COMPARISON OF DIFFERENT CUSTOMER SEGMENTATION MODELS FOR THE FINANCIAL SECTOR VERSUS THE TURKISH FACTORING SECTOR AND A CLUSTERING MODEL PROPOSAL FOR A SME FOCUSED FACTORING COMPANY

A Thesis

By

Mehmet Aktuna

Submitted to the

Graduate School of Sciences and Engineering
In Partial Fulfillment of the Requirements for
the Degree of

Master of Science

in the
Department of Financial Engineering

Özyeğin University
December 2022

COMPARISON OF DIFFERENT CUSTOMER SEGMENTATION MODELS FOR THE FINANCIAL SECTOR VERSUS THE TURKISH FACTORING SECTOR AND A CLUSTERING MODEL PROPOSAL FOR A SME FOCUSED FACTORING COMPANY

Approved by:

Assistant Professor Levent Güntay,
Advisor,
Department of International Finance
Özyeğin University

Assistant Professor Emrah Ahi,
Department of International Finance
Özyeğin University

Professor Özgür Özlük,
Department of Industrial Engineering
MEF University

Date Approved: 28 December 2022

To my family,

To my colleagues,

To all my teachers from elementary school to now

with respect and thanks…

ABSTRACT

Factoring is the process of transferring a forward receivable arising from commercial transactions to a factoring company by assignment. Companies that prefer to do factoring, instead of waiting for the due date of the collection of their receivables arising from their trade, transfer their receivables to the factoring company with a certain discount, with documents such as post-dated cheque, trade invoices and factoring contracts. The use of cheque in Turkey is different from that in the world due to the post-dated maturity written on the cheque and SMEs in Turkey frequently use cheques in their trade with maturity. In terms of segmentation, there are many studies for banks in the literature, but there is no study in the SME focused factoring segment. This study is the first, where a two-stage customer segmentation is implemented using the data of a SME focused Turkish factoring company.

When the clusters generated by the K-Means and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) models are compared, it is observed that the clusters of the K-Means algorithm are more balanced in terms of distribution of the number of customers across clusters. On the other hand, the clusters generated by the DBSCAN model are more homogeneous, especially for outlier clusters. Homogeneity means an effective segmentation, which is there are few outliers within a cluster as defined by the Mahalanobis distance of each observation. The DBSCAN model can generate several homogeneous clusters containing a small number of customers if model parameters are appropriately calibrated. On the contrary, if few number of clusters are preferred, the homogeneity of the DBSCAN algorithm gets worse and the K-Means model gives more balanced clustering results.

ÖZETÇE

Faktoring işlemi, ticari alışverişten doğan vadeli bir alacağın temlik edilmesi yolu ile bir faktoring kuruluşuna devredilmesi işlemidir. Faktoring yapmayı tercih eden firmalar, yapmış oldukları ticaretlerinden doğan alacaklarının tahsilatının vadesini beklemek yerine, ticarete konu bir ödeme araç olan vadeli çek, ticaret faturası ve faktoring sözleşmesi gibi dokümanlar ile aracı faktoring kuruluşuna belirli bir iskonto karşlığında devretmekte ve nakit akışı ihtiyaçlarını düzenlemeyi amaçlamaktadırlar. Türkiye'de kobiler tarafından yaygın bir şekilde kullanılan çek, üzerinde yazan ileri tarihli vade nedeniyle dünyadaki genel kullanımından farklılaşmaktdır. Literatürde bankalar için yapılmış bir çok segmentasyon çalışması mevcuttur ancak kobi sektörüne yönelik bir faktoring müşterisi segmentasyon çalışması bulunmamaktadır. Bu çalışmada faktoring sektöründe faaliyet gösteren bir Türk faktoring firması datası kullanılarak iki aşamalı bir müşteri segmentasyon çalışması yapılmıştır.

Çalışmada kullanılan K-Means ve DBSCAN algoritmaları karşılaştırıldığında, K-Means algoritmasının, kümeler içerisindeki müşteri adetlerinin dağılımı anlamında daha dengeli olduğu gözlemlenmiştir. Öte yandan, DBSCAN modeli tarafından üretilen kümeler, özellikle aykırı kümeler için daha homojendir. Her bir gözlemin Mahalanobis mesafesi ile tanımlanan homojenlik, küme içerisinde daha az aykırı gözlem olduğu zaman etkinlik anlamına gelmektedir. DBSCAN modeli, model parametreleri uygun şekilde kalibre edilirse az sayıda müşteri içeren birçok homojen küme oluşturabilir. Aksine az sayıda küme tercih edilirse DBSCAN algoritmasının homojenliği bozulmakta ve K-Means modeli daha dengeli kümeleme sonuçları vermektedir.

ACKNOWLEDGMENTS

I would like to thank my advisor Assistant Professor Levent Guntay for his contributions and guidance's at every stage of my thesis, my colleagues who contributed to my business knowledge up to this point, my family for their patience and support, and my teachers for all their efforts on me from primary school to the present.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

CHAPTER I

INTRODUCTION

In the literature, different segmentation models have been developed for banks. But in the factoring sector, there is no segmentation study especially for SME focused companies. The K-Means model is generally used in the banking sector. Besides, Customer Life Time Value, which is carried out for the financial gain of the customer throughout his life, and the Recency, Frequency, Monetary Value Model (RFM) is also used frequently.

The use of cheques in Turkey is different from the use in the world. Therefore, a different segmentation study is implemented in this study for a SME focused factoring company operating in Turkey. In this respect, this study is a first in the literature.

In this study, 92.757 different SME Company (customer) who applied to a factoring firm between January 2021 and November 2022 and 761.891 cheque data belonging to different 163.805 drawers are subject to this application data were used. Since a customer, and therefore a drawer, may have been the subject of an application more than once in the mentioned date range, the last observation they came during this period was based on the study in order to get the most up-to-date data. The customer data used includes 40% of factoring customers in Turkey in 2021.

In the second chapter, the factoring transaction process is explained comparatively with the bank. Profiles of factoring customers and drawers in the SME segment are detailed in this section. In the third chapter, findings of the literature research are explained. Turkish SME focused factoring company data is explained in the fourth chapter. In the fifth chapter, the used methodology is explained. In the sixth chapter, the drawers and customers are segmented after variable selection. The clusters are interpreted in this chapter.

CHAPTER II

FACTORING AND FACTORING CUSTOMER

Factoring transactions include three parts as seen in Figure 1;

- The factoring company,
- The customer (can be call as seller in terms of factoring literature),
- The cheque drawer (buyer).

In factoring, the customer needs and applies for financing to the factoring company, and in case the loan transaction is approved and realized, the money outflow is made to the customer's account. On the due date, the factoring firm collects its repayment from the drawer's account. Figure 1 shows that the transaction process proceed in 3 steps.

In Step 1 a trade has taken place between the customer and the drawer (buyer of the trade). In return for this trade, the drawer, instead of paying cash to the customer (seller), writes a cheque with a future dated maturity date on it. Suppose there is a cheque that promises a repayment 90 days after the trade. Buying goods with this method, the drawer undertakes to pay the price of the goods to the seller 90 days after the transaction to the customer (seller).

In Step 2, the customer (seller of the trade) applies to a factoring firm with this cheque and demands cash in return for a certain interest deduction. In this application process, the drawer's is not involved at all, the factoring firm only comes face to face with the customer.

In Step 3, when the cheque is due at maturity date, the factoring firm will take the repayment from the drawer's account, not from the customer. Therefore, the factoring firm's ability to collect the repayment depends on the drawer's strength or financial position on the payment day. So, the factoring firm should consider the drawer's financial and behavioral characteristics at the time of application, (especially on the risk assessment side). Because when the cheque is due, it is important whether there will be money in the account of the drawer who bought the goods in the first day of trade (not the customer to whom the factoring company made the payment on the first day). However, the interest deduction to be applied by the factoring company will be accepted or rejected by the customer. Therefore, in the transaction, the factoring firm should have a structure that evaluates the customer as well as the drawer.

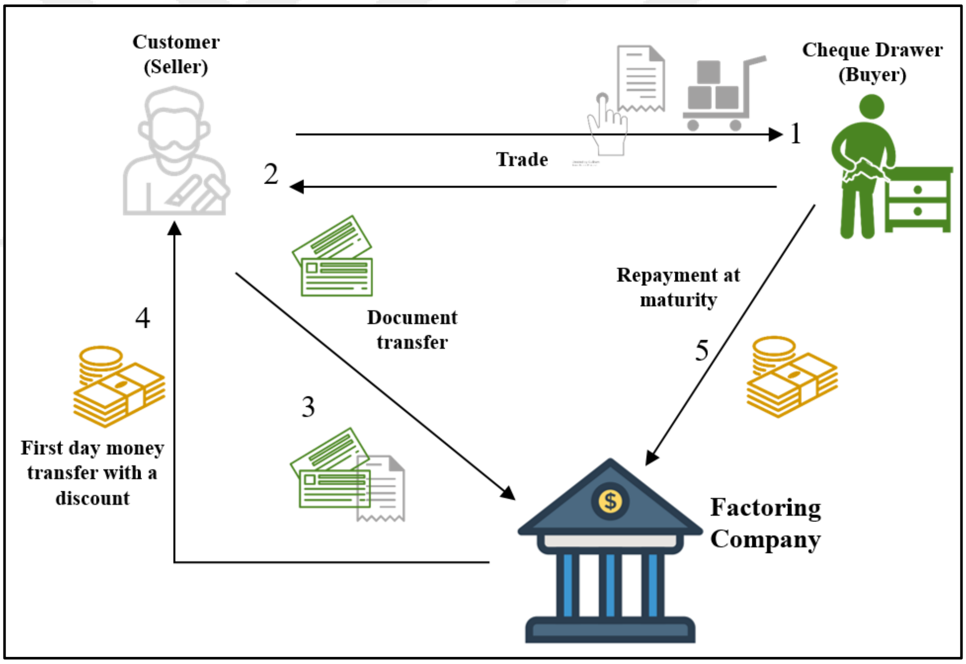

Figure 1: Factoring Company and Customer Transaction Cycle

Figure 2 shows the relationship between a customer who wants borrow a bank loan and the bank. There are only two parts in this transaction;

- The bank,
- The customer

The bank makes the money outflow to the customer account on the day the loan will be used, and on the day the loan is due (at maturity or installment days), the customer is obliged to make the repayment or payments to the bank. The customer's data at the time of application or the transactions that he or she has made with the bank in the past period is the main factor in the assessment of the bank.

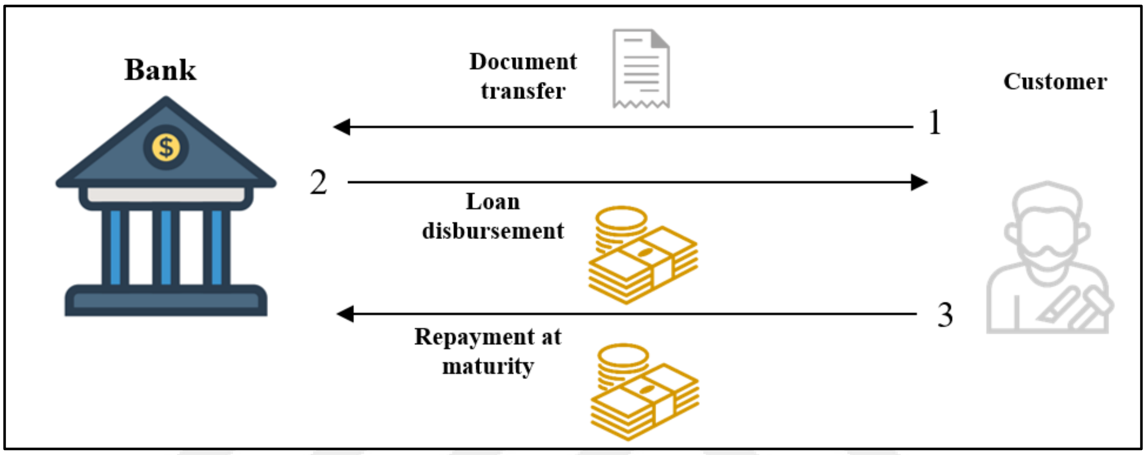

Figure 2: Bank and Customer Transaction Cycle

There are also differences between customer and drawer's profile of a factoring transaction. Figure 3 shows bank limit statistics of 87.514 customers who applied to the factoring firm between January 2021 and August 2022. The bank limits of 40% of these customers at the time of application are below 50.000 TL which is insignificant amount.

count	87,514
mean	914,885
std	6,012,709
min	0
25%	6,000
50%	106,627
75%	537,150
max	819,148,340

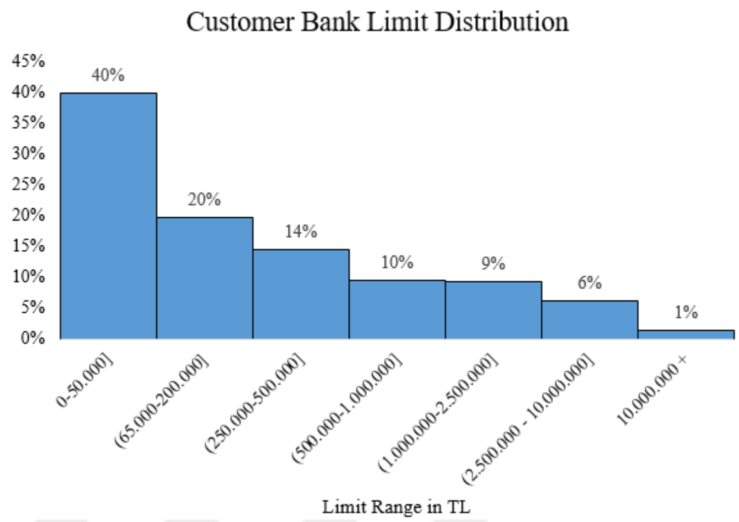

Figure 3: Histogram of the Distribution of the Customer Bank Limits

In the same applications, there are 154.507 drawers. Figure 4 shows drawer bank limit statistics. Unlike customers, only 6% of the buyers have a bank limit of less than 50.000 TL and about 57% of them have a bank limit of 1 million TL and above.

count	154,507
mean	14,862,477
std	221,362,005
min	0.00
25%	430,000
50%	1,300,900
75%	4,416,620
max	60,109,551,194

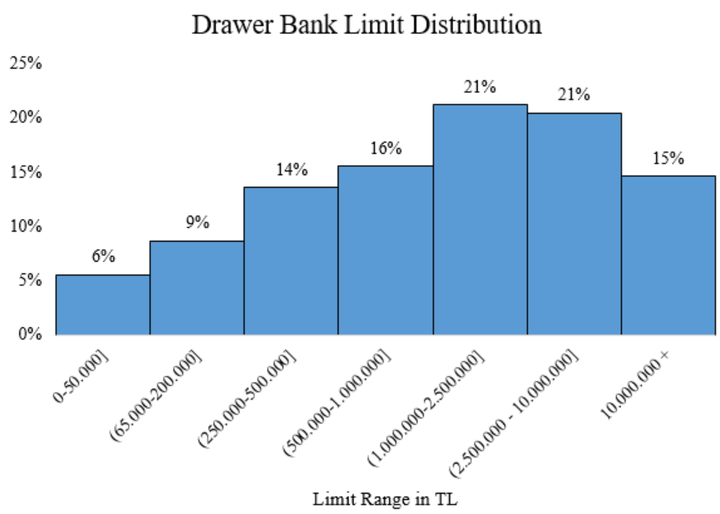

Figure 4: Histogram of the Distribution of the Drawers' Bank Limits

Table 1 shows both data comparatively. The drawers have significantly 1525% higher limit in banks rather than customers.

Table 1: Comparison of Customer and Drawer in Terms of Bank Limits

	Customer	Drawer	Δ
mean	914,885	14,862,477	1525%
std	6,012,709	221,362,005	35.82
25%	6,000	430,000	70.67
50%	106,627	1,300,900	11.20
75%	537,150	4,416,620	7.22
max	819,148,340	60,109,551,194	72.38

Figure 5 and Figure 6, shows factoring customer and drawers bank limits and "credit utilization ratio" on a monthly and per company basis. As of October 2022, the average bank limit of the customers is 1.3 million TL and the drawer limits is 36.2 million TL. The limit usage ratio of customers is higher than that of drawers, which indicates that customers have a higher financing need and try to use their limits more effective.

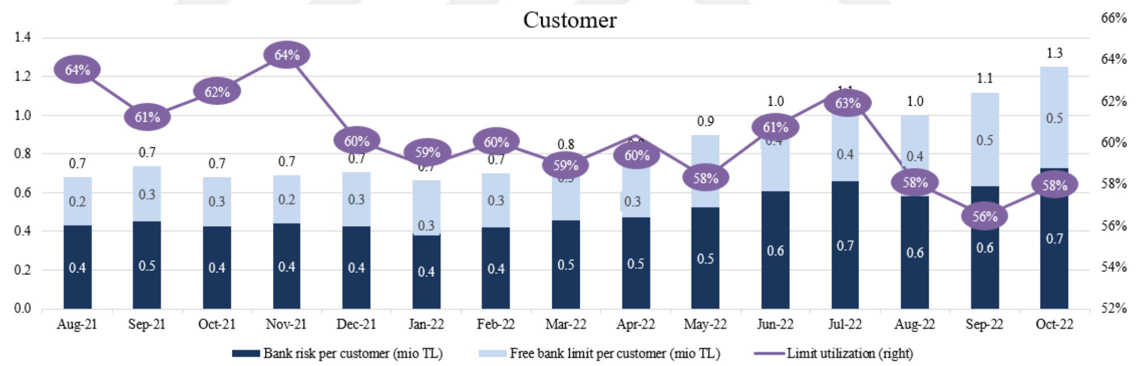

Figure 5: Trend of Monthly and per Customer Bank Limit and Limit Utilization Ratio

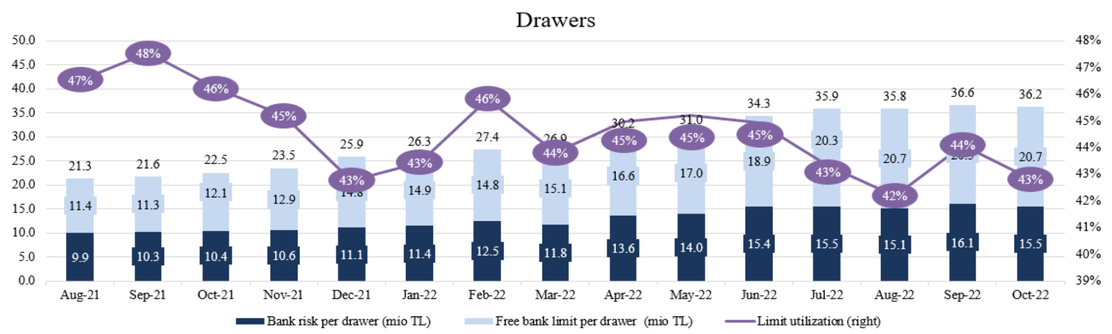

Figure 6: Trend of Monthly and per Drawer Bank Limit and Limit Utilization Ratio

Factoring is a more expensive product than bank loans. The basis of this is the ability of the customers to reach the bank loan. As seen in the right pie chart of Figure 7, 29% of the customers out of total 17.091 customers (that the factoring company made transactions for the first time in 2021), were established in the last 12 months. As a natural consequence of this, banks cannot make a risk assessment to these customers. This results in banks not allocating credit limits to these customers. Left pie chart of the Figure 7 also shows for the same customer groups bank limits on the application date. 39% of these customers do not have any bank limit on application date on which they can use loans from banks. So they need factoring financing, a more expensive product, in order to manage their cash flows.

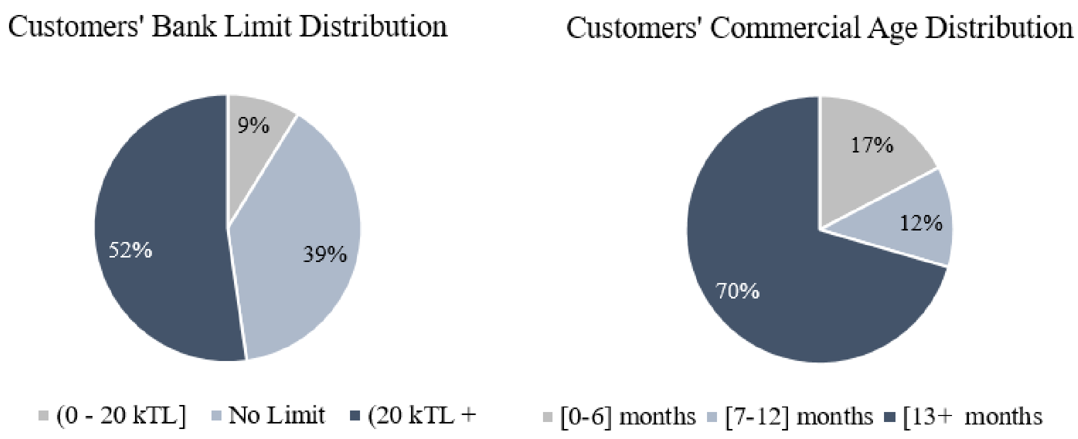

Figure 7: Customers’ Bank Limits and Age Distribution

Considering both the difference between the factoring bank processes and the fact that the customer and the drawers of a factoring transaction have different profiles from each other, in the factoring sector, drawers should also be segmented as well as customer segmentation.

CHAPTER III

LITERATURE RESEARCH

In sales, commerce and economics, a customer (sometimes known as a client, buyer, or purchaser) is the recipient of a good, service, product or an idea, obtained from a seller, vendor, or supplier via a financial transaction or exchange for money or some other valuable [1]. This definition consists of two sides:

- Customer (sometimes known as a client, buyer, or purchaser)
- Seller, vendor, or supplier

Seller, vendor, or supplier, can be generally called as companies. All companies aims to get benefit within this existing relationship, they also try to maximize this benefit.

Another important element in the definition is good, service, product or an idea that bring these two parties together and make a relationship. These elements bring together the parties of the relationship at a common point. Maximizing the utility is accomplished through these common elements. The customer will prefer the company that provides the most benefit from the products him /her purchases. Besides, the more sales companies can make, the more they will be able to increase their benefits. But the following questions are also important:

- Does every customer have the same level of risk?
- Does every customer expect the same from a product?
- Is every customer has the same need for every product?
- Is every customer always willing to give the same price for every product?

The answer to all these and similar questions reveals the concept of categorizing customer. Segmentation or clustering is grouping a collection of objects into subsets or "clusters," such that those within each cluster are more closely related to one another than objects assigned to different clusters [2]. So, customers need to be grouped for benefit maximization purposes. A company that wants to categorize its customers to achieve the following objectives:

- Price to apply,
- Campaign to offer,
- Maintain and increase customer loyalty,
- Measuring the level of risk.

Algorithm is a set of rules that precisely defines a sequence of operations for solving a given problem. It starts from an initial input of instructions that describe a computation that proceeds through a finite number of well-defined successive states, producing an output and a final ending state [3]. In simpler terms, it is a procedure for solving a mathematical problem in a finite number of steps that frequently involves repetition of an operation by using computer science. Basically, algorithm is a procedure or formula for solving a problem, based on conducting a sequence of specified actions. Computer science is the study of algorithmic processes and computational machines but it is not just about computers. Computers can only be considered as an assistant element in big and time-consuming operations that are difficult to calculate by human hand. Beside this, each algorithm has a structure that can be explained on paper without the aid of a computer.

In order to better understand that why customer segmentation should be done with the help of algorithm, a few examples are given below:

- The number of internet users in only in the United States who are Amazon Prime members was 124 million in 2019, expected to be 142.5 million in 2020 and 153.1 million in 2022 [4]
- Facebook has over 2.9 billion monthly active users as of the second quarter of 2022 [5]
- Coca-Cola reached to 2.2 billion in Europe, 650 million in Latin America, 360 million in North America, 4.5 billion in Asia Pacific, different consumers in 2019 [6]
- AliBaba’s annual active consumers on online shopping properties in China reached 903 million by the first quarter of 2022 [7]
- Spanish banking giant Banco Santander reached to at approximately 153 million customer in 2019 [8]

According to the examples given in different sectors, the number of customers reached can be expressed in millions or even billions and when we convert these numbers to transaction basis and it is obvious that manual calculations will not work effectively. For example, in the financial sector where price differentiation between customers is an important issue, it is not possible for a bank with 153 million customers to decide which customer it will offer which price by manual methods. So, algorithms should be used.

Machine learning (ML) is the study of computer algorithms that improve automatically through experience [9]. Machine learning algorithms build a model based on sample data, known as "training data", in order to make predictions or decisions. The algorithms used in the customer categorization process also have an important place in machine learning. How will machine learning algorithms can find solution from the training data offered to produce results? Machine learning, as the name suggests, is a learning process, and this process can be categorized in two ways, either supervised or unsupervised. In supervised learning, the goal is to predict the value of an outcome measure based on a number of input measures; in unsupervised learning, there is no outcome measure, and the goal is to describe the associations and patterns among a set of input measures [2]. So, unsupervised learning is descriptive and supervised learning is predictive; four different types of analytics are:

- Descriptive analytics: What happened in the past?
- Diagnostic analytics: Why did an event happen?
- Predictive analytics: What is most likely to happen next?
- Prescriptive analytics: What is the most optimal action given the event happened?

In the supervised learning system, there is a data so that the algorithms provides what output might occur next (predictive); however, in unsupervised learning models, algorithms aim to only create similar clusters based on the data.

The use of unsupervised or supervised study depends on the structure of the data available. For the machine learning algorithm, the training data must be suitable for a supervised study. In supervised models, the variable targeted for prediction, has its results for each past observation within its set. This is a guiding principle that the algorithm can train and test on the data. Table 2 is a small example of bank with 10 different customers that we have information on age, gender and annual income of these customers as training set. The label column in the right table of Table 2 provides the opportunity to conduct a supervised study. For the left table, since this data does not contain information about customer’s payment performance, the algorithm will not be able to predict a customer's ability to pay according to their specifications. The algorithm developed will only explain the differences / similarities between groups or customers, hence it will be an unsupervised and descriptive model. But with a labeled data on the right table, which shows customer has successfully paid their loan in the past, or has been in trouble, machine learning algorithm can predict the probability of paying the loan of any customer regarding to the age, gender and annual income.

Table 2: Unlabeled Data versus Labeled Data

Unlabeled Data			
Customer ID	Age	Gender	Annual Income (TL)
1	25	Male	270,000
2	24	Male	300,000
3	32	Female	1,200,000
4	50	Female	7,500,000
5	33	Male	630,000
6	19	Female	190,000
7	25	Female	450,000
8	44	Male	750,000
9	65	Female	220,000
10	22	Male	175,000

Labeled Data				
Customer ID	Age	Gender	Annual Income (TL)	Default
1	25	Male	270,000	0
2	24	Male	300,000	0
3	32	Female	1,200,000	1
4	50	Female	7,500,000	0
5	33	Male	630,000	1
6	19	Female	190,000	0
7	25	Female	450,000	0
8	44	Male	750,000	1
9	65	Female	220,000	1
10	22	Male	175,000	0

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) [10]. Segmentation terminology is also used for cluster analysis. The aim of cluster analysis is to divide a collection of objects into subsets which are called clusters, where elements in each clusters are more similar to each other and differs most from other cluster elements. Figure 8 shows a group of clustered data into three separate groups such as blue, green and yellow together with an algorithm. In short, segmenting aims minimizing the differences in each clusters, but maximizing distances in each cluster.

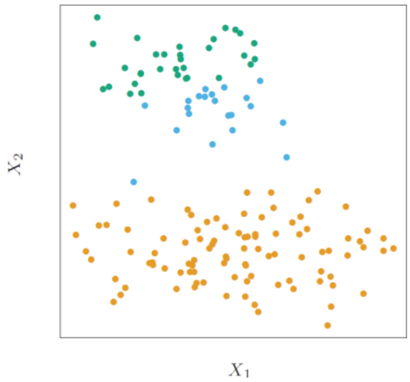

Figure 8: Clustered Customer Visual Example

Segmentation methods can be divided into classes in more than one classification according to the need. For example, a classification is made as partitional versus hierarchical clustering where hierarchical clustering algorithms can be summarized as iterative loop algorithms by combining small groups or, on the contrary, dividing large groups. New clusters are formed with groups that are similar or not similar to each other until optimization is achieved at each step on the other hand, partitional clustering starts with determining the number of groups to be created first. Then, optimization is tried to be achieved by the method of shifting the centers of these groups. K-Means is the most common example for the partitional clustering model.

Figure 9 visually shows the difference between partitional clustering and hierarchical clustering [11]. Those close to each other were clustered in three groups (which is pre-defined number of group) to provide optimization in partitional clustering then two yellow sub-group hierarchically grouped to yellow group.

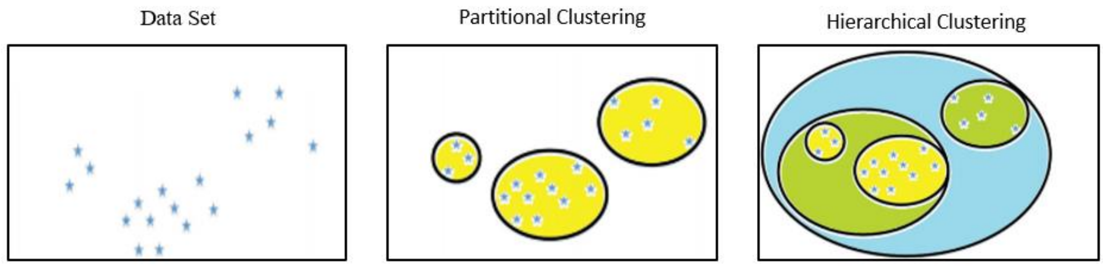

Figure 9: Partitional versus Hierchical Clustering Visual Example

Another classification is fuzzy versus non-fuzzy classifications. In non-fuzzy clustering methods, data is divided into sharp and specific clusters. An observation can only be in one cluster. However, in fuzzy models, an element can be in different clusters with a total weight of 1. Figure 10 visually shows the difference between non-fuzzy and fuzzy clustering [12]. All observations are distributed in cluster A or B with a weight of 1 or 0. In this model, however, the observation can take place between both groups according to the coefficients ranging from 0 to 1.

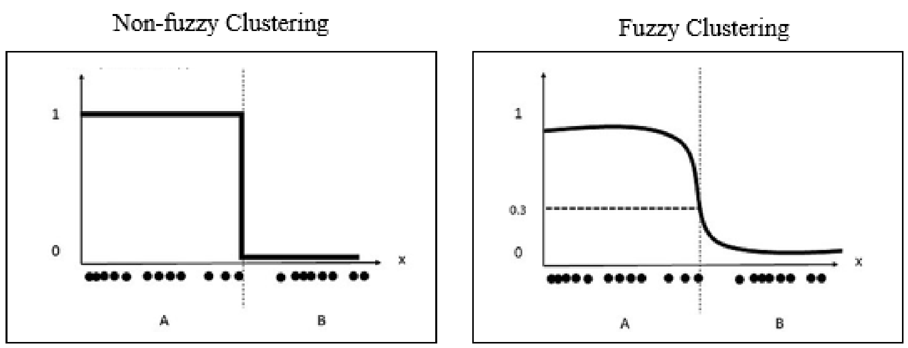

Figure 10: Non-fuzzy versus Fuzzy Clustering Visual Example

Partial versus complete clustering are another classification methods. As the name suggests, while complete cluster algorithms all elements will be assigned to a defined cluster, some elements are not included in a cluster in partial models. The main reason for this is the noise and outlier elements in the data set.

In this thesis, the following two clustering methods are implemented and results are interpreted:

- K-Means Clustering
- Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

Within the statistical results, a clustering algorithm is considered to be good if it satisfies the following requirements:

- Minimum requirements of domain knowledge,
- Making effective shaped clusters,
- Good efficiency on large data sets
- Easy to understand and interpret results

but none of the existing algorithms can meet these three basic needs [13]. Therefore, each algorithm has advantages and disadvantages.

K-Means clustering is one of the most used methods due to its ease of application. It starts with assigning each sample to a cluster by determining the closest distance between the sample and centroid [14]. The aim of the K-Means algorithm is to divide M points in N dimension into K cluster. In the simplest sense, the clusters each point to the k centers it has determined randomly at the beginning, then determines a new center for these clusters and makes the same clustering again. This process, which is like the most step-by-step iteration, continues until the group centers remain unchanged. [15]. In addition to being easy and fast to apply, the biggest disadvantage of this method is that it is not known in the first place how many K groups the group will be divided into. Also, the negative effect of outliers on this model is one of its disadvantages [16].

The key idea of Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is that for each object of a cluster the neighborhood of a given radius (Eps) has to contain at least a minimum number of objects (MinPts) [17]. This model groups objects according to specific density objective functions like other Density-based clustering algorithms. [13]. Thanks to the specified Eps and MinPts, this model carries noise and outliers together with the clusters it eventually forms. However, although this method is advantageous in determining the outliers, it is also possible to use methods other than clustering for determining outliers. In this direction, for the detection of anomalies; there is a study made by a Turkish factoring company using Mahalanobis distance, Minimum Covariance Determinant (MCD) and Orthogonalized Gnanadesikan-Kettenring (OGK) methods based on specific scenarios rather than clustering [18].

The literature research reveals that financial institutions, including banks, use the main clustering models described above, especially K-Means [19]. K-Means and DBSCAN models are also applicable for a factoring company. However, along with the customers, the drawers should also be taken into account at the segmentation stage. In this thesis, K-Means and DBSCAN algorithms are used in accordance with factoring needs.

In the segmentation studies for banks, Customer Life Time Value (CLV) and Recency, Frequency, Monetary Value Model (RFM) models are also frequently used. Since these two models do not contain complex algorithms, they are very effective methods and can be used also by factoring companies. Although they seem like separate methods, CLV and RFM are sometimes seen as intertwined and used together. As there are applications of this in the banking sector, there are similar studies in the literature [20]. CLV models, aim to measure the financial benefit that a customer will provide to the company throughout its life. The goal is estimating future value of different customer segments. This modeling can be done using many different methods, one of which includes outputting RFM parameters or RFM model itself. It means that, after calculating each customer segment value based on adapted weighted RFM model for example, using time series method for prediction of each segment future value (such as ARIMA), Customer Lifetime Value can be calculated and customers can be categorized. [21]. The RFM generates a score (for ex from 0 to 5) by ranking customers based on the parameters of when the customer last purchase (R), how often they purchase (F), and how much money they spend (M). The scores obtained from here can be used for purposes such as campaign management, pricing, or they can be used in CLV calculation. The RFM model, which is relatively simple and understandable, has been developed over time, and it has been tried to be made more effective by using different parameters according to the needs. For example, in the LRFM models, L (length, loyalty) parameter was tried to be added in the calculation, while in the LRFMP model, it was also taken into account whether the customers behaved regularly with the variable P (periodicity) [22].

On the other hand, this thesis proposes a two-stage clustering model. First, the drawers have been clustered, then the customers will are clustered by using the drawer clusters. Although there is no example is the banking sector, there are some two-stage clustering studies in different sectors. For example, in a research based on a Fintech company operating in the field of payment processor, the companies were first clustered with the K-Means and LRFMP method, and in the second stage, the clusters were re-clustered into subgroups to make sense of them. [23].

CHAPTER IV

DATA SET

Banks and factoring companies operating in Turkey have different licenses and are regulated by the Turkish official governmental intuition called Banking Regulation and Supervision Agency (BRSA). Due to the different licenses they have, the services they can offer to their customers and, accordingly, the information they can obtain about their customers through official institutions are also different from each other. In this respect, factoring companies have much more limited product possibilities, and on the other hand, they can access more limited information than banks. Factoring companies have access to the “cheque report” and “risk report” provided by the Credit Records Bureau (KKB) thanks to the tax identification number they will obtain from their customers (for example, through the tax certificate). Similarly, when customers apply to the factoring firm, the tax identification number of the drawer of that cheque is also on the cheque that they use. In this way, the factoring company has the opportunity to access the risk and check report of the drawer.

Among the mentioned reports, the risk report; it provides information about the financial results of companies arising from their relations with banks, factoring companies and leasing companies. The risk report, which contains nearly 100 different information, provides information on the following topics, with sub-details of each, and three different financial institutions separately (bank, factoring, leasing):

- The company tax number (identification),
- The company name,
- The company type, individual or commercial (Firms operating in Turkey can be divided into two; commercial and individual companies),

- The number of institutions that the company is associated with,
- The company's total credit limit information in financial institutions,
- The company's total credit usage in financial institutions (called "risk" in financial institutions terminology),
- The company's total legal follow-up credit (non-performing loan, NPL) in financial institutions,
- The company's first - last limit allocation date,
- The company's first last follow-up date,
- Individual Credit Score calculated by the Credit Records Bureau (KKB) on the past credit payment performance of an individual person (not the firm)

The information provided is generally given in terms of both number and amount, depending on the content. Factoring companies, unlike banks, cannot access the customer's first or last loan application date.

The Individual Credit Score is only available for individual companies and shows the credit rating of the owner of the company. It takes a value between 1 and 1900, where 1 indicates the weakest score and 1900 indicates the strongest. In order to have an individual credit score, it is necessary that the company have a history of providing data within financial institutions related to enabling credit scoring, such as credit or credit card usage. For this reason, the individual credit score in the database has been made categorical in the following breakdowns with the help of domain information; (N/A, 1-500, 501-999, 1000-1399, 1400+)

On the other hand, the cheque report, contains nearly 100 data like the risk report, another service provided by the Credit Records Bureau, provides information about the payment performance of the cheques used by the firms only in their trade. Therefore, due to the nature of the business, this report does not have a breakdown by bank, leasing or factoring company. The use of cheque in its trade by a company and using factoring finance are completely independent from each other. While factoring includes only cheques given to factoring companies for financing purposes, cheque information provides information about all checks issued. In summary, factoring is a small subset of the cheque used in Turkey. The information included in the cheque report provided by the Credit Records Bureau can be summarized as follows:

- The company tax number (identification),
- The company name,
- Total number and amount of cheques due in the past (with 3, 6, 12 and 12 + month breakdowns)
- Total number and amount of cheques paid successfully in the past,
- Total number and amount of cheques defaulted and not paid in the past,
- Date of first and last cheque paid successfully in the past,
- Date of first and last check unpaid in the past,
- Total number and amount of cheques to be paid in the future (with 3, 6, 12 and 12 + month breakdowns),
- Cheque Index (Score) calculated by the Credit Records Bureau (KKB) on the past 36 months cheque payment performance of the company.

While the Check Index takes values between 0 and 1000, the absence of a firm's Cheque Index and its zero are two different situations. Cheque Index is not calculated for users who have not drawn (used) 3 or more checks in the last 3 years. Like the Individual Credit score, the Cheque Index in the database has been made categorical in the following breakdowns with the help of domain information; (N/A, 0, 1-500, 501-750, 751-999, 1000), where 0 means all cheques in the past were not paid on time (at maturity or at due date) and 1000 means all cheques were paid on time.

Apart from the cheque report and the risk report, Credit Records Bureau provides factoring companies the opportunity to access data on cheque that were due but not paid the previous day throughout Turkey every day. In addition to the various information contained in this data, for each cheque following data are also available:

- Cheque identification number,
- Company (drawer) tax number (identification),
- Cheque amount,
- Cheque due (maturity) date,

Thanks to the cheque identification number and the tax number of the drawer, the factoring firm has the information on whether or not each cheque application in the past has been paid, whether it accepts the application or not. This information is an important resource for a supervised modeling later in the thesis. The same data also shows the cheque information that was not paid on due date but was paid later. Since no data is shared about cheque paid on due date by Credit Records Bureau, cheque that are not included in this report are considered to be paid on due date. Table 3 shows the summary of the monthly report published by the Turkey Risk Center. According to this report, a total of 152.576 cheque were not paid on due date in 2021, and 40.026 of them were paid later [24] . Based on these data, it is seen that for the year 2021, the Credit Records Bureau shares an average of 850 cheque data per day (in 226 working days).

Table 3: Turkey Cheque Payment Performance in 2021

Yıllar	Aylar	Bankalara İbraz Edilen Çekler (1)			Karşılıksız İşlemi Yapılan Çekler (2) (*)			Karşılıksız İşlemi Yapıldıktan Sonra Ödenen Çekler (3)			Karşılıksız İşlemi Yapılan Çekler / Bankalara İbraz Edilen Çekler (%)		
		Adet	Tutar (Milyon TL) (**)	Tekil Kişi Sayısı (4)	Adet	Tutar (Milyon TL) (**)	Tekil Kişi Sayısı (4)	Adet	Tutar (Milyon TL)	Tekil Kişi Sayısı (4)	Adet	Tutar (**)	Tekil Kişi Sayısı
2021	1	751,233	75,265	187,864	7,559	681	2,320	1,768	158	1,075	1.0	0.9	1.2
	2	1,133,048	88,979	255,952	10,547	865	3,066	2,748	214	1,537	0.9	1.0	1.2
	3	1,632,974	121,634	288,168	15,233	1,172	3,689	3,938	256	1,897	0.9	1.0	1.3
	4	1,070,007	93,690	247,801	7,610	674	2,441	1,693	144	1,120	0.7	0.7	1.0
	5	1,140,919	91,901	258,514	-	-	-	-	-	-	0.0	0.0	0.0
	6	1,392,280	119,841	293,439	31,009	2,498	5,995	9,857	676	3,594	2.2	2.1	2.0
	7	949,023	92,174	240,618	12,108	1,082	3,530	3,130	258	1,855	1.3	1.2	1.5
	8	1,588,862	137,059	296,439	17,233	1,499	4,293	4,549	340	2,263	1.1	1.1	1.4
	9	1,310,731	125,017	294,276	13,590	1,318	3,789	3,440	277	1,869	1.0	1.1	1.3
	10	792,426	93,730	205,751	8,184	892	2,792	1,879	180	1,238	1.0	1.0	1.4
	11	1,930,062	176,561	316,074	16,755	1,684	4,222	4,283	378	2,126	0.9	1.0	1.3
	12	1,466,354	163,145	309,539	12,746	1,470	3,588	2,741	313	1,625	0.9	0.9	1.2
	2021 Toplam	**15,157,919**	**1,378,995**	**477,212**	**152,576**	**13,836**	**16,796**	**40,026**	**3,194**	**11,483**	**1.0**	**1.0**	**3.5**

As of September 2022, there are 49 different companies operating in Turkey. While 12 of these companies are subsidiaries of banks, the remaining 37 companies are called non-bank factoring companies. The SME focused factoring company used in the thesis is also takes place among the non-bank factoring companies. According to the BRSA reports, the total amount of factoring receivables on the balance sheets of factoring companies in 2021 year and is TL 59.9 million [25]. Considering the amount of cheques due in Turkey in 2021, which is in the second column of Table 3, it is observed that less than 1% of the cheques used in Turkey are subject to factoring. So, factoring is a small subset of the cheque used in Turkey. On the other hand, according to the Risk Center data, 92.000 different companies made factoring transactions in 2021. The factoring company, whose data are used in the thesis, made transactions with 36.000 different factoring companies in 2021 and made transactions with 39.1% of the factoring customers in the sector. According to this data, the data used includes the data of a very important part of the factoring companies in Turkey. On the other hand, in terms of factoring receivables (total 59.9 million in sector), the factoring firm has a share of only 3.2% in the sector. This is due to the fact that the factoring firm carries out SME focused transactions, that is, with small amount of transactions, but a large number of transactions.

The whole study of this thesis is based on the data of between January 2021 and November 2022, where 92.757 different customers applied to the factoring company with 761.891 cheques. There are 163.805 different drawer cheques in these applications. As of November 2022, 643.480 of the same cheques are due, so the factoring firm has the information whether these cheques have been be paid or not. 23.130 of these cheques have not been paid on due date (4.7%). This data is used in the supervised study in determining variables of the drawers in the later parts of this thesis.

Out of total of approximately 200 variables, as stated above, 45 variables in Table 4 and Table 5 were selected from with the help of business knowledge. Due to their different profiles, the same variables were not used in customer segmentation and drawer segmentation. So, the drawers were segmented and these clusters were used as a variable in customer segmentation. At this stage, while determining the drawer variables; thanks to the cheque payment information provided by the Credit Records Bureau, a supervised study is conducted, but variables have been selected with business information on the customer side.

The Python program and the following libraries were used in the study: Pandas, Numpy, Sklearn, Matplotlib, Min Max Scaler, Plot Confusion Matrix, Train Test Split, Decision Tree Classifier, Random Forest Classifier, Kneighbors Classifier, SVC, XGB Classifier, Extra Trees Classifier, Accuracy Score, Classification Report, Logistic Regression, Cross Val Score, Cx Oracle, Os, Tree, Metrics.

Table 4: Drawer Variables Selected with Domain Information

Variable	Explanation	Source	Drawer	Customer
CEKENDEKSI	The firm's check score produced by Credit Bureau (based on past cheque payments)	Cheque Report	√	
KSD_IBRAZ_EDILEN_ILK_CEK_GUN	The date between the first cheque paid by the drawer and the application date in days	Cheque Report	√	
BKTOPLAMTAKIPBAKIYESI	Total number of non-performing personal loans of the drawer in TL	Cheque Report	√	
SON3BTURUCEKSAYISI	Total number of cheques not paid on due date of the issuer in the last three months in TL	Cheque Report	√	
KSD_TK_ILK_KREDI_KULLANIM_GUN	The number of days between the drawers first loan usage date and the application date in days	Cheque Report	√	
KSD_TK_SON_KREDI_KULLANIM_GUN	The number of days between the drawers first loan usage date and the application date in days	Cheque Report	√	
KSD_ODENEN_SON_CEK_GUN	The number of days between the date of the last paid check by the drawer and the date of application in days	Cheque Report	√	
KARSILIKSIZBTURUTOPLAMCEKSAY	The number of checks that the drawer did not pay in the past at matuirty	Cheque Report	√	
KARSILIKSIZBTURUTOPLAMCEKTUT	The amount of cheques that the drawer did not pay in the past in TL	Cheque Report	√	
TOPLAMODENMISCEKSAYISI	Total number of cheques matured and the drawer has paid successfully in the past	Cheque Report	√	
SON3ODENENCEKSAYISI	Total number of cheques the drawer has paid in the last 3 months	Cheque Report	√	
KSD_ILERI_3AY_ODEYECEGICEKADEDI	Total number of cheques the drawer will pay in the next 3 months	Cheque Report	√	
TOPLAMVADELICEKADEDI	Total number of checks the drawer will pay in the future	Cheque Report	√	
TOPLAMCEKSAYISI	Total number of cheques which has matured in the past whether it is paid or not	Cheque Report	√	
KSD_ILERI_3_AY_CEK_TUTAR_ORAN	Total number of cheques the drawer has paid in the last 3 months / the total number of cheques the drawer will pay in the next 3 months	Cheque Report	√	
KSD_ILERI_3AY_ODEYECEGICEKTUT	Total value of cheques the drawer will be bay in the next 3 months in TL	Cheque Report	√	
SON3ODENENCEKTUTARI	Total value of cheques the drawer has paid in the last 3 months in TL	Cheque Report	√	
KSD_ILERI_3_AY_CEK_ADET_ORANI	Total value of cheques the drawer has paid in the last 3 months / the total value of cheques the drawer will pay in the next 3 months	Cheque Report	√	
TOPLAMODENMISCEKTUTARI	Total value of cheques the drawer has paid in the past at the maturity date in TL	Cheque Report	√	
BKKREDINOTU	Credit score generated by the individual firm's past loan payment produced by Credit Bureau	Risk Report	√	
KSD_GERCEKTUZEL_T	Firm type (individual/commercial)	Risk Report	√	
KSD_TK_EN_GUNCEL_LIMIT_TAHSIS_GUN	The day between the last limit date allocated to the drawer by the bank and the application date in days	Risk Report	√	
BKTAKIPBILBULKRMSAY	Number of financial institutions with non-performing loan in individual loans	Risk Report	√	
KSD_LEASING_TOK_TOPLAM	The drawers total leasing usage in TL	Risk Report	√	
KSD_FAK_TOK_TOPLAM	The drawers total factoring usage in TL	Risk Report	√	
BKKURUMSAYISI	Number of financial institutions that the drawer has financial statements on individual loans	Risk Report	√	
BKKREDILIHESAPSAYISI	The number of financial accounts of the drawer in individual loans	Risk Report	√	
KSD_TOPLAM_BANKA_LIMIT_TUM	Total bank limit of the drawer in TL	Risk Report	√	
KSD_BANKA_LIMIT_KULLANILIM_ORAN	The drawers' bank loan usage ratio (bank loan/ bank limit)	Risk Report	√	

Table 5: Customer Variables Selected with Domain Information

Variable	Explanation	Source	Drawer	Customer
STC_BKKREDILIHESAPSAYISI	The number of financial accounts of the customer in individual loans	Risk Report		√
STC_BKTOPLAMTAKIPBAKIYESI	Total amount of non-performing personal loans of the cstomer in TL	Risk Report		√
STC_BKKREDINOTU	Credit score generated by the individual firm's past loan payment produced by Credit Bureau	Risk Report		√
STC_TOPLAM_BANKA_LIMIT_TUM	Total bank limit of the customer in TL	Risk Report		√
STC_TOPLAM_BANKA_RISK_TUM	Total bank loan of the customer in TL	Risk Report		√
STC_LEASING_TOK_TOPLAM	Total leasing usage of the customer in TL	Risk Report		√
STC_LEASING_TAK_TOPLAM	Total leasing non-performing loan of the customer in TL	Risk Report		√
STC_FAK_TOK_TOPLAM	Total factoring usage of the customer in TL	Risk Report		√
STC_FAK_TAK_TOPLAM	Total factoring non-performing loan of the customer in TL	Risk Report		√
STC_BANKA_LIMIT_KULLANILIM_ORAN	The customers' bank loan usage ratio (bank loan/ bank limit)	Risk Report		√
STC_SON_KANUNI_TAKIP_GUN	The day between the last legal follow up dade of the customer and the application date	Risk Report		√
STC_TK_EN_GUNCEL_LIMIT_TAHSIS_GUN	The day between the last limit date allocated to the customer by the bank and the application date	Risk Report		√
STC_BK_SON_KREDI_KULLANIM_GUN	Number of days between the customers last commercial loan usage date and the application date	Risk Report		√
STC_TK_SON_KREDI_KULLANIM_GUN	Number of days between the customers last personal loan usage date and the application date	Risk Report		√
CALISTIGI_KURUM_TOPLAM	Number of financial institutions that the customer has financial statements on individual loans	Risk Report		√
TAKIP_KURUM_TOPLAM	Number of financial institutions with non-performing loan in individual and commercial loans	Risk Report		√

CHAPTER V

METHODOLOGY

The aim of the thesis is to segment the customers of a factoring firm effectively. However, this process is carried out in two steps; in Step 1, the drawers subject to the factoring transaction have been segmented, and in Step 2, the customers have been segmented by using these drawer segmentations. In both segmentation processes, unsupervised work have been carried out and K-Means and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) models were used. Figure 11 simply illustrates these two steps.

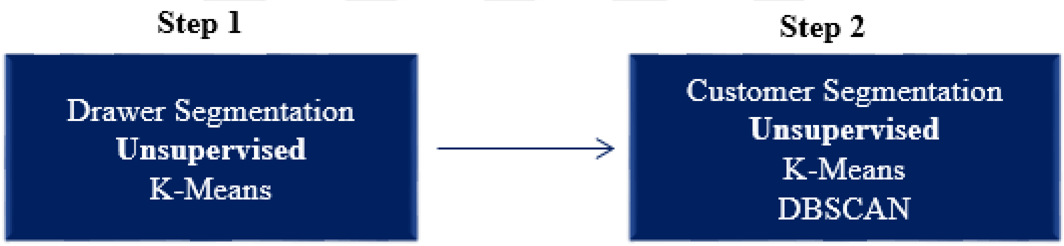

Figure 11: Two Steps of Customer Segmentation

Different variables were used in customer and drawer segmentation, both due to the data structure available and because customers and drawers have different profiles from each other. While determining the variables used in Step 1 of Figure 11, that is, in the drawer segmentation, a supervised study was implemented first, as it is known whether the cheque used in the past applications are paid by the drawers. To determine variables for the drawers, 6 different algorithms (Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Extra Trees Classifier, K-Nearest Neighbor's Classifier and XGBoost Classifier) will be used. As stated in the dataset chapter, 643.480 cheque data which has matured (which means that can be labeled as 0 or 1) will be allocated as 70% training sample and remaining 30% test sample. The results of the test data for each 6 model will be compared and significant variables have been used in the drawer segmentation according to the most appropriate model results.

While determining the variables to be used in Step 2 in the Figure 11, that is, in the customer segmentation, variables have been selected with the help of domain information, since the cheque payment performance data which allows to conduct supervised work is not meaningful to the customer. However, in customer segmentation, the drawers segmented in the previous step also added as a variable among other variables. Therefore, the Figure 11 summarized above is like the Figure 12 below when we add these variable selection steps:

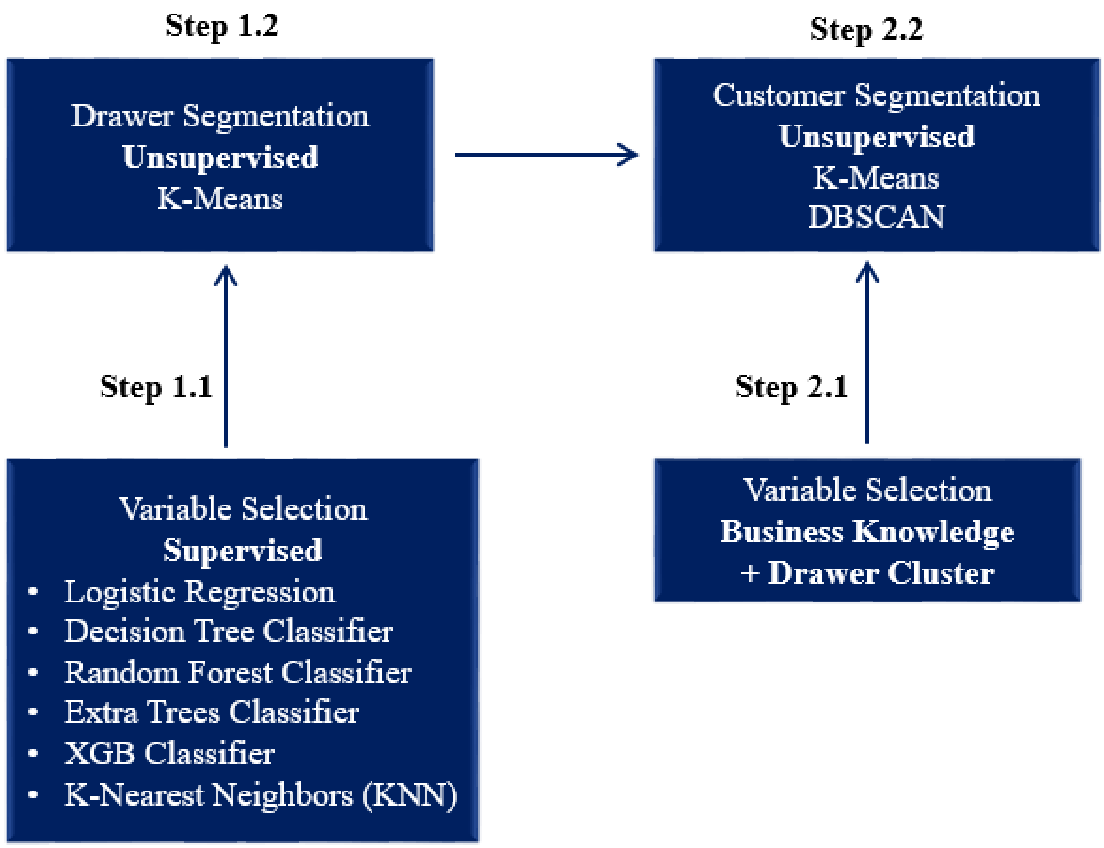

Figure 12: Two Steps of Customer Segmentation Including Variable Selection Process

Figure 13 shows the same process in three steps including the models used, the methods used, data source and evaluation processes. The results obtained by customer and drawer segmentation are interpreted separately.

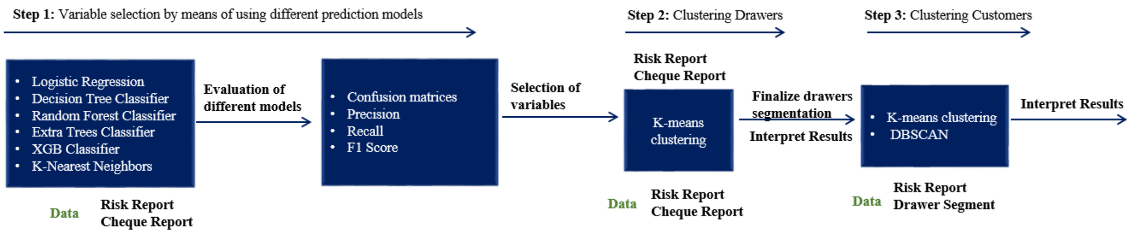

Figure 13: Flowchart of the Methodology

CHAPTER VI

RESULTS OF CLUSTERING ANALYSIS

6.1 Clustering Drawers

6.1.1 Variable Selection for Drawer Clustering

Before clustering drawers, variables to be used must be selected. The factoring firm can access more than hundred data belonging for both customer and drawer from Credit Records Bureau at the time of an application. Apart from the domain information about which of these data to be used, there is also the possibility for a statistical model for drawers. Even the drawer segmentation is a clustering (unsupervised model), the factoring firm has the information whether the matured 643.480 cheques in the past applications are paid or not when they are due (at payment day). Therefore, with the help of a supervised forecasting model, it is possible to gain knowledge about the factors affecting a drawers' payment performance, which are the variables used for the segmentation model. Then, Step 1 determines the best supervised model, using a randomly selected training sample, 70% of the total data, and 30% of which is the rest part of the data as the test sample. Step 2 chooses the set of independent variables for the estimation of this model.

The following six commonly used prediction models have been studied with their basic features explained below.

Logistic Regression: Logistic regression is a method of estimating a binary outcome with the help of a list of independent variables. The estimation provided by the logistic regression is a probability which predicts the binary outcome variable equals 1, given the baseline state is labeled as 0. In addition to being easily understandable and interpretable, it is one of the most preferred models due to its speed.

Decision Tree Classifier: Decision tree model is the process of separating large data sets into smaller data sets, thanks to the decision stages it applies. In this model, we can say that the more effectively the variables used separate the data set at each stage, the more significant that variable is. The Gini Coefficient produced by these models helps in interpreting how pure the separated group is. Random Forest Classifier, Extra Trees Classifier, XGB Classifier are actually decision tree algorithms and there are slight differences between them. While these algorithms essentially reach the final result by producing more than one tree model, there are differences such as trees working independently from each other or each tree providing information transfer to another tree.

K-Nearest Neighbors (KNN Classifier): This method assigns results based on the results of the observations closest (similar to) it. Each data point receives a prediction value based on the closest cluster to that point. Alternative distance measures such as Euclidean or Mahalanobis distance can be used in the measurement of proximity.

Figure 14 gives the structure of a confusion matrix, which is used to evaluate the results of a predictive model. With the help of confusion matrices, we can evaluate a model within itself and also effectively compare performances of alternative models. The fields specified in the sample matrix below can be explained as follows:

True Positive: True prediction for true actual result

True Negative: False prediction for false actual

False Positive: True prediction for false actual

False Negative: False prediction for true actual

	Positive (1)	Negative (0)
Positive (1)	True Positive	False Positive
Negative (0)	False Negative	True Negative

Figure 14: Confusion Matrix Example

The confusion matrices of all models are shown in Figure 15. The sum of 190.044 data in the whole matrix show how each model predicted true and false results in the test sample.

Logistic Regression

	0	1
0	182,449	142
1	7,270	183

Decision Tree Classifier

	0	1
0	178,327	4,264
1	3,836	3,617

Random Forest Classifier

	0	1
0	181,297	1,294
1	4,068	3,385

Extra Trees Classifier

	0	1
0	181,215	1,376
1	4,039	3,414

XGB Classifier

	0	1
0	181,858	733
1	5,608	1,845

Kneighbors Classifier

	0	1
0	180,913	1,678
1	5,221	2,232

Figure 15: The Comparison of Confusion Matrices for Six Models

Three different metrics are used to interpret the models in the matrix and make them comparable, calculated using the data in the matrix.

Precision: The ratio of true predicted data to total true prediction data or the ratio of false predicted data to total false predicted data.

$$Precision = \frac{\text{True Positive}}{\text{True Poisitive + False Positive}} \quad (1)$$

Recall: The ratio of true predicted data to total true actual data or the ratio of false predicted data to total false actual data.

$$Recall = \frac{\text{True Positive}}{\text{True Poisitive + False Negative}} \quad (2)$$

F1 Score: Harmonic mean of precision and recall results. As in our example used, there is a cost in estimating the paid cheque as not paid, as well as correctly estimating the cheques payable in the cheque payable model. Therefore, the F1 Score is an important metric as it takes into account the cost other than false negative and false positive.

$$F1\ Score = \frac{\text{2 x Precision x Recall}}{\text{Precision + Recall}} \quad (3)$$

Table 6 comparatively shows the evaluation metrics calculated from the confusion matrix of the six different models used. In Table 6, Random Forest Classifier has the best statistics of precision, recall and F1 Score for 0 and 1. Hence, Random Forest Classifier is chosen in the rest of this analysis for the purpose of variable selection.

Table 6: Precision, Recall and F1 Score Comparison of six Models

	Logistic Regression	Decision Tree Classifier	Random Forest Classifier	Extra Trees Classifier	XGB Classifier	Kneighbors Classifier
precision (0)	0.96	0.98	0.98	0.98	0.97	0.97
precision (1)	0.56	0.46	0.72	0.71	0.72	0.57
recall (0)	1.00	0.98	0.99	0.99	1.00	0.99
recall (1)	0.02	0.49	0.45	0.46	0.25	0.30
f1-score (0)	0.98	0.98	0.99	0.99	0.98	0.98
f1-score (1)	0.05	0.47	0.56	0.56	0.37	0.39

The Gini Coefficient used as feature importance item in evaluating the variables to be selected in random forest model. The closer the Gini Coefficient is to 0, the more similar the elements in that group are. In parallel with this, since Gini Coefficient close to 0 have no distinguishing feature. Variables with a Gini Coefficient of 0.35 and above are considered significant and 11 variables marked below red dotted line below Table 7 were selected to be used in the drawer segmentation.

Table 7: Gini Coefficients of Random Forest Classifier

Feature importance items (Gini Coefficient)
BKKREDILIHESAPSAYISI': 0.016
BKKREDINOTUcut_a4': 0.002
BKKURUMSAYISI': 0.012
BKTAKIPBILBULKRMSAY': 0.004
BKTOPLAMTAKIPBAKIYESI': 0.001
CEKENDEKSIcut_a5': 0.003
KARSILIKSIZBTURUTOPLAMCEKSAY': 0.002
KARSILIKSIZBTURUTOPLAMCEKTUT': 0.002
KSD_GERCEKTUZEL_T': 0.002
KSD_IBRAZ_EDILEN_ILK_CEK_GUNcut_a5': 0.005
KSD_LEASING_TOK_TOPLAM': 0.009
KSD_ODENEN_SON_CEK_GUNcut_a5': 0.001
KSD_TK_EN_GUNCEL_LIMIT_TAHSIS_GUNcut_a5': 0.002
KSD_TK_SON_KREDI_KULLANIM_GUNcut_a5': 0.002
KSD_TOPLAM_BANKA_LIMIT_TUM: 0.02
KSD_FAK_TOK_TOPLAM': 0.033
KSD_BANKA_LIMIT_KULLANILIM_ORAN': 0.039
TOPLAMODENMISCEKSAYISI': 0.042
SON3ODENENCEKSAYISI': 0.05
KSD_ILERI_3AY_ODEYECEGICEKADEDI': 0.05
TOPLAMVADELICEKADEDI': 0.052
TOPLAMCEKSAYISI': 0.057
KSD_ILERI_3_AY_CEK_TUTAR_ORANI': 0.063
KSD_ILERI_3AY_ODEYECEGICEKTUT': 0.072
SON3ODENENCEKTUTARI': 0.074
KSD_ILERI_3_AY_CEK_ADET_ORANI': 0.083
TOPLAMODENMISCEKTUTARI': 0.091

6.1.2 K-Means Clustering for Drawers

One of the disadvantages of the K-Means clustering model is the determination of the K (number of clusters) that needs to be determined at the beginning. One of the most common methods in determining K, is the "Elbow Method”. The Elbow Method defines number of clusters by looking at the percentage of the comparison between the numbers of clusters that will form an angle on the curve. In essence, this method determines a center for each cluster it creates, and evaluates the distances of all elements in the cluster from this center with Within-Clusters Sum of Squares (WSS). Equation 4 calculates WSS for each cluster, where greater the number of clusters k, the WSS value will be smaller or vice versa. [26]

$$WSS = \sum_{j=1}^{k}\sum_{i=1}^{n} \parallel x_i^{(j)} - c_j \parallel ^2 \quad (4)$$

n = the number of objects,

x_i = ith element in the cluster,

c_j = the centroid of jth cluster.

In Figure 16, the point at which the WSS effect of an additional cluster is not significant (regarding to the slope of the curve) is determined as the optimum number of clusters which is 5. Drawers segmented into 5 separate clusters with the K-Means method, together with the variables chosen.

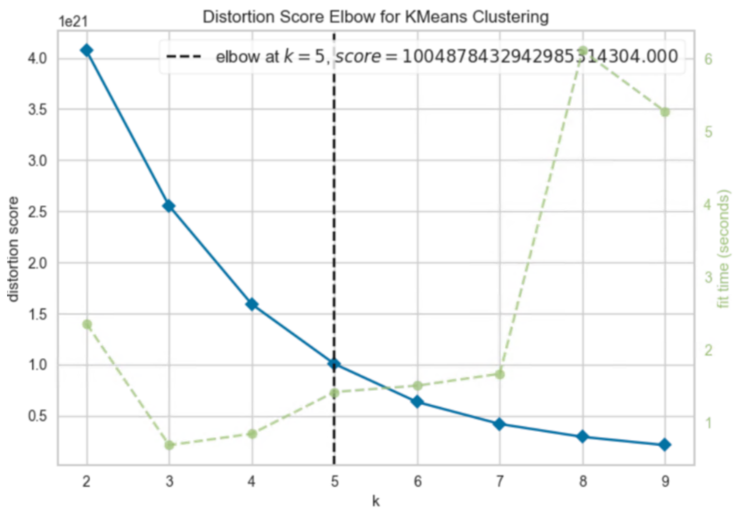

Figure 16: Elbow Method Result for Drawers

The K-Means model is run with 5 clusters, and the results are given in Table 8 and Figure 17. A three-dimensional space representation is needed to visualize a clustering algorithm in which more than two variables are used. Alternatively, transforming the existing variables into two synthetic variables by Pricing Component Analysis (PCA) provides the opportunity to fit this data into a two-axis superimposition. Figure 17 were created with this method in order to visualize the distribution of the explorer clusters in Table 8. In Figure 17, the scale of the graph is arranged to leave out the extreme points and it is aimed to show the separation of the clusters better.

Table 8: K –Means Clustering Results for Drawers

Cluster 1	66,623	40.7%
Cluster 2	38,667	23.6%
Cluster 3	25,790	15.7%
Cluster 4	23,786	14.5%
Cluster 5	8,939	5.5%
Total	163,805	100%

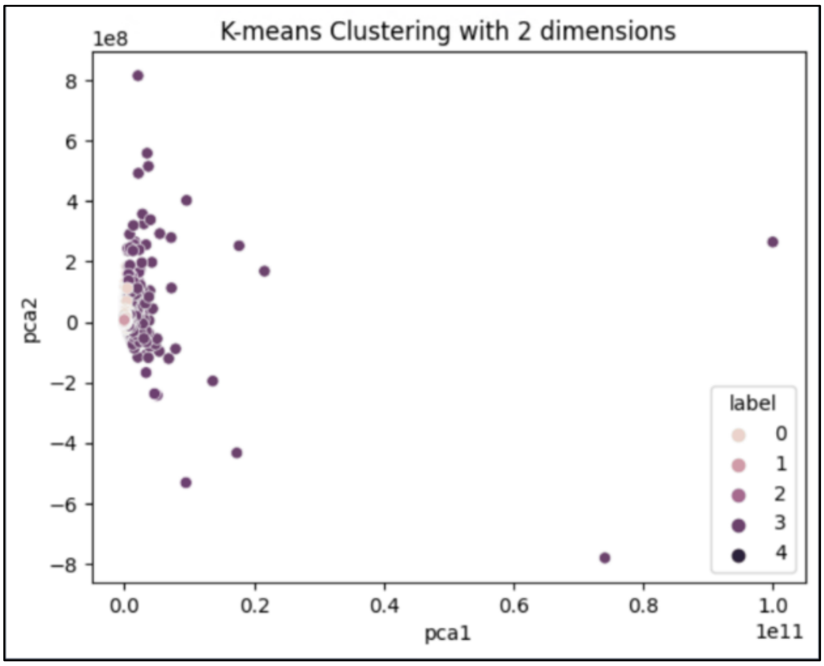

Figure 17: Drawer Clusters Distribution Visual

While we investigate the current analysis in the figure and the table, it is also useful to interpret each cluster. Table 9 is a summary table of the mean statistics of each variable classified by clusters.

Table 9: Variable Means for Interpreting Drawer Clusters

Variable Means	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
KSD_BANKA_LIMIT_KULLANILIM_ORAN	2374%	54%	141%	44%	71%
TOPLAMCEKSAYISI	51	254	517	1,089	4,569
TOPLAMODENMISCEKSAYISI	50.6	253	516	1,088	4,566
SON3ODENENCEKSAYISI	3.8	10	17	32	107
KSD_ILERI_3AY_ODEYECEGICEKADEDI	2.6	6.3	10.8	19.7	56.7
TOPLAMVADELICEKADEDI	3.6	8.7	14.9	26.3	69.7
KSD_ILERI_3_AY_CEK_TUTAR_ORANI	134%	228%	261%	314%	550%
KSD_ILERI_3AY_ODEYECEGICEKTUT (mio TL)	0.1	0.4	0.8	2.0	12.0
SON3ODENENCEKTUTARI (mio TL)	0.2	0.5	1.1	2.9	20.6
KSD_ILERI_3_AY_CEK_ADET_ORANI	61%	72%	71%	70%	63%
TOPLAMODENMISCEKTUTARI (mio TL)	0.7	4.0	10.5	32.3	274
Average Cheque size	0.05	0.06	0.08	0.10	0.21

Interprets about drawers from Table 9 are as follows:

- Cluster 5, the smallest group (in terms of number of drawers), is the group of drawers that use cheques the most in their trade.
- Relatively high percentages of next three months cheques pay ratio is observed in Clusters 2, 3 and 4. It is inferred that they have grown in the last three months with slightly more in cheque usage.
- The average cheque size is significantly different across clusters. According to this variable, Cluster 5 drawers' uses the largest size cheque size among the

As a further analysis, we evaluate the clusters in the model within themselves. We investigate and compare the homogeneity of the clusters in Table 10 using the z-scores calculated by the equation 5.

$$z - score = \frac{\text{cluster standard deviation} - \text{cluster mean}}{\text{cluster standard deviation}} \tag{5}$$

Table 10: Variables Z-Scores of Drawers

Variable Z-Scores	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
TOPLAMCEKSAYISI	0.4	0.0	-0.1	-0.1	0.7
TOPLAMODENMISCEKSAYISI	0.4	0.0	-0.1	-0.1	0.7
SON3ODENENCEKSAYISI	0.1	0.2	0.3	0.2	0.4
KSD_ILERI_3AY_ODEYECEGICEKADEDI	0.2	0.0	-0.1	-0.2	0.4
TOPLAMVADELICEKADEDI	0.2	0.0	-0.1	-0.1	0.4
KSD_ILERI_3AY_ODEYECEGICEKTUT	0.5	0.4	0.5	0.2	0.8
SON3ODENENCEKTUTARI	0.3	0.2	0.1	0.2	0.7
TOPLAMODENMISCEKTUTARI	-0.2	-2.1	-2.9	-1.3	0.8

Table 10 shows that:

- Cluster 5, the smallest group, and Cluster 1, the largest group are two clusters that differ significantly within themselves compared to the other 3 clusters. A possible explanation is that, among the existing drawers, outliers are probably gathered in these two clusters.

Further statistics about cluster characteristics are given in Table 11 and they can be a guide for a more detailed drawer cluster analysis.

Table 11: Drawer Cluster Statistics

TOPLAMCEKSAYISI	Min	Mean	Max	Stdv	Z-score	Count
Cluster 4	7	1,089	19,228	1,001	-0.09	23,786
Cluster 2	1	254	3,875	261	0.03	38,667
Cluster 3	0	517	7,572	467	-0.11	25,790
Cluster 1	0	50.9	1,616	87.03	0.41	66,623
Cluster 5	20	4,569	846,328.00	13,758.68	0.67	8,939

TOPLAMODENMISCEKSAYISI	Min	Mean	Max	Stdv	Z-score	Count
Cluster 4	7	1,088	19,228	1,000	-0.09	23,786
Cluster 2	1	253	3,859	261	0.03	38,667
Cluster 3	0	516	7,572	466	-0.11	25,790
Cluster 1	0	50.6	1,615	86.55	0.42	66,623
Cluster 5	20	4,566	846,328.00	13,756.72	0.67	8,939

SON3ODENENCEKSAYISI	Min	Mean	Max	Stdv	Z-score	Count
Cluster 4	0	32.0	510	26	-0.21	23,786
Cluster 2	0	10.0	132	8.4	-0.18	38,667
Cluster 3	0	17.3	125	14	-0.26	25,790
Cluster 1	0	3.8	72.0	4.4	0.14	66,623
Cluster 5	0	107	7,579.00	174.87	0.39	8,939

KSD_ILERI_3AY_ODEYECEGICEKADE	Min	Mean	Max	Stdv	Z-score	Count
Cluster 4	0	19.7	309	17	-0.15	23,786
Cluster 2	0	6.3	107	6.2	-0.01	38,667
Cluster 3	0	10.8	92	9.5	-0.14	25,790
Cluster 1	0	2.6	70	3.38	0.22	66,623
Cluster 5	0	56.7	3,262	89.51	0.37	8,939

TOPLAMVADELICEKADEDI	Min	Mean	Max	Stdv	Z-score	Count
Cluster 4	0	26	504	24	-0.09	23,786
Cluster 2	0	8.7	199	8.9	0.03	38,667
Cluster 3	0	15	179	14	-0.09	25,790
Cluster 1	0	3.6	81	4.80	0.25	66,623
Cluster 5	0	69.7	4,539	107.41	0.35	8,939

KSD_ILERI_3AY_ODEYECEGICEKTUT	Min	Mean	Max	Stdv	Z-score	Count
Cluster 4	0	2.0	71	2.7	0.24	23,786
Cluster 2	0	0.4	19	0.6	0.39	38,667
Cluster 3	0	0.8	131	1.6	0.48	25,790
Cluster 1	0	0.1	15	0.3	0.53	66,623
Cluster 5	0	12.0	3,586	48.1	0.75	8,939

SON3ODENENCEKTUTARI	Min	Mean	Max	Stdv	Z-score	Count
Cluster 4	0	2.9	59	3.5	0.15	23,786
Cluster 2	0	0.5	5.9	0.6	0.16	38,667
Cluster 3	0	1.1	14.6	1.3	0.13	25,790
Cluster 1	0	0.2	2.0	0.2	0.31	66,623
Cluster 5	0	20.62	5,993	82.20	0.75	8,939

KSD_ILERI_3_AY_CEK_ADET_ORANI	Min	Mean	Max	Stdv	Z-score	Count
Cluster 4	0	0.7	89.0	0.9	0.26	23,786
Cluster 2	0	0.7	100.0	1.1	0.37	38,667
Cluster 3	0	0.7	75.0	0.9	0.23	25,790
Cluster 1	0	0.6	59.0	1.0	0.39	66,623
Cluster 5	0	0.6	55	1.0	0.37	8,939

TOPLAMODENMISCEKTUTARI	Min	Mean	Max	Stdv	Z-score	Count
Cluster 4	14.9	32.3	69	14	-1.33	23,786
Cluster 2	1.1	4.0	6.9	1.3	-2.13	38,667
Cluster 3	0	10.5	16	2.7	-2.93	25,790
Cluster 1	0	0.7	2.2	0.6	-0.15	66,623
Cluster 5	67.3	274	108,058	1,496	0.82	8,939

6.2 Clustering Customers

In this section, customers are clustered with the help of K-Means and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithms. Then, two model compared within themselves to find the more effective clustering model.

6.2.1 Variable Selection for Customer Clustering

One of the independent variable used in clustering customers, is the categorical drawer cluster number obtained from the analysis of the previous section. Other 15 clustering variables, are selected based on domain knowledge from SME factoring and displayed in Figure 18.

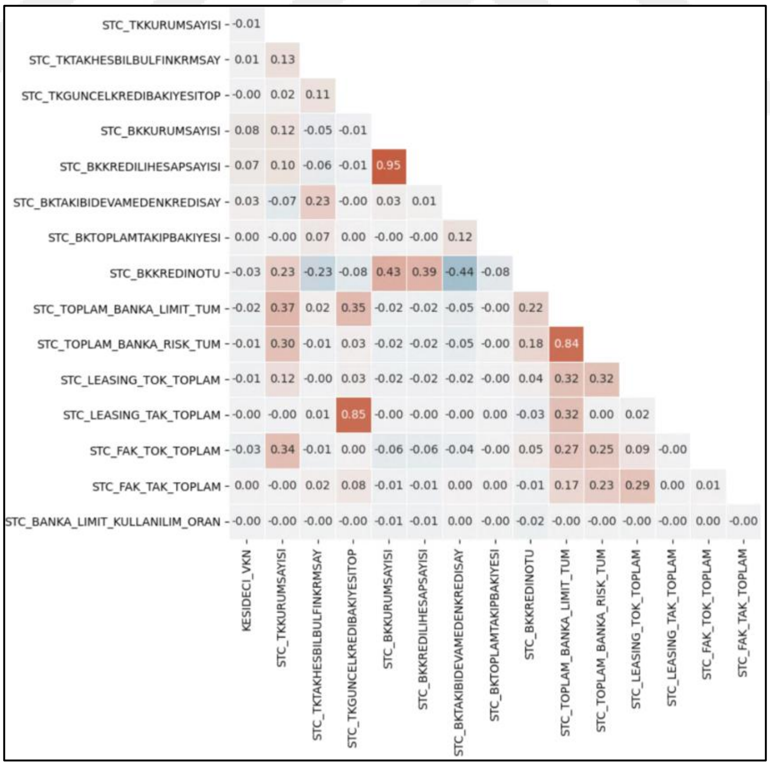

Figure 18: The Correlation Matrix of Customer Variables

6.2.2 K-Means Clustering for Customers

The number of clusters for customers is found to be 4 by the elbow method as seen Figure 19.

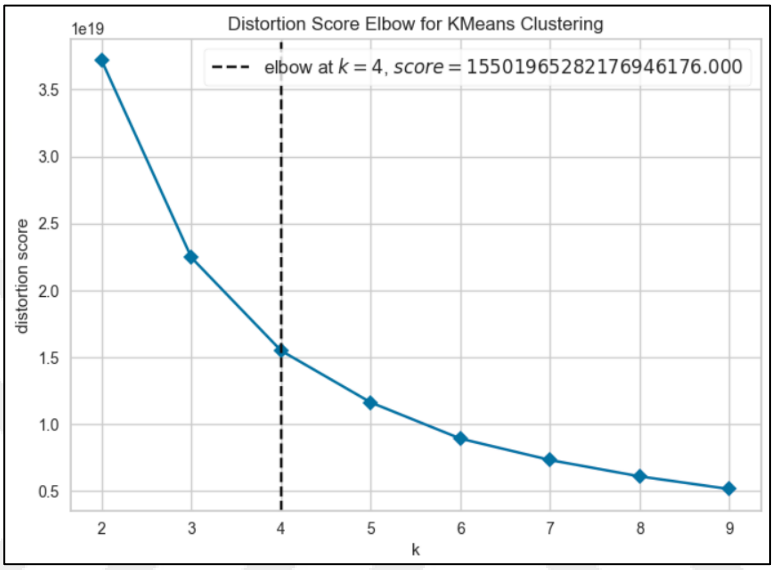

Figure 19: Elbow Method Result for Customers

Table 12 shows the cluster numbers and customer distribution by the K-Means model.

Table 12: K –Means Clustering Results for Customers

	K-means Clustering	
Cluster 1	52,799	56.9%
Cluster 2	25,174	27.1%
Cluster 3	14,116	15.2%
Cluster 4	668	0.7%
Total	92,757	100%

6.2.3 DBSCAN Clustering for Customers

In the DBSCAN algorithm, Eps and MinPts parameters must be provided. MinPts parameter determines the number of minimum observations in a cluster. The Eps parameter, establishes the maximum distance two points can be from one another while still belonging to the same cluster. The default parameters of PHYTON are Eps = 0.5 and MinPts =5. Initially, Eps = 3 and MinPts=100 parameter values are chosen and the clustering model gives the results in the table. The model has divided the sample into 8 clusters and the first two clusters include 99% of all customers.

Table 13: DBSCAN Clustering Results for Customers, with Eps = 3 and MinPts = 100

	DBSCAN Clustering	
Cluster 1	80,129	86.4%
Cluster 2	11,700	12.6%
Cluster 3	250	0.3%
Cluster 4	159	0.2%
Cluster 5	151	0.2%
Cluster 6	143	0.2%
Cluster 7	118	0.1%
Cluster 8	107	0.1%
Total	92,757	100.0%

By trial and error method, the model parameters are changed to Eps = 126 and MinPts = 250, so that there are 4 clusters comparable to the K-Means model. Other parameter values also tried but did not significantly change the allocation into two clusters. It is interpreted that the density-based DBSCAN algorithm creates small clusters for outlier observations away from the center.

The clustering results are shown is Table 14 and it is observed that although four clusters were obtained, the concentration in the two clusters did not change.

Table 14: DBSCAN Clustering Results for Customers, Eps = 126 and MinPts = 250

	DBSCAN Clustering	
Cluster 1	80,402	86.7%
Cluster 2	11,797	12.7%
Cluster 3	306	0.3%
Cluster 4	252	0.3%
Total	92,757	100%

These results imply that the customer profile is concentrated in a certain center, therefore, the density-based DBSCAN algorithm includes many customers in the same two groups, and defines the remaining customers as outlier and noise, and includes them in small clusters.

Table 15: Comparison of Two Clustering Models Distributions for Customers

	K-means Clustering	
Cluster 1	52,799	56.9%
Cluster 2	25,174	27.1%
Cluster 3	14,116	15.2%
Cluster 4	668	0.7%
Total	92,757	100%

	DBSCAN Clustering	
Cluster 1	80,402	86.7%
Cluster 2	11,797	12.7%
Cluster 3	306	0.3%
Cluster 4	252	0.3%
Total	92,757	100%

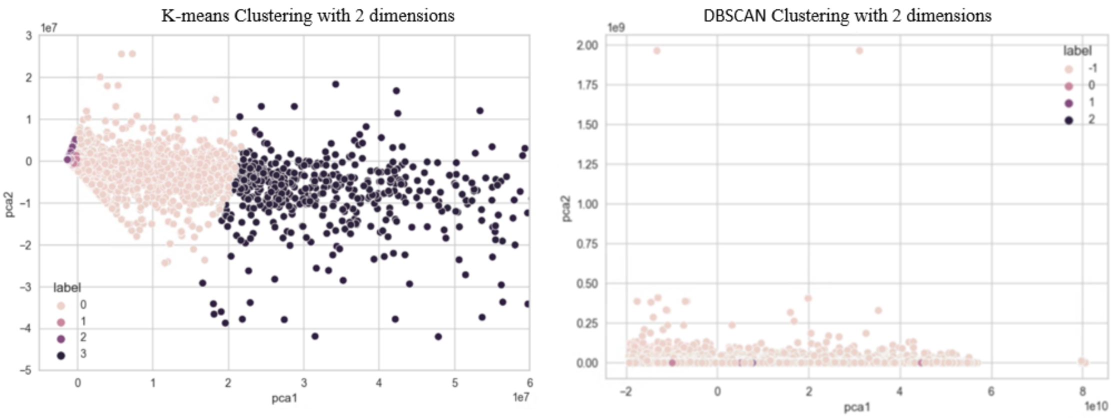

Figure 20: Comparison of two Clustering Models of Customers Visualy

Figure 21 is a hypothetical representation created to better analyze the results of the two models. Accordingly, centroid-based K-Means separated the three groups close to a specific center within itself with the parameter K = 4, and clustered the two outlier groups as a single group. But with the given parameters, DBSCAN has separated the two outlier groups into small groups separately. When DBSCAN parameters are entered as Eps = 3 and MinPts = 100 as in the example above or PHYTON default Eps = 0.5 and MinPts =5, it is normal for the number of outlier groups to increase and to have many small clusters.

Figure 21: Hypothetical representation of of the two algorithms

In order to compare both models, the Mahalanobis distances (which is among the "distance-based" cross-sectional outlier methods for outlier detection) of the observations in the clusters of each model were also calculated. In its simplest terms, Mahalanobis is a measure of the distance between a multidimensional point X and the D distribution [18]. Table 16 shows the Mahalabonis Distance statistics for both models. Weighted average of Mahalanobis means of K-Means is 2.57 which is slightly better than the weighted average of means of DBSCAN of 2.61. This means that the clusters of the K-Means algorithm is more homogenous in total. It has been interpreted that Cluster 3 and Cluster 4, which are two separate small outliers in DBSCAN (with low Mahalanobis scores of 2.86 and 1.51, which is a good indicator), gathered in a single Cluster 4 in the K-Means model, and because these two clusters are far from each other, the combined cluster's Mahalanobis score is high for K-Means (11.64). With different parameters to be selected in the DBSCAN model, various amount of different small clusters will be obtained. It is expected that these small groups will have low Mahalanobis values, that is, they will be efficient in terms of homogeneity, but this will cause the company that performs the segmentation to be faced with a large number of clusters with small customer numbers.

Table 16: Comparison of Two Models' Mahalanobis Distances for Customers

K-Means	Min	Max	Mean	stdv	Median	0.95 Percentile	Count
Cluster 1	0.90	218.12	2.82	2.83	2.30	6.26	52,799
Cluster 2	1.00	20.24	1.85	1.49	1.39	5.06	25,174
Cluster 3	0.58	78.52	2.48	1.79	2.12	5.04	14,116
Cluster 4	2.49	293.01	11.61	24.94	6.10	27.45	668

DBSCAN	Min	Max	Mean	stdv	Median	0.95 Percentile	Count
Cluster 1	0.57	293.01	2.56	3.32	2.00	5.85	80,402
Cluster 2	0.75	151.16	2.98	2.66	2.40	6.12	11,797
Cluster 3	0.99	13.45	2.86	1.80	2.35	6.13	306
Cluster 4	1.01	6.58	1.51	0.90	1.19	3.82	252

Table 17, is a summary table showing K-Means model mean statistics of each customer cluster.

Table 17: Variable Means for Interpreting Customer Clusters

Variable Means	Cluster 1	Cluster 2	Cluster 3	Cluster 4
STC_BKKREDILIHESAPSAYISI	2	4	3	0
STC_BKTOPLAMTAKIPBAKIYESI	779	496	91	0
STC_BKKREDINOTU	439	831	524	84
STC_TOPLAM_BANKA_LIMIT_TUM	40	426	2,908	37,205
STC_TOPLAM_BANKA_RISK_TUM	26	394	3,356	52,210
STC_LEASING_TOK_TOPLAM (k TL)	2.6	10	44	1,031
STC_LEASING_TAK_TOPLAM (k TL)	0.1	0.1	0.7	32
STC_FAK_TOK_TOPLAM (k TL)	110	67	433	4,265
STC_FAK_TAK_TOPLAM (k TL)	0.2	0.2	0.1	29
STC_SON_KANUNI_TAKIP_GUN	284	135	41	0.0
STC_TK_EN_GUNCEL_LIMIT_TAHSIS_GUN	364	218	92	45
STC_BK_SON_KREDI_KULLANIM_GUN	373	398	226	36
STC_TK_SON_KREDI_KULLANIM_GUN	153	82	23	14
CALISTIGI_KURUM_TOPLAM	2.5	6.1	8.2	13.2
TAKIP_KURUM_TOPLAM	0.6	0.4	0.2	0.2

Interprets about customers from Table 17 are as follows:

- Customers in Cluster 4 have significantly more credits and limits in banking, leasing and factoring intuitions compared to other clusters. This suggests that the customers in Cluster 4 are financially strong companies.
- The same cluster has the lowest mean data as Individual Credit Score (BKN). The fact that the Individual Credit Score (BKN) will be high for companies with high credibility contradicts the previous comment. However, at the same time, the Individual Credit Score (BKN) is calculated only for individual firms, and commercial firms do not have Individual Credit Score (BKN).

Table 18 and Figure 22 shows the distribution of all customers in the dataset according to company type. Although the individual and commercial distribution is almost equal in total (54% vs 46%), 95% of the companies in Cluster 5 are commercial companies.

Table 18: Customers Breakdown According to Company Type

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	**Total**
Indivdual	27,733	16,849	5,383	36	50,001
Commercial	25,066	8,325	8,733	632	42,756
Total	52,799	25,174	14,116	668	92,757

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
Indivdual	53%	67%	38%	5%	54%
Commercial	47%	33%	62%	95%	46%
Total	100%	100%	100%	100%	100%

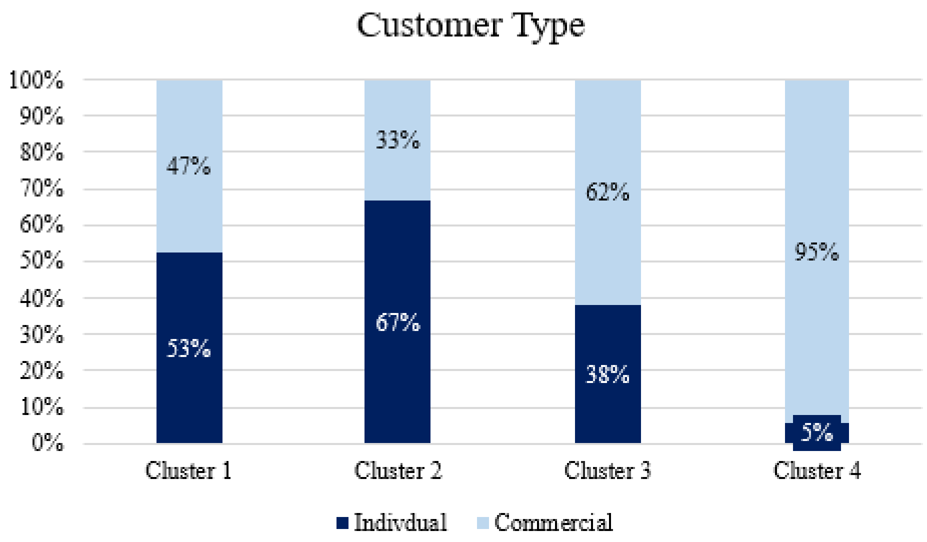

Figure 22: Distribution of Customers According to Company Type

Table 19 shows the number of customers in the cluster customers separately as individual and commercial. Clusters are interpreted differently according to the company type in the following Table 20.

Table 19: Cluster Breakdown to Company Type

	Individual	Commercial	**Total**
Cluster 1	27,733	25,066	52,799
Cluster 2	16,849	8,325	25,174
Cluster 3	5,383	8,733	14,116
Cluster 4	36	632	668
Total	50,001	42,756	92,757

Table 20: Variable Means for Interpreting Customer Clusters on Company Type Basis

Individual Companies

Variable Means for Individual Companies	Cluster 1	Cluster 2	Cluster 3	Cluster 4
STC_BKKREDILIHESAPSAYISI	3	6	7	8
STC_BKTOPLAMTAKIPBAKIYESI	1,484	741	239	0
STC_BKKREDINOTU	836	1,242	1,375	1,549
STC_TOPLAM_BANKA_LIMIT_TUM	59	397	2,262	22,369
STC_TOPLAM_BANKA_RISK_TUM	34	376	2,683	30,522
STC_LEASING_TOK_TOPLAM (k TL)	1.3	5	17	89
STC_LEASING_TAK_TOPLAM (k TL)	0.1	0.0	0.0	0
STC_FAK_TOK_TOPLAM (k TL)	55	44	197	750
STC_FAK_TAK_TOPLAM (k TL)	0.2	0.0	0.0	0
STC_TK_EN_GUNCEL_LIMIT_TAHSIS_GUN	452	202	75	40
CALISTIGI_KURUM_TOPLAM	3.5	7.3	11.0	16.0
TAKIP_KURUM_TOPLAM	1.2	0.4	0.2	0.1

Commercial Companies

Variable Means for Commercial Companies	Cluster 1	Cluster 2	Cluster 3	Cluster 4
STC_BKKREDILIHESAPSAYISI	0	0	0	0
STC_BKTOPLAMTAKIPBAKIYESI	0	0	0	0
STC_BKKREDINOTU	0	0	0	0
STC_TOPLAM_BANKA_LIMIT_TUM	20	484	3,306	38,050
STC_TOPLAM_BANKA_RISK_TUM	16	431	3,771	53,445
STC_LEASING_TOK_TOPLAM (k TL)	4.2	20	61	1,084
STC_LEASING_TAK_TOPLAM (k TL)	0.1	0.3	1.1	33
STC_FAK_TOK_TOPLAM (k TL)	171	112	579	4,466
STC_FAK_TAK_TOPLAM (k TL)	0.2	0.5	0.1	31
STC_TK_EN_GUNCEL_LIMIT_TAHSIS_GU	267	252	103	46
CALISTIGI_KURUM_TOPLAM	1.4	3.6	6.5	13.0
TAKIP_KURUM_TOPLAM	0.1	0.2	0.1	0.2

- For both types of companies, it is observed that Cluster 1 has the least relationship with banks. Individual customers' works with more financial credit institutions.
- On the contrary, Bank limits and risks of commercial customers are higher in Cluster 2, 3 and 4 than individual customers.
- It is interpreted that individual customers meet their financial needs by working with small limits but more institutions due to weaker financial strength than commercial customers.
- The Individual Credit Score individual customers increases significantly from Cluster 1 to Cluster 4. This indicates that customers in Cluster 4 are in a stronger financial position.

The drawer cluster variable that was used was in customer clustering not added to the analysis in Table 20. Instead, in Figure 23, the drawer clusters in each cluster were analyzed. It is observed that the drawer distribution among customer clusters does not change and remain in line parallel with the drawer cluster distribution handled in the drawer cluster.

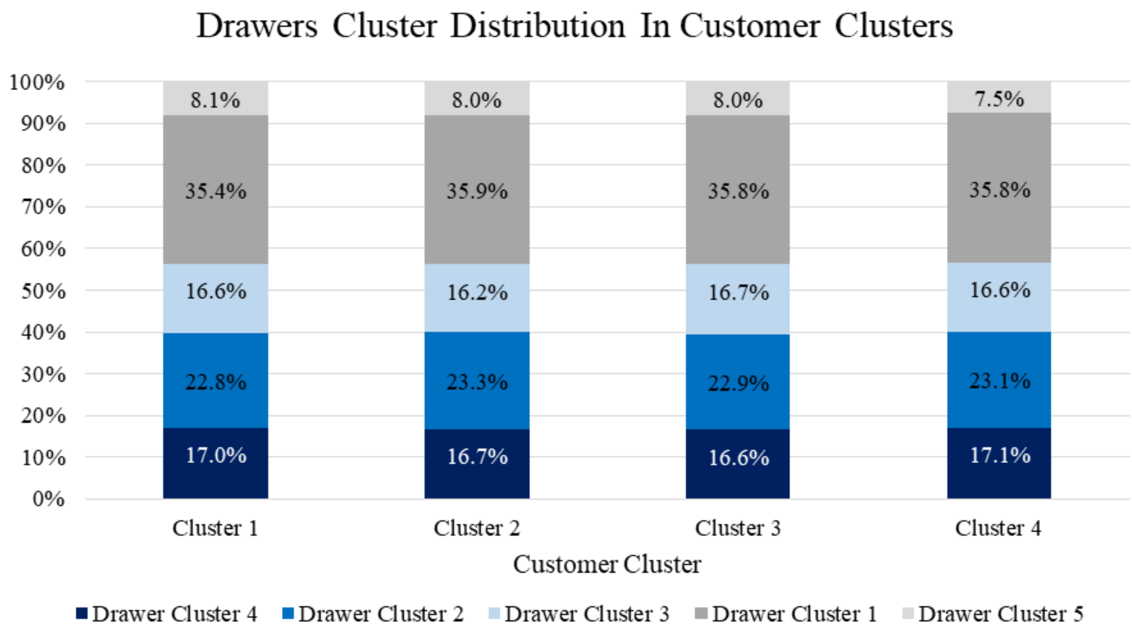

Figure 23: Drawers Cluster Distribution in Customer Clusters

Customers are aware of the quality of the cheque they bring (from the drawer) during the factoring process. Even though a customer with low credibility like Cluster 1, brings a cheque of a very strong drawer (Cluster 5 in our model) knows the quality of this cheque and for example, a price strategy should be made accordingly.

Further statistics about cluster characteristics are given in Table 21 (for individual companies) and Table 22 (for commercial companies). They can be a guide for a more detailed customer cluster analysis.

Table 21: Individual Customer Cluster Statistics

BKKREDILIHESAPSAYISI	Min	Mean	Max	Stdv	Z-score	Count
Cluster 3	0.00	6.92	25	3.90	-0.78	5,383
Cluster 2	0.00	5.81	30	3.60	-0.61	16,849
Cluster 1	0.00	2.86	22	2.81	-0.02	27,733
Cluster 4	3.00	7.61	16	3.18	-1.39	36

BKTOPLAMTAKIPBAKIYESI	Min	Mean	Max	Stdv	Z-score	Count
Cluster 3	0.00	239.27	128,601	3,746	0.94	5,383
Cluster 2	0.00	740.69	366,562	8,636	0.91	16,849
Cluster 1	0.00	1,483.78	4,717,206	30,034	0.95	27,733
Cluster 4	0.00	0.00	0	0	0.00	36

BKKREDINOTU	Min	Mean	Max	Stdv	Z-score	Count
Cluster 3	0.00	1,374.87	1,900	344	-2.99	5,383
Cluster 2	0.00	1,241.87	1,900	425	-1.92	16,849
Cluster 1	0.00	836.06	1,900	611	-0.37	27,733
Cluster 4	1,082.00	1,549.47	1,900	210	-6.37	36

TOPLAM_BANKA_LIMIT_TUM	Min	Mean	Max	Stdv	Z-score	Count
Cluster 3	0.00	2,262,011	20,462,090	1,952,063	-0.16	5,383
Cluster 2	0.00	396,609	2,021,968	268,356	-0.48	16,849
Cluster 1	0.00	58,605	619,533	60,457	0.03	27,733
Cluster 4	6,463,289	22,369,259	52,487,500	9,659,058	-1.32	36

TOPLAM_BANKA_RISK_TUM	Min	Mean	Max	Stdv	Z-score	Count
Cluster 3	0.00	2,683,217	20,047,229	2,440,419	-0.10	5,383
Cluster 2	0.00	375,837	1,612,474	273,779	-0.37	16,849
Cluster 1	0.00	34,254	457,026	51,223	0.33	27,733
Cluster 4	9,929,488	30,522,212	68,368,012	14,213,429	-1.15	36

LEASING_TOK_TOPLAM	Min	Mean	Max	Stdv	Z-score	Count
Cluster 3	0.00	16,578	5,841,054	159,910	0.90	5,383
Cluster 2	0.00	4,759	6,954,512	74,751	0.94	16,849
Cluster 1	0.00	1,271	1,616,746	27,383	0.95	27,733
Cluster 4	0.00	89,367	1,685,808	348,080	0.74	36

LEASING_TAK_TOPLAM	Min	Mean	Max	Stdv	Z-score	Count
Cluster 3	0.0	0	0	0	0.00	5,383
Cluster 2	0.0	39	282,408	2,933	0.99	16,849
Cluster 1	0.0	79	288,565	3,568	0.98	27,733
Cluster 4	0.0	0	0	0	0.00	36

FAK_TOK_TOPLAM	Min	Mean	Max	Stdv	Z-score	Count
Cluster 3	0.0	196,788	47,308,684	1,070,219	0.82	5,383
Cluster 2	0.0	44,099	3,149,999	140,913	0.69	16,849
Cluster 1	0.0	55,354	6,998,305	220,234	0.75	27,733
Cluster 4	0.0	750,136	7,709,999	1,925,464	0.61	36

FAK_TAK_TOPLAM	Min	Mean	Max	Stdv	Z-score	Count
Cluster 3	0.0	3	15,000	204	0.99	5,383
Cluster 2	0.0	24	133,910	1,242	0.98	16,849
Cluster 1	0.0	197	1,330,212	10,719	0.98	27,733
Cluster 4	0.0	0	0	0	0.00	36

BANKA_LIMIT_KULLANILIM_(	Min	Mean	Max	Stdv	Z-score	Count
Cluster 3	0.0	1.4	71	2	0.10	5,383
Cluster 2	0.0	131.7	521,984	7,154	0.98	16,849
Cluster 1	0.0	63.5	334,138	3,918	0.98	27,733
Cluster 4	0.3	1.5	4	1	-1.13	36

SON_KANUNI_TAKIP_GUN	Min	Mean	Max	Stdv	Z-score	Count
Cluster 3	0.0	108	8,466	494	0.78	5,383
Cluster 2	0.0	202	6,096	618	0.67	16,849
Cluster 1	0.0	540	8,587	916	0.41	27,733
Cluster 4	0.0	0	0	0	0.00	36

TK_EN_GUNCEL_LIMIT_TAHSI	Min	Mean	Max	Stdv	Z-score	Count
Cluster 3	0.0	75	3,208	120	0.38	5,383
Cluster 2	0.0	202	6,265	324	0.38	16,849
Cluster 1	0.0	452	8,324	786	0.42	27,733
Cluster 4	1.0	40	175	43	0.07	36

BK_SON_KREDI_KULLANIM_G	Min	Mean	Max	Stdv	Z-score	Count
Cluster 3	0	591.51	9,421	943	0.37	5,383
Cluster 2	0	594.55	8,769	943	0.37	16,849
Cluster 1	0	710.38	10,753	1,160	0.39	27,733
Cluster 4	5	664.58	4,253	925	0.28	36

TK_SON_KREDI_KULLANIM_G	Min	Mean	Max	Stdv	Z-score	Count
Cluster 3	0	14.24	1,540	58	0.75	5,383
Cluster 2	0	61.47	5,821	223	0.72	16,849
Cluster 1	0	146.34	8,092	511	0.71	27,733
Cluster 4	1	6.61	131	22	0.70	36

CALISTIGI_KURUM_TOPLAM	Min	Mean	Max	Stdv	Z-score	Count
Cluster 3	1	10.95	36	4.14	-1.64	5,383
Cluster 2	1	7.29	30	2.99	-1.43	16,849
Cluster 1	0	3.49	22	2.39	-0.46	27,733
Cluster 4	7	16.03	24	4.60	-2.48	36

TAKIP_KURUM_TOPLAM	Min	Mean	Max	Stdv	Z-score	Count
Cluster 2	0	0.19	20	0.83	0.77	5,383
Cluster 3	0	0.44	30	1.37	0.68	16,849
Cluster 4	0	1.16	19	1.92	0.40	27,733
Cluster 1	0	0.08	1	0.28	0.70	36

Table 22: Commercial Customer Cluster Statistics

TOPLAM_BANKA_LIMIT_TUM	Min	Mean	Max	Stdv	Z-score	Count
Cluster 3	0.00	3,305,986	31,000,000	3,039,643	-0.09	8,733
Cluster 2	0.00	484,466	2,070,729	323,780	-0.50	8,325
Cluster 4	0.00	19,892	686,289	53,266	0.63	25,066
Cluster 1	389,200	38,050,063	819,148,340	60,865,216	0.37	632

TOPLAM_BANKA_RISK_TUM	Min	Mean	Max	Stdv	Z-score	Count
Cluster 3	0.00	3,771,093	22,730,974	3,503,462	-0.08	8,733
Cluster 2	0.00	431,145	1,723,190	323,820	-0.33	8,325
Cluster 4	0.00	16,415	538,632	44,472	0.63	25,066
Cluster 1	194,434	53,445,402	1,794,554,845	127,945,904	0.58	632

LEASING_TOK_TOPLAM	Min	Mean	Max	Stdv	Z-score	Count
Cluster 3	0.00	60,806	21,509,909	509,389	0.88	8,733
Cluster 2	0.00	19,628	35,722,482	428,208	0.95	8,325
Cluster 4	0.00	4,162	7,263,123	90,816	0.95	25,066
Cluster 1	0.00	1,084,246	108,550,345	6,595,122	0.84	632

LEASING_TAK_TOPLAM	Min	Mean	Max	Stdv	Z-score	Count
Cluster 3	0.0	1,103	4,449,307	60,646	0.98	8,733
Cluster 2	0.0	350	1,171,397	15,237	0.98	8,325
Cluster 4	0.0	76	1,261,566	8,135	0.99	25,066
Cluster 1	0.0	33,372	20,741,798	825,125	0.96	632

FAK_TOK_TOPLAM	Min	Mean	Max	Stdv	Z-score	Count
Cluster 3	0.0	578,515	62,649,198	2,286,402	0.75	8,733
Cluster 2	0.0	112,309	3,203,308	277,463	0.60	8,325
Cluster 4	0.0	171,047	11,845,726	586,810	0.71	25,066
Cluster 1	0.0	4,465,629	686,168,694	29,737,453	0.85	632

FAK_TAK_TOPLAM	Min	Mean	Max	Stdv	Z-score	Count
Cluster 3	0.0	83	397,200	4,544	0.98	8,733
Cluster 2	0.0	464	2,334,100	26,793	0.98	8,325
Cluster 4	0.0	229	583,005	6,110	0.96	25,066
Cluster 1	0.0	31,052	14,108,248	602,252	0.95	632

BANKA_LIMIT_KULLANILIM_(	Min	Mean	Max	Stdv	Z-score	Count
Cluster 3	0.0	1.5	404	5	0.72	8,733
Cluster 2	0.0	44.7	358,300	3,927	0.99	8,325
Cluster 4	0.0	47.6	297,500	2,995	0.98	25,066
Cluster 1	0.0	1.8	124	5	0.68	632

SON_KANUNI_TAKIP_GUN	Min	Mean	Max	Stdv	Z-score	Count
Cluster 3	0.0	0	0	0	0.00	8,733
Cluster 2	0.0	0	0	0	0.00	8,325
Cluster 4	0.0	0	0	0	0.00	25,066
Cluster 1	0.0	0	0	0	0.00	632

TK_EN_GUNCEL_LIMIT_TAHSI	Min	Mean	Max	Stdv	Z-score	Count
Cluster 3	0.0	103	4,288	196	0.47	8,733
Cluster 2	0.0	252	6,235	445	0.43	8,325
Cluster 4	0.0	267	6,977	695	0.62	25,066
Cluster 1	0.0	46	1,436	111	0.59	632

TK_SON_KREDI_KULLANIM_GI	Min	Mean	Max	Stdv	Z-score	Count
Cluster 3	0	29.02	3,776	132	0.78	8,733
Cluster 2	0	123.77	6,572	393	0.68	8,325
Cluster 4	0	161.11	8,920	581	0.72	25,066
Cluster 1	1	14.84	1,154	82	0.82	632

CALISTIGI_KURUM_TOPLAM	Min	Mean	Max	Stdv	Z-score	Count
Cluster 3	0	6.51	30	3.67	-0.77	8,733
Cluster 2	1	3.58	18	2.24	-0.60	8,325
Cluster 4	0	1.37	27	1.74	0.21	25,066
Cluster 1	1	13.00	43	5.98	-1.18	632

TAKIP_KURUM_TOPLAM	Min	Mean	Max	Stdv	Z-score	Count
Cluster 3	0	0.13	13	0.71	0.82	8,733
Cluster 2	0	0.16	9	0.63	0.75	8,325
Cluster 4	0	0.06	7	0.30	0.80	25,066
Cluster 1	0	0.19	13	1.12	0.83	632

CHAPTER VII

CONCLUSION

This study is a first in the literature, clustering Turkish SME factoring customers. Since the factoring company whose dataset is used represent 39.1% of the total companies using for financing, an important data for the sector is included in the study.

This thesis has been applied as two-stage clustering methods for the first time in the area of factoring sector. In order to produce meaningful or interpretable results, first the drawers were clustered and then these clusters were used as a variable in customer segmentation.

When the clusters generated by the K-Means and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) models are compared, it is observed that the clusters of the K-Means algorithm are more balanced in terms of distribution of the number of customers across clusters. On the other hand, the clusters generated by the DBSCAN model are more homogeneous, especially for outlier clusters. Homogeneity means an effective segmentation, which is there are few outliers within a cluster as defined by the Mahalanobis distance of each observation. The DBSCAN model can generate several homogeneous clusters containing a small number of customers if model parameters are appropriately calibrated. On the contrary, if few number of clusters are preferred, the homogeneity of the DBSCAN algorithm gets worse and the K-Means model gives more balanced clustering results. On the other hand, it has been observed that the drawer clusters within the customer clusters created did not differentiate from the total drawer cluster distribution.

BIBLIOGRAPHY

[1] Wikipedia, "Customer," [Online]. Available: wikipedia.org. https://en.wikipedia.org/wiki/Customer (accessed Nov 01, 2022).

[2] T. Hastie, R. Tibshirani and J. Friedman, The Elements Of Statistical Learning, 2nd ed. 2001, pp. 501-502.

[3] F. T. Lima, J. R. Kós and R. C. Paraizo, Handbook of Research on Visual Computing and Emerging Geometrical Design Tools, 2016.

[4] Statista, "Number of amazon prime subscription households usa," [Online]. Available: https://www.statista.com/statistics/504687/number-of-amazon-prime-subscription-households-usa/ (accessed May 02, 2021).

[5] Statista, "Number of monthly active facebook users worldwide," [Online]. Available: https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/#:~:text=With%20over%202.7%20billion%20monthly,network%20ever%20to%20do%20so (accessed Nov 01, 2022).

[6] Coca-Cola, "Cocacola Company Overview," [Online]. Available: https://d1io3yog0oux5.cloudfront.net/_c395ef695ff6760498b6445d7afd64f4/cocacolacompany/db/706/6349/pdf/IR+Overview+-+Updated+for+3Q+_10.28.pdf (accessed May 02, 2021).

[7] Statista, "Alibaba cumulative active online buyers taobao tmall," [Online]. Available: https://www.statista.com/statistics/226927/alibaba-cumulative-active-online-buyers-taobao-tmall/#:~:text=The%20largest%20Chinese%20e%2Dcommerce,the%20fisrt%20quarter%20of%202020 (accessed Nov 01, 2022).

[8] Statista, "Banco santander customer numbers globally," [Online]. Available: https://www.statista.com/statistics/416413/banco-santander-customer-numbers-globally/ (accessed Nov 01, 2022).

[9] T. Mitchell, Machine Learning, McGraw-Hill, 1997.

[10] Wikipedia, "Cluster analysis," [Online]. Available: https://en.wikipedia.org/wiki/Cluster_analysis (accessed Nov 01, 2022).

[11] U. Kutbay, Recent Applications in Data Clustering, 2018.

[12] Wikipedia, "Fuzzy clustering," [Online]. Available: https://en.wikipedia.org/wiki/Fuzzy_clustering (accessed Nov 01, 2022).

[13] A. Fahim, A. Salem, F. Torkey and M. Ramadan, "Density Clustering Based on Radius of Data," International Journal of Applied Mathematics and Computer Sciences, vol. 3, pp. 80-85.

[14] B. Simhachalam and G. Ganesan, "Performance comparison of fuzzy and non-fuzzy classification methods," Egyptian Informatics Journal, 2015.

[15] J. Hartigan and M. Wong, "A K-Means Clustering Algorithm," Journal of the Royal Statistical Society. Series C (Applied Statistics), vol. 28, pp. 100-108, 1979.

[16] M. Celebi, H. A. Kingravi and P. A. Vela, "A comparative study of efficient initialization methods for the k-means clustering algorithm," Expert Systems with Applications, pp. 200-210, 2013.

[17] K. Khan, S. U. Rehman, K. Aziz, S. Fong and S. A. Vishwa, "DBSCAN Past, Present and Future," Rehman, S. U., Asghar, S., Fong, S., & The Fifth International Conference on the Applications of Digital Information and Web Technologies, 2014, pp. 232-238.

[18] L. Güntay and M. Aktuna, "Finansal Kurumlarda Senaryo Bazlı Aykırı Gözlem Tespiti: Türk Faktoring Sektörü Üzerine Bir Çalışma," BDDK Bankacılık ve Finansal Piyasalar Dergisi, pp 83-113, 2021.

[19] A. Çalış, A. Boyacı and B. Kasım, "Data mining application in banking sector with clustering and classification methods," International Conference on Industrial Engineering and Operations Management, Dubai, 2015.

[20] E. Nikumanesh and A. Albadvi, "Customer's life-time value using the RFM model in the banking industry a case study," International Journal of Electronic Customer Relationship Management, vol. 8, pp. 15-30, 2014.

[21] M. Khajvand and M. J. Tarokh, "Estimating customer future value of different customer segments based on adapted RFM model in retail banking context," Procedia Computer Science 3, pp. 1327–1332, 2011.

[22] S. Peker, A. Koçyiğit and P. E. Eren, "LRFMP model for customer segmentation in the grocery retail industry: a case study," Emerald Insight, pp. 544-559, 2017.

[23] A. Sheikh, T. Ghanbarpour and D. Gholamiangonabadi, "A Preliminary Study of Fintech Industry: A Two-Stage Clustering Analysis for Customer Segmentation in the B2B Setting," Journal Of Business-To-Business Marketing, vol. 26, pp. 197-207, 2019.

[24] Riskmerkezi, "Istatistikler," [Online]. Available: riskmerkezi.org https://www.riskmerkezi.org/Content/Upload/istatistikiraporlar/ekler/3519/Ban

kalara_Ibraz_Edilen_ve_Karsiliksiz_Islemi_Yapilan_Cek_Bilgileri_-_Ekim_2022.pdf (accessed Nov 01, 2022).

[25] BDDK, "https://www.bddk.org.tr/BultenAylikBdmk/tr/Gosterim/Faktoring," Bulten Aylik Bdmk, [Online]. Available: bddk.org. https://www.bddk.org.tr/BultenAylikBdmk/tr/Gosterim/Faktoring (accessed Nov 01, 2022).

[26] R. Gustriansyah, N. Suhandi and F. Antony, "Clustering optimization in RFM analysis based on k-means," Indonesian Journal of Electrical Engineering and Computer Science, vol. 18, pp. 470-477, 2020.

VITA

After graduating from Hacettepe University, Department of Business Administration, he worked as a financial controller in various companies and joined to the factoring company where he is still working in 2012 as a budget and planning specialist. In the company where he has been working since its establishment, he expanded his field of duty over time and carried out the duties of senior analytics specialist, treasury manager and financial control and reporting manager under the finance department. He served as the analytics manager between September 2019 and October 2022 and is currently holding the strategic planning manager position.

While working in the analytics department, he took part in risk-oriented in-house scorecard studies developed for the SME segment in factoring segment. On the other hand, he took part in the risk analysis studies of the company and also worked on the development of the pricing strategy including interest rate and maturity risk. Within the scope of the experience gained in the field of risk and finance, the works within the 5-year business plan of the company are under his responsibility.