, financial, and life quality variables. Demographic variables contain data about population share in Turkey, general fertility rate, migration rate, and average household size. Employment variables contain data about the ratio of the working-age population to district population, share of manufacturing industry employment, actively working women insured rate, and the ratio of population premium paid by the state to district population. Education variables contain data about female literacy ratio, the

number of students per classroom in primary education, pre-school enrollment rate, and the ratio of the population graduated from college or faculty to 22 plus age. Health variables contain data about the number of hospital beds per thousand people, number of physicians per thousand people, number of dentists per thousand people, and number of family medicine applications per person. Competitiveness variables contain data about agricultural production per capita, the share of the number of beds with tourism investment, the share of the investment amount with incentive certificate in Turkey, the share of parcels produced in industrial zones in Turkey, and the share of industrial electricity consumption in Turkey. Financial variables contain data about the number of bank branches per ten thousand people, bank deposit amount per person, bank loan amount per person, municipal revenue per capita, and municipal expenses per person. Life quality variables contain data about residential electricity consumption per person, amount of social aid per person, movie theater availability, and fixed broadband Internet subscribers per person.

The principal component analysis, which is a technique for reducing the dimensionality of a data set, is used to derive a single attribute. Thirty-two attributes reduce to a single attribute as a score determining the district's socio-economical rank. Because the publication data date is recent, the SEGE score is one of the dataset's attributes. The correlation between the SEGE score and the delivery amount is positive, which means it can be meaningful data to be used. The distribution of the SEGE score in branches can be seen in Figure 4.3.8.1. The figures indicate branches' available service areas are mostly well-developed regions. The intervals are between [-1.741,7.73] all over Turkey. The available intervals districts' SEGE score within the branches are approximately [-0.95,7.04]

Figure 4.3.8.1: Distribution of SEGE score



4.4. Normalization

Acquired data attributes have different intervals. Using different intervals can cause inefficient results. Because of that, normalization is one of the critical steps before starting teaching the model. The dataset includes only quantitative data, in other words, numerical data. Data attributes are normalized to use the same scales with min-max normalization. The new scale of the attributes is calculated using min-max normalization with the given intervals. Every attribute of the dataset is transformed into a scale between [0,1] with min-max normalization. In machine learning models, using normalized data is crucial because it affects the model's efficiency and accuracy. Min-max normalization is used because the data set does not contain too many noisy data points.

After applying the min-max normalization on the data set, the distribution between the delivery count and the data set attributes can be seen in Figure 4.4.1, Figure 4.4.2, and Figure 4.4.3. Order of the graphs respect to (a), (b), (c), (d), (e), (f), (g), (h), (i), (j), (k), (l), (m), (n), (o), (p), (q), (r), (s), (t) is normalized distribution of data features total population, uneducated population, lower education population, higher education population, 0-29 age population, 30-59 age population, 60-90 plus population, agricultural density area, number of Internet subscribers, other loans amount, total deposit amount, credit card usage amount, cash loans amount, overdraft account amount, registered unemployed population, registered labor force population, compulsory insured population, number of artisans, number of workplaces and SEGE score.

It can be concluded that some features have similar trends with each other such as credit card usage and cash loans. Feature selection is an essential process to reduce the number of input variables in machine learning techniques. “The advantages of the feature selection are facilitating data visualization and data understanding, reducing the measurement and storage requirements, reducing training and utilization times, defying the curse of dimensionality to improve prediction performance” [52]. Distributions show that the feature selection process must be applied in this dataset because some features have similar trends.

Figure 4.4.1: Data Features Normalized Distributions (a) to (i)

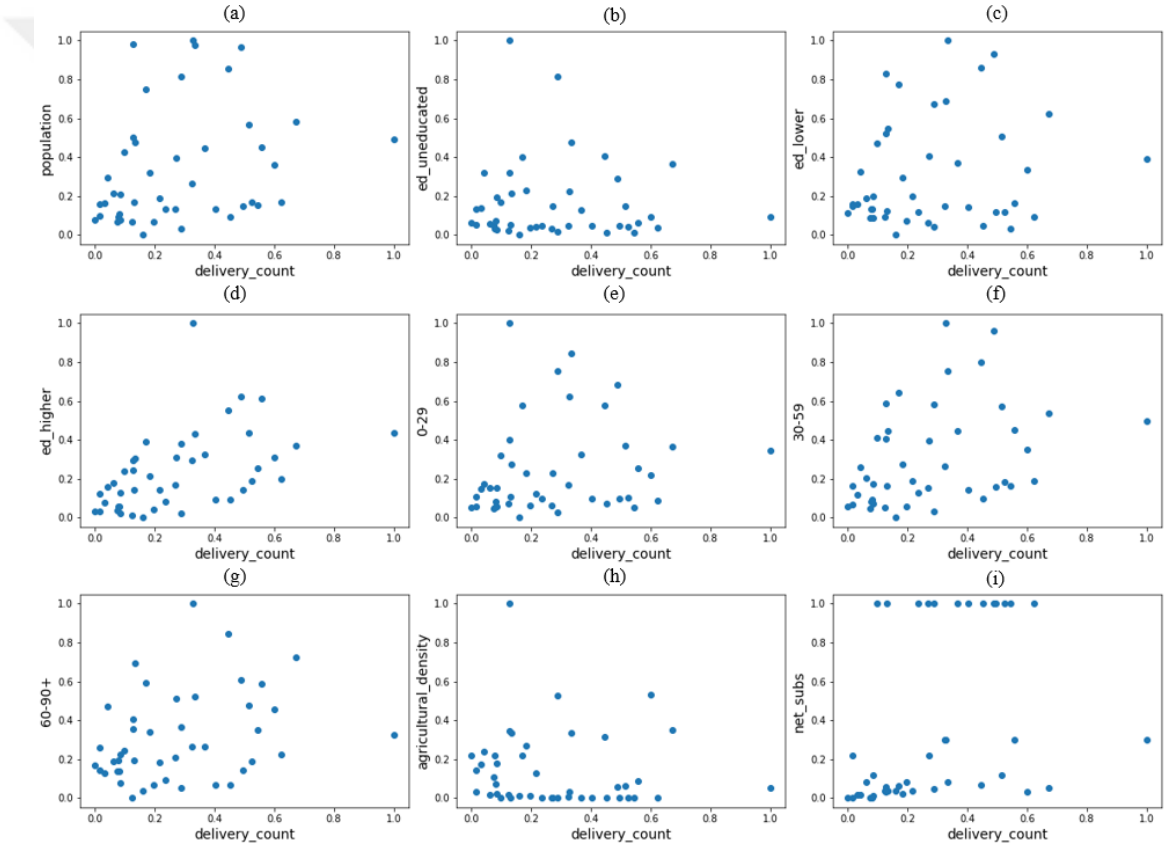


Figure 4.4.2: Data Features Normalized Distributions (j) to (r)

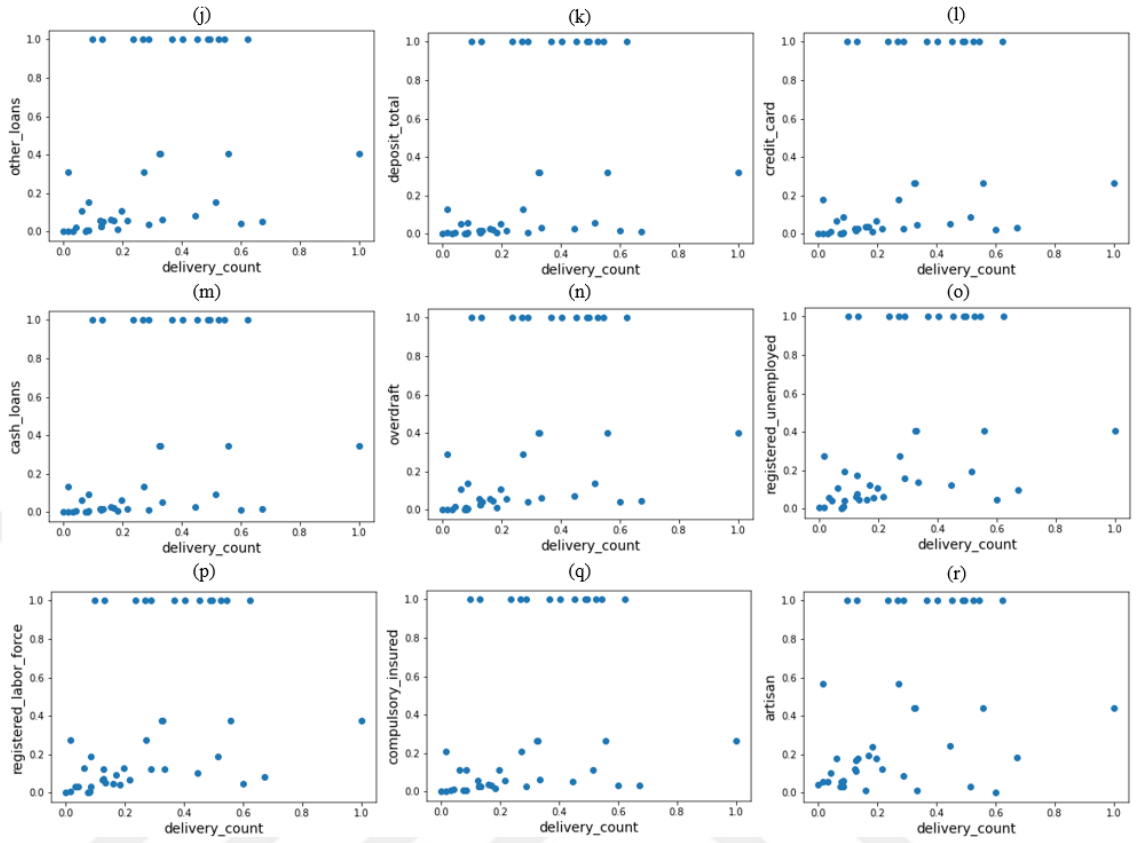
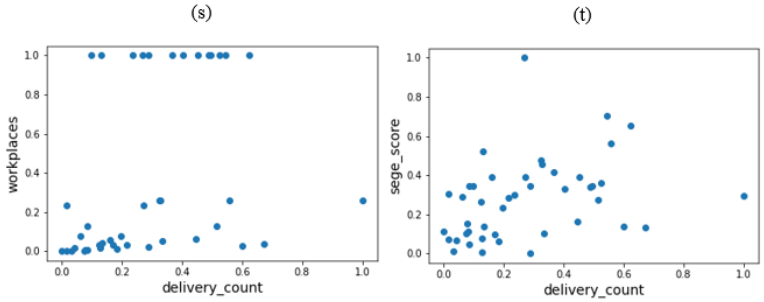


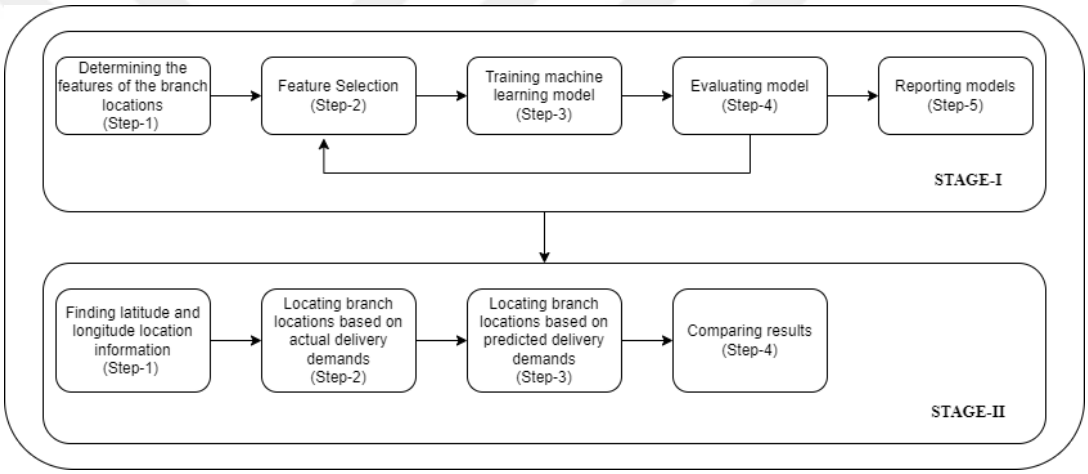
Figure 4.4.3: Data Features Normalized Distributions (s) and (t)



5. METHODS

This chapter explains which methods are used and how to apply those methods. The workflow of this thesis consists of two main stages. The workflow can be seen in Figure 5.1. In the first stage, e-commerce delivery demand potentials are predicted using machine learning techniques with geodemographic's data to select the best model. In the second stage, a new branch location is proposed combining regression analysis and k-means clustering.

Figure 5.1:Workflow

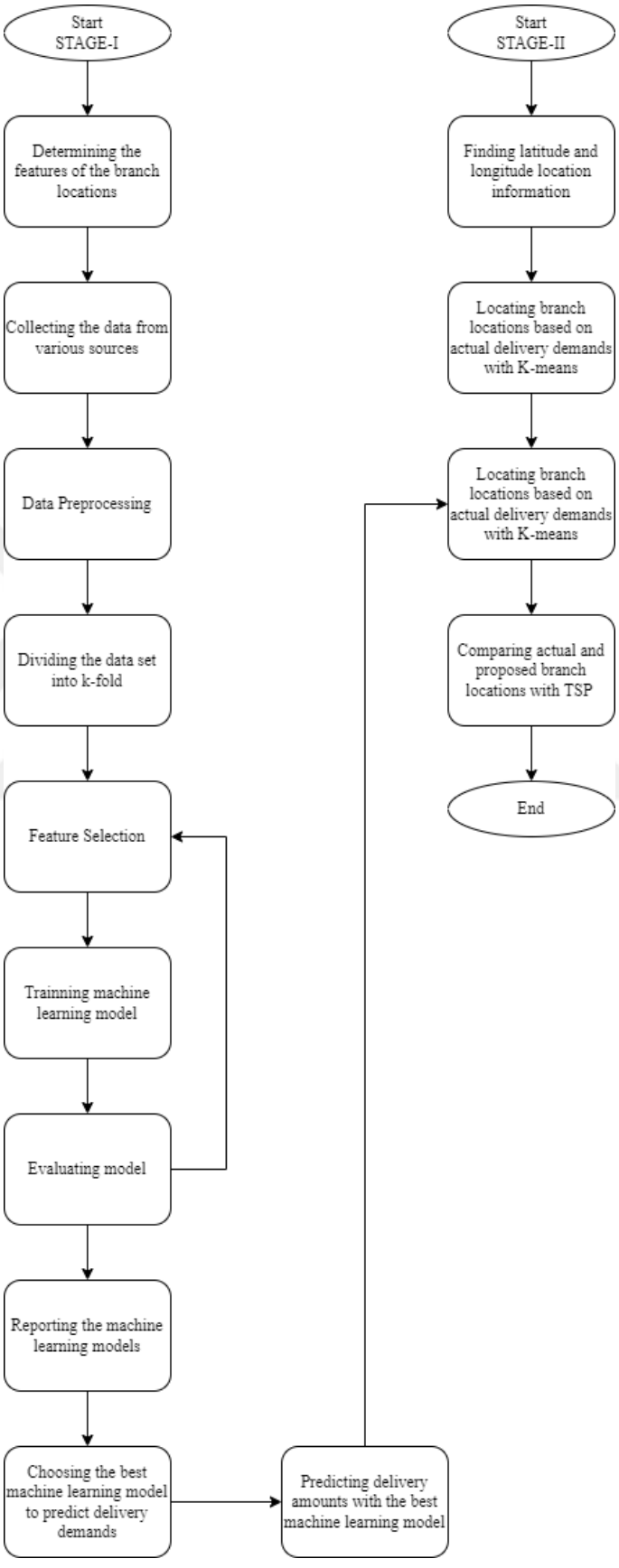


A detailed flowchart of the thesis can be seen in Figure 5.2. Firstly, the features of the data set are determined in Stage-I. Twenty features are selected to be used in the data set after that data set is collected from various sources. Before teaching the models, the data set was preprocessed. Data features are manipulated and normalized in the data preprocessing step. The data set was divided into k-fold to validate the machine learning methods. After that, feature selection which chooses the best features to predict delivery, is applied to increase the reliability of the model. The curse of dimensionality, which causes the problem of is one of the important problems in machine learning. Feature selection methods must be applied to overcome that problem. Several machine learning methods, LR, SVR, and DTR, are used to predict delivery potentials. After teaching an ML model, several evaluation metrics were used to evaluate the models: mean absolute,

mean squared, root mean squared, r-squared, and maximum residual error. Between feature selection and evaluation, steps are repeated for every ML model. In the last step of Stage-I, models are reported, and the best model is chosen to use in Stage-II.

Firstly, location information which is longitude and latitude data is found because the data set only includes the address of the delivery demands. After finding the location of the delivery demands, the K-means clustering algorithm is applied to locate the actual branch location. The best model is selected in Stage-I, and the model predicts the delivery amounts to propose a new branch location with the K-means algorithm. After finding the location of actual and proposed branch locations, locations are compared with the Traveling Salesperson problem (TSP) optimization. TSP optimization solves the problem of giving a set of n nodes and the distances for each pair of nodes, finding a roundtrip of minimum total length visiting each node exactly once [53]. TSP optimization is applied to test the reliability of the proposed location of a new branch.

Figure 5.2: Flowchart



5.1. KNIME

KNIME is a free and open-source data analytics, reporting, and integration platform. It can be used for data preparation, manipulation, cleaning, visualizing, etc. Most of the data preprocessing steps are done in the KNIME software. KNIME supports programming language, but most functions are prebuilt in KNIME. It calls those functions as a node.

The raw dataset includes all of the information about Turkey. The first step of the data preparation was combining the district's data to branch coverage. Most of the district-level data values are the number of the population except agricultural area. All of the values for the individual feature are summed by grouping branch coverage. The second step of the data preparation was joining Turkey's district-based data with the city-based data. As mentioned before in chapter 4.2, acquiring district-level data is one of the challenges in Turkey. Because of that, some of the features are district level and city level. Thus, branches located in the same city have the same value as other branches.

After preparing the data set, min-max normalization was applied in KNIME. All of the features' scales are rearranged between [0,1].

5.2. JetBrains DataSpell IDE

DataSpell is a software application for data science with intelligent Jupyter notebooks, which open-source software for interactive computing across dozens of programming languages. In addition to Jupyter notebooks, it can directly run Python scripts.

5.3. Python Programming Language

Python is an interpreted high-level, general-purpose programming language. Unlike other programming languages, Python uses indentation to increase readability. It also follows the object-oriented approach, which is a method for storing the data. Python has become one of the most used languages for data science with the growing community. Communities create libraries for specific project areas, such as the Scikit-learn library.

5.4. Python Libraries

Python libraries are useful functions that you do not need to write the same algorithm again. There are thousands of libraries which is available to use with Python.

5.4.1. Pandas library

Pandas is a software library written for the Python programming language for data manipulation and analysis. Pandas DataFrame object is a tabular data structure with labeled axes. The data set, which was formatted as comma-separated values, is converted into Pandas DataFrame object to analyze machine learning techniques. The basic usage of Pandas DataFrame object:

```
import pandas as pd
d = {'col1': [1, 2], 'col2': [3, 4]}
df = pd.DataFrame(data=d)
```

Where;

d: Python built-it data type dictionary

df: Pandas DataFrame object

5.4.2. Scikit-learn library

Scikit-learn library, which David Courbapeau developed as a Google Summer of Code project in 2007, is one of Python's most extensive data science tools [54]. Library documentation is accessible through their website. The commonly used machine learning algorithms are supported. Scikit-learn library can be imported with file dependencies sklearn.

Feature selection

Feature selection is the first step before starting to train the models. Feature selection is one of the critical topics for machine learning research areas. They must include meaningful features to train the model. Feature selection methods provide the required features to be used in the model. Models used in this thesis are trained with at least three attributes by selecting the best features. The total number of data set's features is twenty without the targeted label. Thus, models are trained with the best three to twenty features that mean the models are trained without using the feature selection method.

Feature selection is implemented under the feature_selection module. Inside the module, the select best k method with the f regression is used for selecting the best features. Select best k method chooses the best features accordingly to the given selection method, which is f regression. F regression methods return the F-statistics and p-values. F test is one of the standard methods in statistics, which gives the importance of the parameters. F regression method is done in two simple steps. The first step is to compute the cross-correlation between each regressor and the target. The second step is converting into F-score and then a p-value. After that, select best k returns the selected features according to the F-score and the p-value with the given number of k.

K-fold cross-validation

K-fold cross-validation is used to test the reliability of the models. The model is divided into four k because instances of this data set are too small to increase the number of k. Overall, 25% of the data set is used for testing the models in each iteration. K-fold cross-validation is implemented under model_selection module. Each iteration model is trained and tested after splitting into k number of folds. The basic usage of K-fold cross-validation:

```
from sklearn.model_selection import KFold
kf = KFold(n_splits=2)
for train_index, test_index in kf.split(X):
  X_train, X_test = X[train_index], X[test_index]
  y_train, y_test = y[train_index], y[test_index]
```

Where;

X: data set

n_split: number of folds

Linear regression

Linear regression is one of the commonly used regression analysis in machine learning algorithms. In this thesis, linear regression is one of the selected models because the approach of the problem is simple with linear regression. Created model is tested if the data set is suitable with the solution. Linear regression is implemented under linear_model module. The first step is creating an object for the model. After that, the model can be used for making predictions. LR is used to predict branches' e-commerce logistic deliveries. The basic usage of LR:

```
from sklearn.linear_model import LinearRegression
reg = LinearRegression().fit(X, y)
reg.predict(z)
```

Where;

X: training data without targeted label

y: training data for targeted label

z: the data used for predicting by using a trained model.

Support vector regression

Another regression model, which is support vector regression, is applied for the data set. One of the chosen regression models was the support vector regression model because the model tries to minimize the error between predicted and actual values. SVR tries to fit the best line within a boundary value. Support vector regression is implemented under the svm module. The first step is creating an object for the model. After that, the model

can be used for making predictions. SVR is used to predict branches' e-commerce logistic deliveries. The basic usage of SVR:

```
from sklearn.svm import SVR
from sklearn.pipeline import make_pipeline
from sklearn.preprocessing import StandardScaler
regr = make_pipeline(StandardScaler(), SVR(C=1.0, epsilon=0.2)).fit(X, y)
regr.predict(z)
```

Where;

X: training data without targeted label

y: training data for targeted label

z: The data used for predicting by using a trained model.

Decision tree regression

The decision tree algorithm is one of the powerful machine learning techniques. It creates a tree with nodes. Nodes include decisions to predict targeted output by making decisions on the node. Decision tree regression is implemented under the tree module. The first step is creating an object for the model. After that, the model can be used for making predictions. DTR is used to predict branches' e-commerce logistic deliveries. The basic usage of DTR:

```
from sklearn.tree import DecisionTreeRegressor
regr = DecisionTreeRegressor(max_depth=d).fit(X,y)
regr.predict(z)
```

Where;

X: training data without targeted label

y: training data for targeted label

z: the data used for making predictions by using a trained model.

d: maximum depth of the tree

K-nearest neighbor regression

K-nearest neighbor regression (KNNR) is applied to predict delivery amounts. The model looks nearest neighbor's similarities. The basic usage of KNNR:

```
from sklearn.neighbors import KNeighborsRegressor
neigh = KNeighborsRegressor(n_neighbors=2).fit(X, y)
```

Where;

X: training data without targeted label

y: training data for targeted label

n_neigbors: number of nearest neigbors

Ridge regression

Ridge regression (RR), another linear model, is used to predict delivery amounts. The basic usage of the RR:

```
from sklearn.linear_model import Ridge
clf = Ridge().fit(X, y)
```

Where;

X: training data without targeted label

y: training data for targeted label

Kmeans clustering

Kmeans clustering uses the euclidian distance to cluster the data point. Kmeans is implemented under the cluster module. The first step is creating an object for the model. After that, the model can assign the new data points to their clusters and get the centroid's location of clusters. K-means clustering is used after making predictions of e-commerce deliveries. After determining the delivery potential using several regression analyses, the

location of new branches is predicted using K-means clustering. The location of a new branch is found with the predicted delivery amount by using machine learning techniques. The basic usage of Kmeans:

```
from sklearn.cluster import KMeans
kmeans = KMeans(n_clusters=k, random_state=0).fit(d)
centroids = kmeans.cluster_centers_
```

Where;

k: number of clusters

d: training data

centroids: center locations of clusters

Evaluation metrics

In every iteration of the k-fold, the model must be evaluated to decide its reliability. Evaluation metrics are implemented in the metrics module. Several evaluation metrics are calculated to determine the errors. MAE, MSE, RMSE, R2, and maximum residual errors were calculated to observe the model's reliability. The basic usage of the evaluation metrics:

```
from sklearn.metrics import *
mse = mean_squared_error(y_true=ytest, y_pred=y_predicted)
rmse = math.sqrt(mse)
mae = mean_absolute_error(y_true=ytest, y_pred=y_predicted)
r2 = r2_score(y_true=ytest, y_pred=y_predicted)
max_ras_error = max_error(y_true=ytest, y_pred=y_predicted)
```

Where;

ytest: testing target labels

y_predicted : predicted target labels

5.4.3. Geopy library

The location-allocation problem is solved with a simple K-means cluster algorithm. Geographic Information System (GIS) location is used on the clustering model. The acquired data has no information about exact latitude and longitude. Finding the latitude and longitude information is another challenge to apply to k-means clustering. Spatial data is obtained by using the geopy library. Geopy library can be used to find latitude and longitude from an address or an address from latitude and longitude. The basic usage of geopy library:

```
from geopy.geocoders import Nominatim
geolocator = Nominatim(user_agent="geolocator")
location = geolocator.geocode(s)
```

Where;

s: address to find latitude and longitude

location: found latitude and longitude

Nominatim: Free address web service

5.4.4. Local-TSP library

Traveling Salesperson Problem is heuristically solved with Local-TSP library. “In the Vehicle Routing Problem (VRP), the goal is to find optimal routes for multiple vehicles visiting a set of locations” [55]. If there's only one vehicle, it reduces to the TSP. Local-TSP performs the Lin-Kernighan algorithm, one of the heuristic solutions to find the shortest route to cover all the nodes in TSP. The distance matrix, which includes every distance with the nodes, is determined with the GeoPy library. GeoPy library calculates the distance between two locations in terms of the geodesic. Geodesic is a curve that represents the shortest path between two points in a surface.

5.4.5. Gurobipy library

Gurobi is a mathematical optimization solver. Gurobipy is the Python API of Gurobi solver. Gurobi solver is used for the exact solution of the TSP by utilizing the cutting planes algorithm to solve TSP optimally. Implementation of the Gurobi solver can be found in their documentation [56].

6. DATA ANALYSIS

In this chapter, models are evaluated with each other. Evaluation methods are explained in chapter 3.5. The best model is chosen by evaluation metrics. Propose a new location for branch, K-Means algorithm is applied to compare with the actual delivery amount and predicted delivery amounts.

Number abbreviation of features can be seen in Table 6.1. Features are abbreviated into numbers to show in other tables, as seen in Table 6.1.1.1, Table 6.1.2.1, Table 6.1.3.1, Table 6.1.4.1, and Table 6.1.5.1.

Table 6.1: Feature Number List

Feature	Feature Number
agricultural_density	1
sege_score	2
0-29	3
30-59	4
60-90+	5
ed_higher	6
ed_uneducated	7
ed_lower	8
credit_card	9
cash_loans	10
deposit_total	11
overdraft	12
other_loans	13
artisan	14
workplaces	15
registered_unemployed	16
registered_labor_force	17
compulsory_insured	18

net_subs	19
population	20

6.1. Evaluating Machine Learning Techniques To Predict Delivery Amounts

Several machine learning techniques, LR, DTR, KNNR, RR, and SVR, are applied to be chosen to allocate a new branch location. Models are created by using the k-fold cross-validation method. The feature selection method is applied in each training with select best k features by f regression method. Several evaluation metrics are observed to compare the reliability of the machine learning techniques. Mean absolute error, mean squared error, root mean squared error, r2, and maximum residual error metrics are evaluated to select the best model for finding a new branch location.

6.1.1. Linear regression model

Evaluation metrics of the LR model can be seen in Table 6.1.1.1. The best LR model is created using ed_higher, overdraft, and registered_unemployed features. The best model of the LR is illustrated in bold letters in the table. The reason of R2 evaluation metric gives zero is that the LR model can not follow the data with the fitted line. Also, the overall performance of this model is very weak when we compare the results with other models' evaluation metrics. In addition to others, when a given number of features increases, the LR model's performance decreases. The LR model can not predict delivery amounts with this data set.

Table 6.1.1.1: LR Model Evaluation Metrics

Features	**MAE**	**MSE**	**RMSE**	**R2**	**Max Error**
[6, 12, 16]	**0.14158**	**0.03745**	**0.18419**	**-0.20025**	**0.38049**
[6, 12, 13, 16]	0.16490	0.04870	0.20578	-0.55000	0.43565
[6, 12, 13, 16, 17]	0.17235	0.05135	0.21241	-0.59777	0.43880
[6, 10, 12, 13, 16, 17]	0.19555	0.07056	0.24795	-1.06072	0.56545
[6, 10, 11, 12, 13, 16, 17]	0.19810	0.07871	0.26251	-1.18397	0.54099

[6, 10, 11, 12, 13, 16, 17, 19]	0.28809	0.20991	0.39647	-4.15734	0.89119
[6, 10, 11, 12, 13, 14, 16, 17, 19]	0.39223	0.47392	0.53612	-9.83613	1.14426
[4, 6, 10, 11, 12, 13, 14, 16, 17, 19]	0.34301	0.35554	0.48216	-7.46078	1.08819
[4, 6, 9, 10, 11, 12, 13, 14, 16, 17, 19]	0.75622	3.18833	1.14279	-66.23177	2.27974
[4, 6, 9, 10, 11, 12, 13, 14, 15, 16, 17, 19]	0.77493	3.14374	1.17068	-66.59891	2.32554
[4, 6, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19]	0.86315	4.65521	1.32347	-94.31900	2.55130
[2, 4, 6, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19]	0.93740	4.80933	1.44708	-98.54764	2.82498
[2, 4, 5, 6, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19]	0.97709	4.95511	1.50673	-103.13072	2.96321
[2, 4, 5, 6, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]	1.13368	5.90453	1.75511	-123.91262	3.59159
[2, 3, 4, 5, 6, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]	1.13368	5.90453	1.75511	-123.91262	3.59159
[2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]	1.20420	6.85829	1.85103	-142.94363	3.75968
[1, 2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]	1.12875	6.17371	1.72415	-128.03405	3.44427
[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]	1.29171	6.81909	1.96309	-142.36566	4.14777

6.1.2. Decision tree regression model

Evaluation metrics of DTR can be seen in Table 6.1.2.1. The best model of DTR was obtained with ed_higher, overdraft, and registered_unemployed attributes. The best model of the DTR is illustrated in bold letters in the table. One of the essential issues to be careful about in the DTR model is the maximum depth of the tree. If the given maximum depth input is too high, it can cause the problem of overfitting. In every creation of the tree, the maximum depth of the tree is given the number of features in that model to overcome the overfitting problem. DTR model has the same problem as the LR model, which is R2 evaluation metrics give negative results. Thus, this model should not be

applied to predict delivery demands. So, the overall performance of the DTR is not good enough to be used for finding a new branch location when we compared it with other models' evaluation metrics. The maximum residual error is high, so the model cannot predict one of the location's delivery amounts.

Table 6.1.2.1: DTR Model Evaluation Metrics

Features	MAE	MSE	RMSE	R2	Max Error
[6, 12, 16]	**0.12592**	**0.03184**	**0.17662**	**-0.4946**	**0.41805**
[6, 12, 13, 16]	0.14879	0.04468	0.20891	-1.5454	0.48842
[6, 12, 13, 16, 17]	0.13382	0.03873	0.19466	-1.2022	0.47035
[6, 10, 12, 13, 16, 17]	0.14132	0.04138	0.20247	-1.4855	0.48317
[6, 10, 11, 12, 13, 16, 17]	0.14407	0.04441	0.20904	-1.9424	0.49901
[6, 10, 11, 12, 13, 16, 17, 19]	0.14426	0.04318	0.20575	-1.4412	0.47524
[6, 10, 11, 12, 13, 14, 16, 17, 19]	0.13327	0.03762	0.18947	-0.8641	0.48251
[4, 6, 10, 11, 12, 13, 14, 16, 17, 19]	0.17001	0.05326	0.22833	-2.4829	0.51171
[4, 6, 9, 10, 11, 12, 13, 14, 16, 17, 19]	0.15926	0.05111	0.22101	-2.3122	0.51884
[4, 6, 9, 10, 11, 12, 13, 14, 15, 16, 17, 19]	0.16836	0.05456	0.22839	-2.5172	0.5144
[4, 6, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19]	0.16626	0.05502	0.22732	-2.3706	0.50672
[2, 4, 6, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19]	0.16432	0.04854	0.21467	-1.607	0.48507
[2, 4, 5, 6, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19]	0.15896	0.0449	0.20663	-1.1794	0.46477
[2, 4, 5, 6, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]	0.1798	0.07004	0.25435	-4.0613	0.61312
[2, 3, 4, 5, 6, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]	0.17199	0.05448	0.22762	-2.4511	0.51526
[2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]	0.17971	0.05552	0.23062	-2.5714	0.48586
[1, 2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]	0.16572	0.05074	0.21673	-1.6144	0.50576
[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]	0.1551	0.04607	0.20715	-1.0886	0.49708

6.1.3. K-nearest neighbor regression model

Evaluation metrics of KNNR can be seen in Table 6.1.3.1. The best model of the KNNR is created with ed_higher, overdraft, and registered_unemployed features. The best model of the KNNR is illustrated in bold letters in the table. The most critical input of KNNR is the number of neighbors. The number of neighbors affects the results. If the number of neighbors is too low, it can cause the problem of underfitting. If it is too high, it can cause the problem of overfitting. The number of neighbors is given as five in all of the KNNR models because the number of instances is too small to provide a larger number of neighbors in this data set. The overall performance of KNNR was the second-best model when we compared it with other models' evaluation metrics. However, the R2 error is not acceptable for finding a new location for a branch because the model can not follow the trend of the data set.

Table 6.1.3.1: KNNR Model Evaluation Metrics

Features	MAE	MSE	RMSE	R2	Max Error
[6, 12, 16]	**0.13452**	**0.03562**	**0.17956**	**-0.0552**	**0.41421**
[6, 12, 13, 16]	0.14337	0.04067	0.19343	-0.2577	0.44857
[6, 12, 13, 16, 17]	0.13955	0.04043	0.19079	-0.1615	0.45399
[6, 10, 12, 13, 16, 17]	0.13775	0.03971	0.18899	-0.1423	0.45399
[6, 10, 11, 12, 13, 16, 17]	0.13495	0.03794	0.18398	-0.0917	0.42561
[6, 10, 11, 12, 13, 16, 17, 19]	0.1384	0.04006	0.1895	-0.1434	0.45399
[6, 10, 11, 12, 13, 14, 16, 17, 19]	0.13955	0.04239	0.19012	-0.0726	0.44206
[4, 6, 10, 11, 12, 13, 14, 16, 17, 19]	0.14298	0.04072	0.1919	-0.2014	0.44509
[4, 6, 9, 10, 11, 12, 13, 14, 16, 17, 19]	0.14281	0.0407	0.19186	-0.201	0.44509
[4, 6, 9, 10, 11, 12, 13, 14, 15, 16, 17, 19]	0.14577	0.04346	0.19719	-0.2454	0.46222
[4, 6, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19]	0.14532	0.04342	0.19709	-0.2445	0.46222
[2, 4, 6, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19]	0.13847	0.04099	0.19196	-0.1913	0.44714
[2, 4, 5, 6, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19]	0.13922	0.043	0.19243	-0.154	0.45856

[2, 4, 5, 6, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]	0.14475	0.04084	0.18991	-0.1334	0.43139
[2, 3, 4, 5, 6, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]	0.14314	0.04032	0.18898	-0.125	0.43202
[2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]	0.14364	0.04055	0.19079	-0.2026	0.42699
[1, 2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]	0.13916	0.03927	0.18498	-0.0604	0.41031
[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]	0.14086	0.04342	0.19114	-0.1534	0.43722

6.1.4. Ridge regression model

Evaluation metrics of the RR model can be seen in Table 6.1.4.1. The best model of the RR is created with ed_higher, overdraft, and registered_unemployed features. The best model of the RR is illustrated in bold letters in the table. RR's overall performance is weaker than KNNR and SVR models. In addition to that case, it is not good enough to be used because the R2 error has negative values, which indicates the RR model can not follow the trend of the data set.

Table 6.1.4.1:RR Model Evalutaion Metrics

Features	MAE	MSE	RMSE	R2	Max Error
[6, 12, 16]	**0.13578**	**0.03507**	**0.17899**	**-0.1244**	**0.36134**
[6, 12, 13, 16]	0.13623	0.03518	0.1793	-0.1272	0.36131
[6, 12, 13, 16, 17]	0.13666	0.03536	0.17984	-0.1313	0.35907
[6, 10, 12, 13, 16, 17]	0.1495	0.04131	0.19499	-0.3344	0.39721
[6, 10, 11, 12, 13, 16, 17]	0.15093	0.04289	0.19828	-0.3546	0.40799
[6, 10, 11, 12, 13, 16, 17, 19]	0.15175	0.04359	0.20011	-0.3578	0.41413
[6, 10, 11, 12, 13, 14, 16, 17, 19]	0.14793	0.04245	0.19752	-0.3416	0.41867
[4, 6, 10, 11, 12, 13, 14, 16, 17, 19]	0.16276	0.05577	0.22075	-0.6506	0.50043
[4, 6, 9, 10, 11, 12, 13, 14, 16, 17, 19]	0.16298	0.05589	0.22142	-0.6481	0.49788

[4, 6, 9, 10, 11, 12, 13, 14, 15, 16, 17, 19]	0.16227	0.05526	0.22026	-0.6297	0.49163
[4, 6, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19]	0.16281	0.05511	0.22008	-0.6224	0.4877
[2, 4, 6, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19]	0.16555	0.05698	0.22453	-0.6864	0.49624
[2, 4, 5, 6, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19]	0.15983	0.05446	0.22024	-0.556	0.47996
[2, 4, 5, 6, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]	0.16869	0.05748	0.22681	-0.6597	0.49434
[2, 3, 4, 5, 6, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]	0.17073	0.058	0.22798	-0.6758	0.49568
[2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]	0.18015	0.06599	0.23978	-0.8328	0.52724
[1, 2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]	0.16304	0.0537	0.21706	-0.52	0.48842
[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]	0.17394	0.05806	0.22993	-0.9223	0.52037

6.1.5. Support vector regression model

Evaluation metrics of the SVR can be seen in Table 6.1.5.1. The best model is obtained with ed_higher, cash_loans, overdraft, other_loans, registered_unemployed, and registered_labor_force features. There are two crucial inputs, C and epsilon, in the SVR model to make the prediction reliable. “Epsilon specifies the epsilon-tube within which no penalty is associated in the training loss function with points predicted within a distance epsilon from the actual value” [57]. All of the models are given as 0.01 to decrease the errors of the model. C is the regularization parameter 0.9 in all of the models to reduce errors within all the models. The best model of the SVR is illustrated in bold letters in the table. SVR is the best model among other used machine learning techniques. The maximum residual error is still one of the problems, but it can be tolerable to find a new branch location. The error of R2 shows that the data set and the SVR can be used for determining a new location of the branch because it can follow the trend of the data set.

Table 6.1.5.1: SVR Model Evaluation Metrics

Features	MAE	MSE	RMSE	R2	Max Error
[6, 12, 16]	0.11379	0.02787	0.15733	0.25087	0.36596
[6, 12, 13, 16]	0.11421	0.02794	0.15733	0.25458	0.36128
[6, 12, 13, 16, 17]	0.11290	0.02748	0.15610	0.26514	0.35724
[6, 10, 12, 13, 16, 17]	**0.11212**	**0.02703**	**0.15503**	**0.27216**	**0.35247**
[6, 10, 11, 12, 13, 16, 17]	0.11271	0.02708	0.15527	0.26795	0.35040
[6, 10, 11, 12, 13, 16, 17, 19]	0.11482	0.02744	0.15643	0.25426	0.35077
[6, 10, 11, 12, 13, 14, 16, 17, 19]	0.11514	0.02722	0.15624	0.24767	0.35172
[4, 6, 10, 11, 12, 13, 14, 16, 17, 19]	0.11847	0.02928	0.16187	0.19748	0.35743
[4, 6, 9, 10, 11, 12, 13, 14, 16, 17, 19]	0.11881	0.02890	0.16094	0.20436	0.35367
[4, 6, 9, 10, 11, 12, 13, 14, 15, 16, 17, 19]	0.11934	0.02866	0.16028	0.21052	0.35191
[4, 6, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19]	0.11997	0.02870	0.16041	0.20911	0.35199
[2, 4, 6, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19]	0.12226	0.02872	0.15920	0.22409	0.34098
[2, 4, 5, 6, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19]	0.12562	0.03255	0.16863	0.15095	0.37901
[2, 4, 5, 6, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]	0.12630	0.03317	0.17142	0.10249	0.38348
[2, 3, 4, 5, 6, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]	0.13094	0.03412	0.17390	0.06687	0.38296
[2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]	0.13602	0.03580	0.17779	0.02178	0.38441
[1, 2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]	0.14294	0.04041	0.19033	-0.15420	0.42265
[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]	0.14753	0.04162	0.19549	-0.30151	0.43370

6.2. Locating A New Branch

The best model is support vector regression for predicting e-commerce logistic service potential. SVR is used for locating a new branch. Branches of Izmir are removed from

the training data set to compare the results with the actual center location, which is determined with the K-means algorithm's centroid location. A new model is trained with the best obtained SVR model, including ed_higher, cash_loans, overdraft, other_loans, registered_unemployed, and registered_labor_force features. The predicted branch location can be seen in Figure 6.2.1 and Figure 6.2.2. Blue circles are the generalized district centers of deliveries. The black circle is the actual center location from the K-means algorithm with the actual deliveries. The red circle is the determined new location for a branch from K-means with the predicted delivery amounts. The distance between the actual center and the predicted center is approximately 810 and 5400 meters.

The scope of this thesis is locating a new branch location based on the center of the delivery demands. Finding a storage place is another challenge, and it is not considered in this thesis. The predicted branch location may not have an available place to store deliveries. It finds the center of the delivery demands, which means finding a close place to store deliveries for the determined location will increase the potential of the delivery amounts.

Figure 6.2.1:Actual and Predicted Branch Locations I

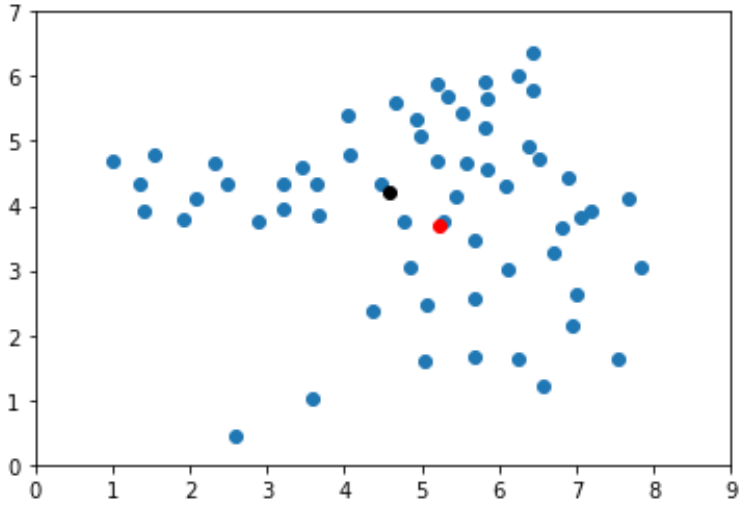
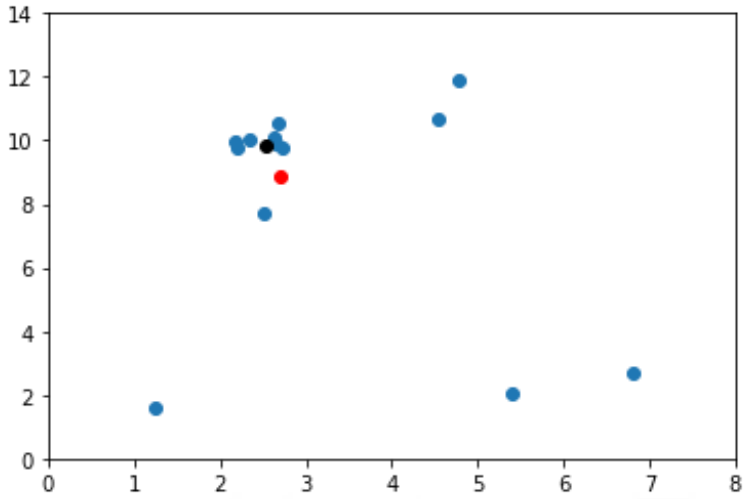


Figure 6.2.2: Actual and Predicted Branch Locations II



6.3. Proposed Model Results

After determining new locations for branches, Traveling Salesperson Problem optimization is applied to compare the locations with round trip distance (RTD) to determine the reliability of the proposed model. Results compared with two methods of the TSP, heuristic and exact solution.

6.3.1. Heuristic results

The Lin-and-Kernighan algorithm is used as a heuristic method. For the branch in Figure 6.2.1, the actual location's annual average RTD is 31171.88 meters, and the predicted location's annual average RTD is 31073.79 meters. Thus, the amount of distance traveled has been reduced with the proposed model. For the branch in Figure 6.2.2, the actual location's annual average RTD is 57585.86 meters, and the predicted location's RTD is 57727.49 meters. Thus, the amount of distance traveled has been increased with the proposed model. When we combine these results, the total cost will be increased by 43.54 meters daily, which is acceptable because it is too small to be considered.

In Figure 6.3.1.1 and Figure 6.3.1.2, the round trip distances between actual and predicted branches can be seen daily basis in meters. For both figures, branches notated actual as A and predicted as P. The positive difference show better performance of the predicted

branch. The negative difference shows better performance of the actual branch. 55.1% of the year performs better in the predicted branch I. The average of the predicted branch's RTD has outperformed the actual branch location, 98.1 meters daily for branch I. 67.2% of the year serves better in the predicted branch II. However, the average of the RTD is performed worse than the actual branch, 141.6 meters daily for branch II. The reason why branch II performed worse even most of the days outperformed the actual location is that some days the difference of RTD is more than 7500 meters.

Figure 6.3.1.1: The difference between actual and predicted RTD branch I with heuristic solution

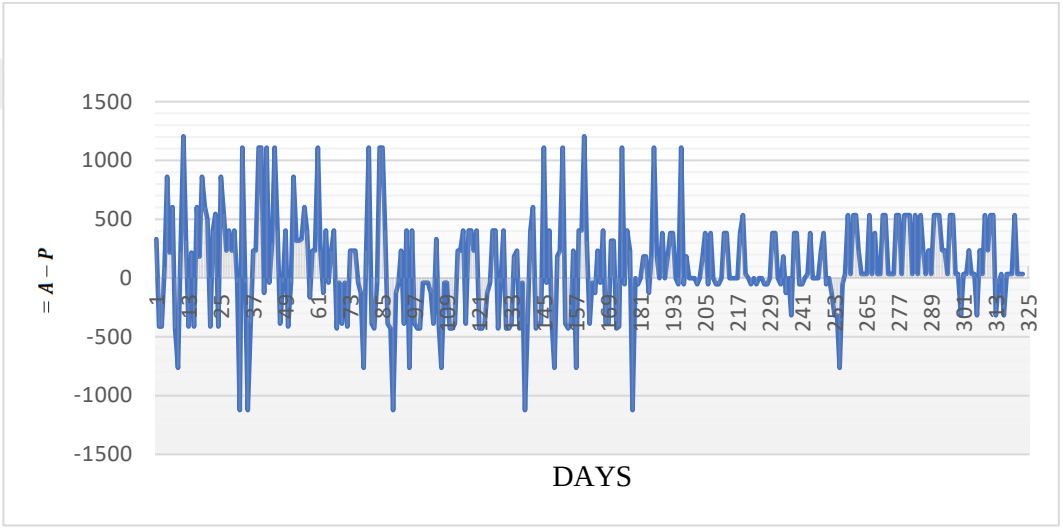
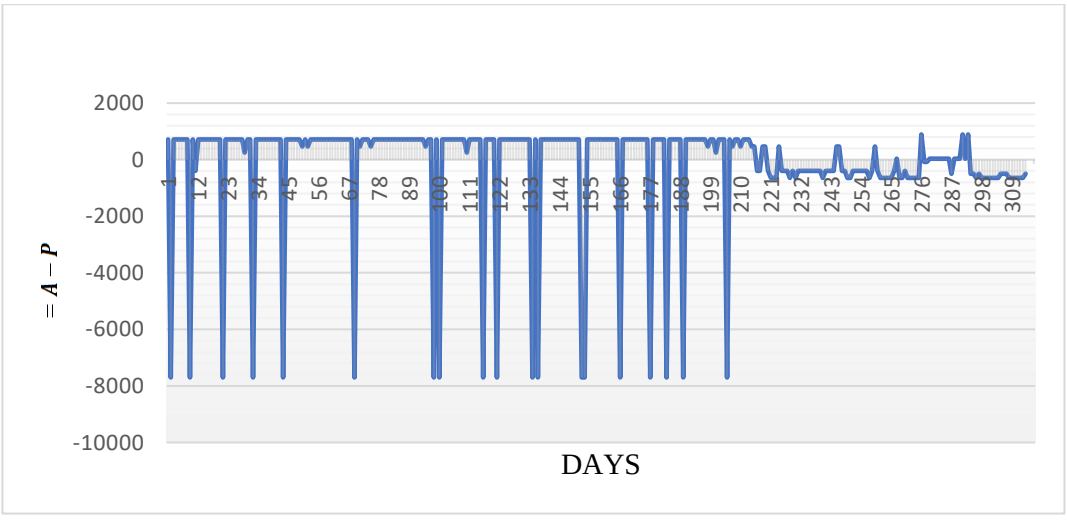


Figure 6.3.1.2: The difference between actual and predicted RTD branch II with heuristic solution



The TSP optimization delivery route can be seen for a day in Figure 6.3.1.3 and Figure 6.3.1.4. Branch locations are annotated as 0 (zero) for both figures. All of the deliveries start from branch location zero end with branch location zero. For predicted branch I, the delivery route is 330.7 meters shorter than the actual branch I. For predicted branch II, the delivery route is 717 meters shorter than the actual branch I. TSP optimization is applied for all the working days to calculate the average RTD for comparison to test the reliabilty of the proposed model.

Figure 6.3.1.3: TSP for branch I with heuristic solution

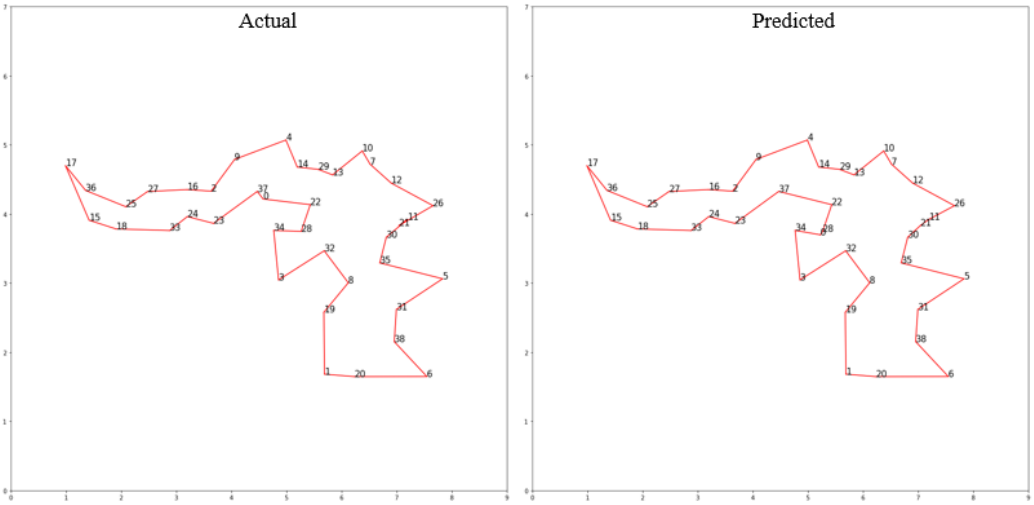
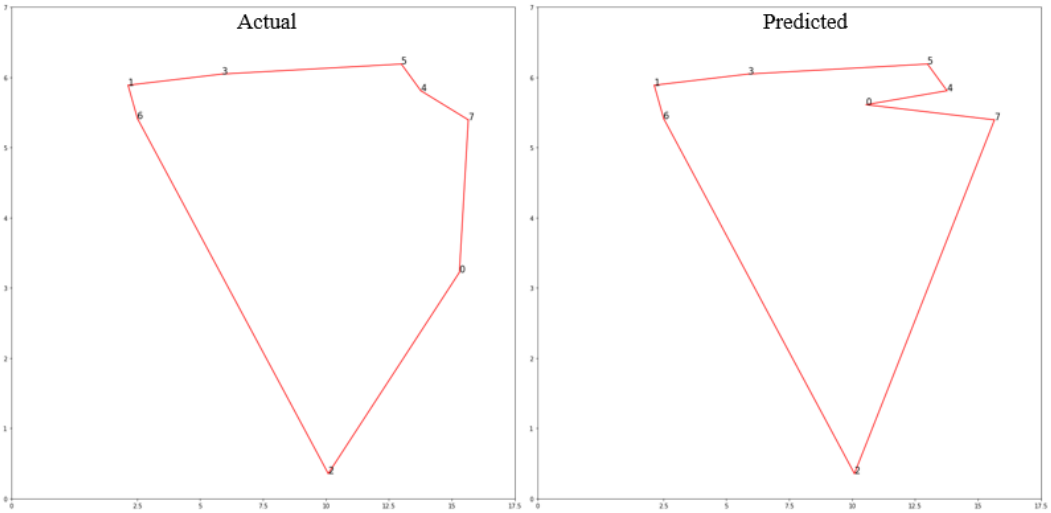


Figure 6.3.1.4: TSP for branch II with heuristic solution



The second use case is dividing one branch into two branches. In Figure 6.3.1.5 and Figure 6.3.1.6, branch division results can be seen for both actual and predicted branches using K-means. For this case, branch I from Figure 6.2.1 is used. When the branch is divided into two, service areas occurred to be different for both cases. Because of this situation, total RTD is compared to observe the reliability of the proposed model. For both graphs, purple and yellow data points are the center of the delivery districts. Also, red and blue data points are for the branch locations. For the branch in Figure 6.3.1.5, the blue branch location's annual average RTD is 8605.19 meters, and the red location's annual average RTD is 16975.84 meters. Thus, the total amount of distance traveled to distribute all the deliveries is 25581.03 meters on a daily basis. For the branch in Figure 6.3.1.6, the blue location's annual average RTD is 6022.25 meters, and the red location's RTD is 18291.78 meters. Thus, the total amount of distance traveled to distribute all the deliveries is 24314.03 meters on a daily basis. Consequently, the total amount of distance traveled to distribute deliveries is decreased by 1267 meters for each working day with the proposed model.

In Figure 6.3.1.7, the round trip distances between actual and predicted branches can be seen daily basis in meters. Branches notated actual as A and predicted as P. The positive difference show better performance of the predicted branches. The negatice difference show better performance of the actual branches. 59.8% of the year performs better in the predicted branches.

Figure 6.3.1.5: Dividing a branch into two actual

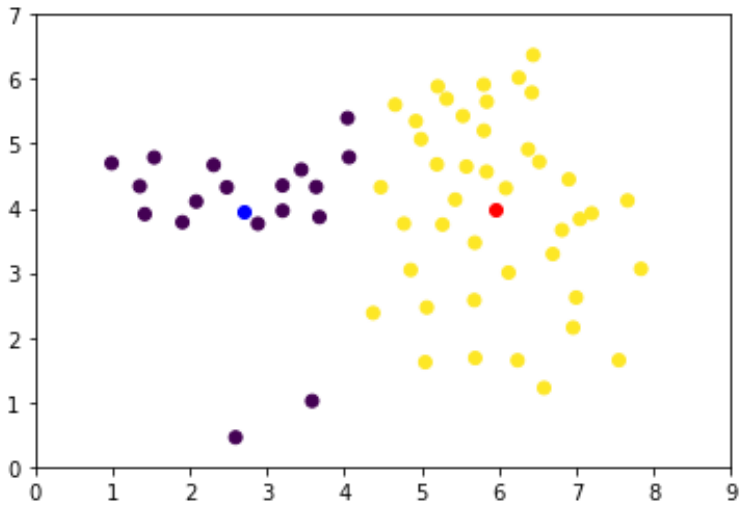


Figure 6.3.1.6: Dividing a branch into two predicted

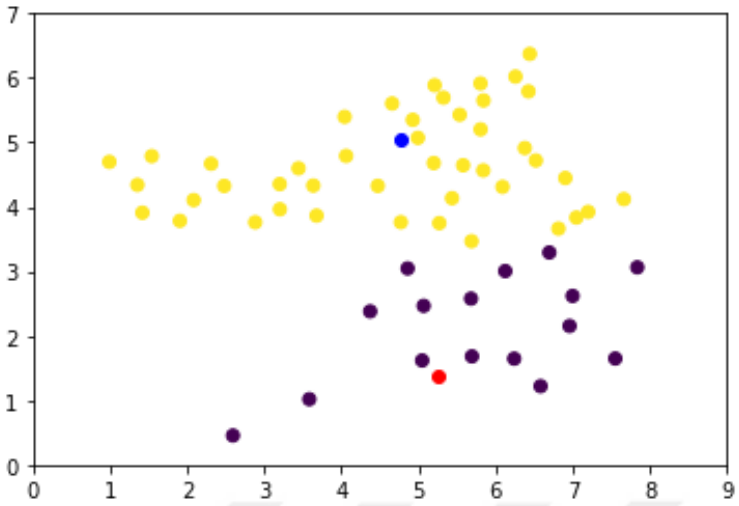
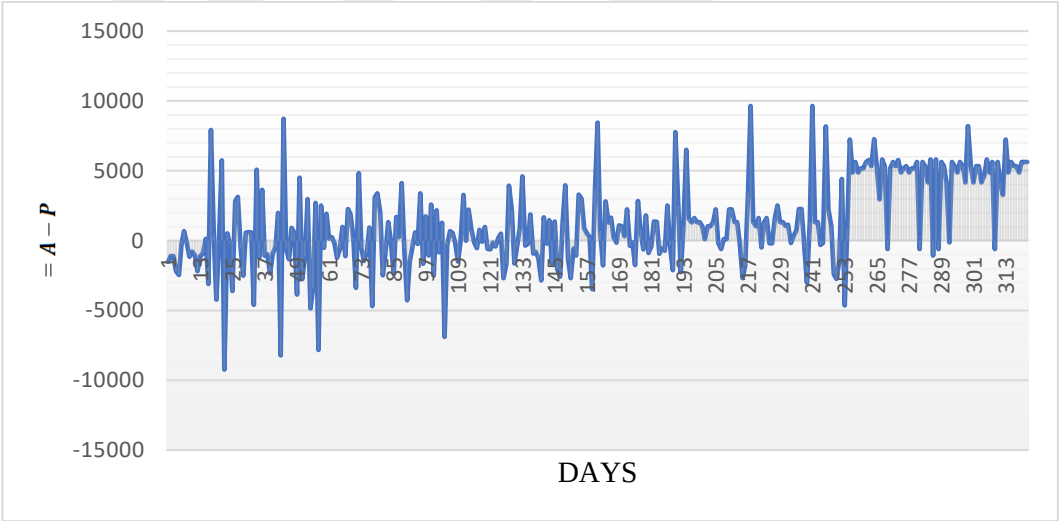


Figure 6.3.1.7: The difference between actual and predicted branches RTD with heuristic solution



6.3.2. Exact solution results

Gurobi solver's cutting planes method is used to obtain exact solutions of TSP for comparing the actual and proposed model results. The For the branch in Figure 6.2.1, the actual location's annual average RTD is 30833.36 meters, and the predicted location's annual average RTD is 30770.35 meters. Thus, the amount of distance traveled has been reduced with the proposed model. For the branch in Figure 6.2.2, the actual location's annual average RTD is 54653.5 meters, and the predicted location's RTD is 54796.9

meters. Thus, the amount of distance traveled has been increased with the proposed model. When we combine these results, the total cost will be increased by 80.39 meters daily, which is still acceptable because it is too small to be considered.

In Figure 6.3.2.1 and Figure 6.3.2.2, the round trip distances between actual and predicted branches can be seen daily basis in meters. The properties of the figures are identical to section 6.3.1. 30.2% of the year performs better in the predicted branch I. However, the average of the predicted branch's RTD has outperformed the actual branch location, 63 meters daily for branch I. 64.74% of the year serves better in the predicted branch II. However, the average of the RTD is performed worse than the actual branch, 143.4 meters daily for branch II. The reason why branch II performed worse even most of the days outperformed the actual location is that some days the difference of RTD is more than 7500 meters.

Figure 6.3.2.1: The difference between actual and predicted RTD branch I with exact solution

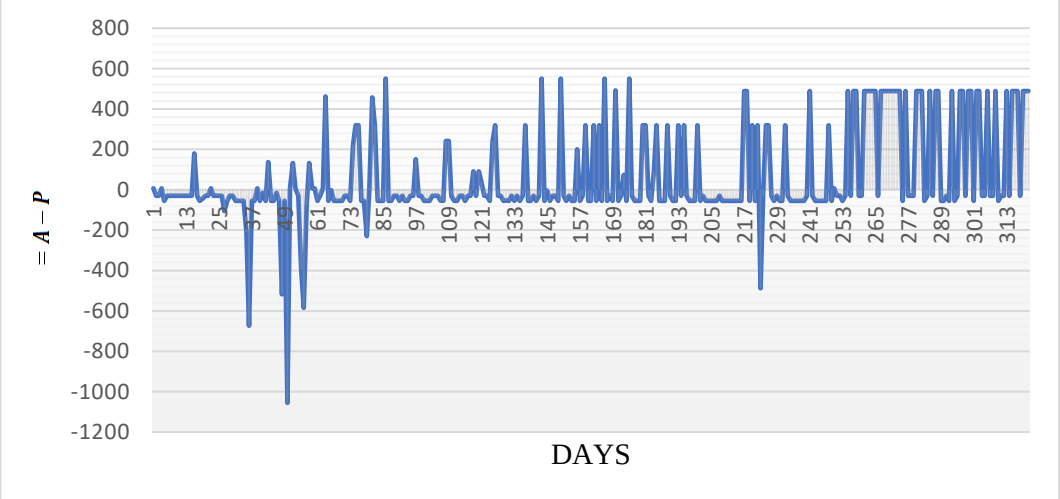
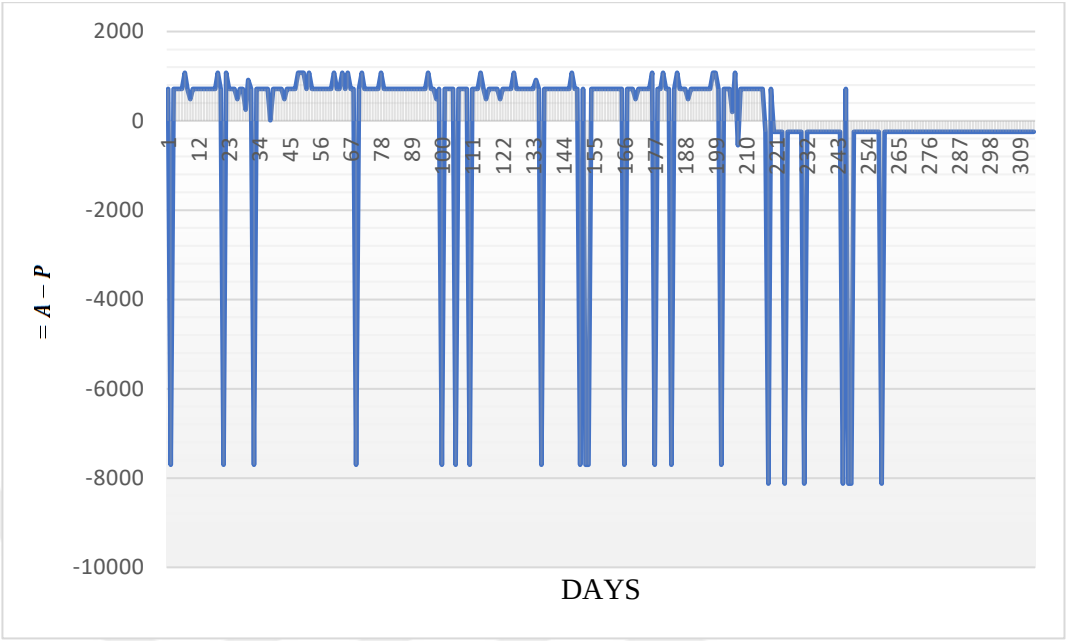


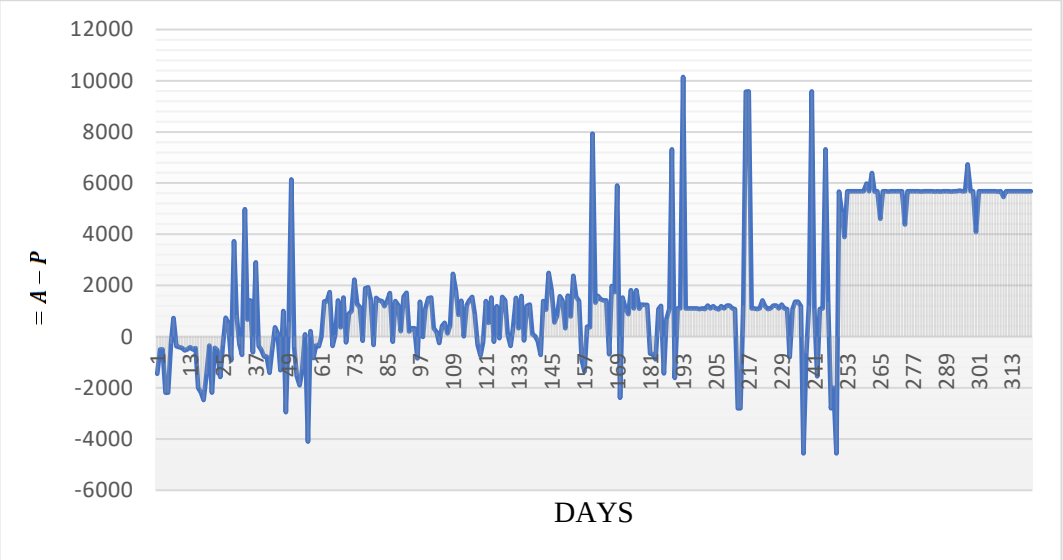
Figure 6.3.2.2: The difference between actual and predicted RTD branch II with exact solution



In Figure 6.3.1.5 and Figure 6.3.1.6, branch division results can be seen for both actual and predicted branches using K-means. For the branch in Figure 6.3.1.5, the blue branch location's annual average RTD is 9205.92 meters, and the red location's annual average RTD is 16801.51 meters. Thus, the total amount of distance traveled to distribute all the deliveries is 26007.43 meters on a daily basis. For the branch in Figure 6.3.1.6, the blue location's annual average RTD is 5986.99 meters, and the red location's RTD is 18225.37 meters. Thus, the total amount of distance traveled to distribute all the deliveries is 24212.36 meters on a daily basis. Consequently, the total amount of distance traveled to distribute deliveries is decreased by 1795.07 meters for each working day with the proposed model.

In Figure 6.3.2.3, the round trip distances between actual and predicted branches can be seen daily basis in meters. Branches notated actual as A and predicted as P. The positive difference shows better performance of the predicted branches. The negative difference shows better performance of the actual branches. 75% of the year performs better in the predicted branches.

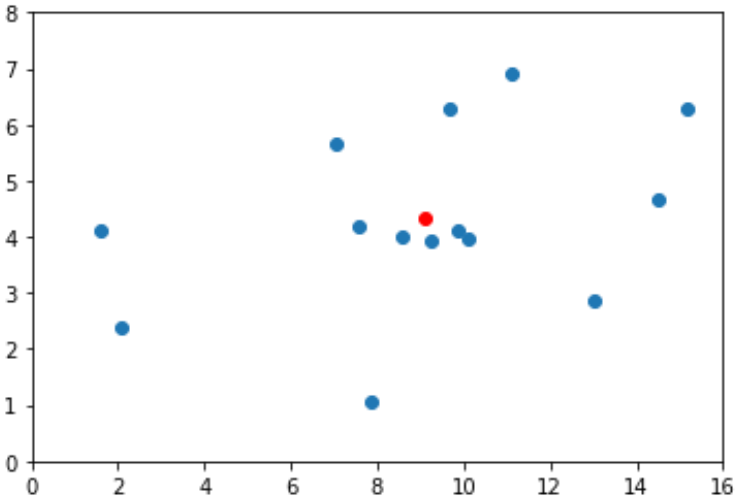
Figure 6.3.2.3: The difference between actual and predicted branches RTD with exact solution



6.3.3. Use case for undiscovered city

Another use case is determining a location for a branch to be opened in the city first time. Bursa is selected to determine a location for a branch because the dataset does not include the data of the city. In Figure 6.3.3.1, the red data point is the determined branch location, and the blue data points are the center of the districts for the city of Bursa. The determined location is not shared because of the confidentiality of the data set.

Figure 6.3.3.1: The first branch for a city



7. CONCLUSION

Logistics is one of the essential variables for e-commerce services. It directly affects both customer loyalty and the revenue of the company. The biggest problem of the logistics is transporting the goods to customers. Before distributing the goods to a customer, a robust distribution network should be established between customers and the warehouses. In this network, branches have essential roles which will affect the delivery time between customers and warehouses in the distribution.

A network can be established by determining the best branch location with the help of machine learning techniques. This problem is another complex location-allocation problem that can be solved with the geodemographic information which affects consumer behavior on e-commerce services. There are some properties that affect the consumer behavior on e-commerce services like income, age, and education, etc.

The problem is solved by dividing the problem into two subproblems. The first problem is predicting the e-commerce on-demand potential. Several machine learning techniques, LR, DTR, KNNR, RR, and SVR, are applied to accomplish that problem. The best model is obtained using the SVR algorithm with ed_higher, cash_loans, overdraft, other_loans, registered_unemployed, and registered_labor_force features. The model is not too reliable because of the insufficiency of data points. However, it promises some potential to predict the number of e-commerce on-demand deliveries.

The second problem is solving the location-allocation problem. Determining a new branch is the based vital decision to make. After finding the best model for the prediction of the on-demand deliveries, the location of the new branch is determined by applying the K-means algorithm, which calculates the centroids of the clusters. Centroids which are the center locations of the clusters, can be used as a branch location. Distances between customers and branches are a much more complex problem, but centroids found with the euclidean distance can be the best location for a new branch. The density of the

on-demand delivery locations will affect the centroid locations, and centroids will be the closest locations for most of the deliveries.

After determining the actual and predicted locations with the K-means algorithm, Traveling Salesperson Problem optimization is used to compare the locations with heuristic and exact solutions. TSP optimization is applied for each working day to observe the average RTD in the branches. Two use case is examined to test the reliability of the proposed model. The first case is opening a new branch in a district, and the second case is dividing a branch into two. For both methods, the proposed model slightly increases the RTD. However, the proposed model can still be used. The reason is that difference is too small to be considered because the proposed model made a small increase as much as it was located in different buildings of the same street. For the second case, the proposed model outperformed the actual locations of the branches in both methods. TSP optimization indicates that the proposed model can be applied in the location-allocation problem because the total round trip distance decreased or slightly increased.

One of the challenges in this work was acquiring the demographic data based on district District-based available data is limited in the Republic of Turkey. Most of the data is collected by the provincial directorates. To make this more reliable provincial data can be obtained as district-based. Obtaining district-based data can be challenging, but data can be obtained by surveying the demographic information. If a demographic data set can be acquired, it will help this proposed solution work more reliably. The data set can also be used for any business area to make decisions without human intervention.

In this thesis, only a limited amount of branch information is used. If the number of the branches is increased in the data set, the proposed solution reliability can increase respectively. The data set includes only the recent year information. If the data set increases with the past years' data set, it can make more reliable decisions to find a location for a new branch.

REFERENCES

[1] G. D. Taylor, *Introduction to Logistics Engineering.* Boca Raton, FL, USA: CRC Press, 2008.

[2] M. Kreak, "The best map ever?", *International Journal of Cartography*, vol. 7, pp. 205-210, 2021.

[3] S. Cheevalier. "Retail e-commerce sales worldwide from 2014 to 2024" Statista. https://www.statista.com/statistics/379043/worldwide-retail-e-commerce-sales/ (accessed Jan. 1, 2022).

[4] Y. Lin, J. Luo, S. Cai, S. Ma, and K. Rong, "Exploring the service quality in the e-commerce context a triadic view", *Industrial Management & Data Systems*, vol. 116, no. 3, pp. 388-415.

[5] C. Doyle, *A Dictionary of Marketing*, 4th ed. Oxford, UK: Oxford Univ. Press, 2016.

[6] M. Figliozzi and A. Unnikrishnan, "Exploring the impact of socio-demographic characteristics, health concerns, and product type on home delivery rates and expenditures during a strict COVID-19 lockdown period: A case study from Portland, OR", *Transportation Research Part A: Policy and Practice*, vol. 153, pp. 1-19, 2021.

[7] "Changes in Online Shopping Trends", U.S. Department of FHWA, 2018.

[8] J. Hou and K. Elliott, "Mobile shopping intensity: Consumer demographics and motivations", *Journal of Retailing and Consumer Services*, vol. 63, 2021.

[9] J. Chacón-García, "Geomarketing techniques to locate retail companies in regulated markets", *Australasian Marketing Journal*, vol. 25, pp. 185-193, 2017.

[10] S. S. Noorian, "A Decision Support System for Sales Territory Planning Using the Genetic Algorithm", M.S. thesis, Dept. Civil, Geo and Env.Eng., Munich Tech. Univ., Munich, Germany, 2015.

[11] Y. Yanga, J. Tangb, H. Luoc, and R. Lawd, "Hotel location evaluation: A combination of machine learning tools and web GIS", *International Journal of Hospitality Management*, vol. 47, pp. 14-24, 2015.

[12] M. A. Salazar-Aguilar, J. L. González-Velarde, and R. Z. Ríos-Mercado, "A Divide-and-Conquer Approach to Commercial Territory Design", *Computación y Sistemas*, vol. 16, no. 3, pp. 309-320, 2012.

[13] S. Noorian, A. Psyllidis, and A. Bozzon, "A time-varying p-median model for location-allocation", *In 21st Conf. on Geo-Information Science*, Delft, NL, 2018.

[14] M. Bastı, “The P-median Facility Location Problem and Solution Approaches”, *Online Academic Journal of Information Technology*, vol. 3, no. 3, 2012.

[15] E. Olivares-Benitez, M. B. Bernábe-Loranca, S. Caballero-Morales, and R. Granillo-Macias, “Multi-objective Design of Balanced Sales Territories with Taboo Search: A Practical case”, *International Journal of Supply and Operations Management*, vol. 8, no. 2, pp. 176-193, 2021.

[16] H. Hsieh, F. Lin, C. Li, E. Ian, and H. Chen, “Temporal popularity prediction of locations for geographical placement of retails stores”, *Knowledge and Information Systems*, vol. 60, pp. 247-273, 2019.

[17] S. Moreno, J. Pereira, and W. Yushimito, “A hybrid K-means and integer programming method for commercial territory design: a case study in meat distribution”, *Annals of Operation Research*, vol. 286, pp. 87-117, 2020.

[18] B. Çavdar and J. Sokol, “A distribution-free TSP tour length estimation model for random graphs”, *European Journal of Operational Research*, vol. 243, no. 2, pp. 558-598, 2015.

[19] C. Gorse, D. Johnston, and M. Pritchard, *A Dictionary of Construction, Surveying and Civil Engineering*, 2nd ed. Oxford, UK: Oxford Univ. Press, 2020.

[20] “Data Never Sleeps 9.0” Domo. https://domo.com/learn /infographic/data-never-sleeps-9 (accessed Nov. 10, 2021)

[21] D. Bashir, G. D. Montanez, S. Sehra, P. S. Segura, and J. Lauw, “An Information-Theoretic Perspective on Overfitting and Underfitting”, *The 33rd Australasian Joint Conf. on Artificial Intelligence*, Canberra, AU, 2020.

[22] J. Daintith and E. Wright, *A Dictionary of Computing*, 6th ed. Oxford, UK: Oxford Univ. Press, 2008.

[23] N. Hashimzade, G. Myles, and J. Black, *A Dictionary of Economics*, 5th ed. Oxford, UK: Oxford Univ. Press, 2017.

[24] M. Tranmer, J. Murphy, M. Elliot, and M. Pampaka, *Multiple Linear Regression*, Manchester, UK: Cathie March Institute, 2020.

[25] M. Awad and R. Khanna, “Support Vector Regression”, in *Efficient Learning Machines*, Berkeley, CA, USA: Apress, 2015, pp. 67-80.

[26] N. Cristianini and E. Ricci, “Support Vector Machines”, in *Encyclopedia of Algorithms*, M. Y. Kao, Ed., Boston, MA, USA: Springer, 2008, pp. 928-932.

[27] M. Xu, P. Watanachaturaporn, P. K. Varchney, and M. K. Arora, “Decision tree regression for soft classification of remote sensing data”, *Remote Sensing of Environment*, vol. 97, no. 3, pp. 322-336, 2005.

[28] “Decision Tree Regression” Scikit-learn. https://scikit-learn.org/stable/auto_examples/tree/plot_tree_regression.html (accessed Jan. 1, 2022).

[29] S. B. Imandoust and M. Bolabdraftar, “Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events Theoretical Background”, *Journal of Engineering Research and Applications*, vol. 3, no. 5, pp. 605-610, 2013.

[30] I. Muhajir, “K-Neighbors Regression Analysis in Python” Analytics Vidhya. https://medium.com/analytics-vidhya_k-neighbors-regression-analysis-in-python-61532d56d8e4 (accessed Jan. 1, 2022).

[31] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, “Least Angle Regression”, *The Annals of Statistics*, vol. 32, no. 2, pp. 407-399, 2004.

[32] A. E. Hoerl and R. W. Kennard, “Ridge Regression: Biased Estimation for Nonorthogonal Problems”, *Technometrics, vol. 42, no. 1, pp. 80-86, 2000.*

[33] G. Tutz and H. Binder, “Boosting Ridge Regression”, *Computational Statistics & Data Analysis*, vol. 51, pp. 6044-6059, 2007.

[34] P. Bühlmann and B. Yu, “Boosting with the L2 loss: regression and classification”, *Journal of the American Statistical Association*, vol. 98, no. 462, pp. 324-339, 2003.

[35] Z. Ghahramani, “Unsupervised Learning”, in *Advanced Lectures on Machine Learning*, O. Bousquet, U. von Luxburg, G. Rätsch, Eds., Berlin, Heidelberg, Germany: Springer, 2004, pp. 72-112.

[36] L. Rokach and O. Maimon, “Clustering Methods”, in *Data Mining and Knowledge Discovery Handbook*, O. Maimon, L. Rokach, Eds., Boston, MA, USA: Springer, 2005, pp. 321-352.

[37] C. Piech, “K Means” Stanford. https://stanford.edu/~cpiech/cs221/handouts/kmeans.html (accessed Jan. 1, 2022).

[38] M. Ester, H. Kriegeli J. Sander, and X. Xu, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”, *Proc. of the 2nd Int. Conf. on Knowledge Discovery and Data Mining*, Menlo Park, NJ, USA, 1996.

[39] N. S. Chauhan, “DBSCAN Clustering Algorithm in Machine Learning” KD Nuggets. https://www.kdnuggets.com/2020/04/dbscan-clustering-algorithm-machine-learning.html (accessed Jan. 1, 2022).

[40] M. Ankerst, M. M. Breunig, H. Kriegel, and J. Sander, “OPTICS: Ordering Points to Identify the Clustering Structure”, *Proc. Of the 1999 ACM SIGMOD Int. Conf. on Management of Data*, Philadelphia, PA, USA, 1999.

[41] DR. K. D. Sree, DR. C. S. Bindu, “Data Analytics: Why Data Normalization”, *International Journal of Engineering & Technology*, vol. 7, no. 4.6, pp. 209-213, 2018.

[42] S. G. K. Patro and K. Kishore, “Normalization: A preprocessing Stage”, *International Advanced Research Journal in Science, Engineering and Technology*, vo. 2, no. 3, pp. 2394-1588, 2015.

[43] T. Hastie, R. Tibshirani, and J. Friedman, “Model Assessment and Selection”, in *The Elements of Statistical Learning*, New York, NY, USA: Springer, 2009, pp. 219-259.

[44] J. Dantas, “The importance of k-fold cross-validation for model prediction in machine learning” Towards Data Science. https://towardsdatascience.com/the-imporatance-of-k-fold-cross-validation-for-model-prediction-in-machine-learning-4709d3fed2ef (accessed Jan. 1, 2022).

[45] D. Chicco, M. J. Warrens, and G. Jurman, “The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation”, *PeerJ Computer Science*, vol. 7, 2021.

[46] “Vehicle Routing” Google. https://developers.google.com/optimization/routing (accessed Jan. 27, 2022)

[47] E. V. Dijck, “A Branch-and-Cut Algorithm for the Traveling Salesman Problem with Drone”, M.S. thesis, Dept. Operations Research & Quantitative Logistics, Erasmus Univ. Rotterdam, Rotterdam, Netherlands, 2018.

[48] K. Helsgaun, “General k-opt submoves for the Lin-Kernighan TSP heuristic”, *Mathematical Programming Computation*, vol 1., no. 2, pp. 119-163, 2009.

[49] F. Yılmaz, S. Acar, Dr. L. Bilen Kazancık, L. Gültekin, M. C. Meydan, Dr. M. Emin Özsan, M. Işık, “İlçelerin Sosyo-Ekonomik Gelişmişlik Sıralaması Araştırması SEGE-2017, Republic of Turkey Ministry of Industry and Technology General Directorate of Development Agency, Ankara, Turkey, 2019.

[50] “Internet usage: users by age group Turkey 2012-2019” Statista. https://www.statista.com/statistcis/998042/internet-users-by-age-group-turkey/ (accessed Dec. 4, 2021).

[51] İ. Koç, “The Timing of Leaving the Parental Home and Its Linkage to Other Life Course Events in Turkey”, *Marriage & Family Review*, vol. 42, no. 1, pp. 29-47, 2007.

[52] I. Guyon and A. Elisseeff, “An Introduction to Variable and Feature Selection”, *Journal of Machine Learning Research*, vol. 3, pp. 1157-1182, 2003.

[53] G. Reinelt, “TSPLIB- A Traveling Salesman Problem Library”, *ORSA Journal of Computing*, vol. 3. no. 3, 1991.

[54] “scikit-learn Machine Learning in Python” Scikit-learn. https://scikit-learn.org/ (accessed Jan. 1, 2022).

[55] “Vehicle Routing Problem” Google. https://developers.goodle.com/optimaztion/routing/vrp/ (accessed Jan. 13, 2022).

[56] “tsp.py” Gurobi. https://www.gurobi.com/documentation/9.5/examples/tsp.py.html (accessed Feb. 9, 2022)

[57] “sklearn.svm.SVR” Scikit-learn. https://scikitlearn.org/stable/modules/generated/sklearn.svm.SVR.html (accessed Jan. 14, 2022.

CURRICULUM VITAE

Personal Information

Name Surname :Tayyip TOPUZ

Education

Bachelor's Degree : Kadir Has University, Computer Engineering, 2014-2019
Third best student in Computer Engineering
Second best final project (Geo-Social Case Analysis)
Kadir Has Honor Scholarship 60%
Kadir Has Honor Scholarship 100%
Kadir Has Honor Scholarship 10%

High School : Maçka Akif Tunçel Anatolian Technical High School, Web Programming, 2010-2014

Work Experince

2021- Present :Kadir Has University, Teaching Assistant