MEF UNIVERSITY

ANOMALY DETECTION IN FACTORING TRANSACTIONS

Capstone Project

Mustafa Kerim Acar

Advisor: Prof. Dr. Özgür Özlük

İSTANBUL, 2019

MEF UNIVERSITY

Name of the project: Anomaly Detection in Factoring Transactions
Name/Last Name of the Student: Mustafa Kerim Acar
Date of Thesis Defense: 19/12/2019

I hereby state that the graduation project prepared by Your Name (Title Format) has been completed under my supervision. I accept this work as a "Graduation Project".

20/12/2019
Prof. Dr. Özgür Özlük

I hereby state that I have examined this graduation project by Your Name (Title Format) which is accepted by his supervisor. This work is acceptable as a graduation project and the student is eligible to take the graduation project examination.

20/12/2019

Prof. Dr. Özgür Özlük
Director
of
Big Data Analytics Program

We hereby state that we have held the graduation examination of __________ and agree that the student has satisfied all requirements.

THE EXAMINATION COMMITTEE

Committee Member	Signature
1. Your Advisor's Name	………………………..
2. …………………………..	………………………..

Academic Honesty Pledge

I promise not to collaborate with anyone, not to seek or accept any outside help, and not to give any help to others.

I understand that all resources in print or on the web must be explicitly cited.

In keeping with MEF University's ideals, I pledge that this work is my own and that I have neither given nor received inappropriate assistance in preparing it.

Kerim Acar 19.12.2019

Name Date Signature

EXECUTIVE SUMMARY

ANOMALY DETECTION IN FACTORING TRANSACTIONS

Mustafa Kerim Acar

Advisor: Prof. Dr. Özgür Özlük

DECEMBER, 2019, 26 Pages

In this study, anomaly detection model performed by using 3 different approaches on factoring dataset which consist of 890.118 observations and 122 features provided by one of the leading factoring companies. In this respect, it is aimed to determine whether the checks brought to the factoring company are fraud or not. In order to enable the company to take precaution against such crimes in a fast and rapid manner, statistics, density and machine learning based algorithms are used and these algorithms are evaluated according to the highest accuracy score and working speed results. Mahalanobis distance, which is a rule-based approach, is strengthened with minimum covariance determinant and gives the fastest result compared to other approaches. In line with the use of this algorithm, the accuracy score threshold determined by the company is passed. The highest accuracy score is achieved by using the Isolation Forest algorithm, however this algorithm is very slow compared to the others. The density-based anomaly detection algorithm scores below the threshold specified by the company and it operates relatively slow.

Key Words: Anomaly Detection, Outlier Detection, Mahalanobis, Local Outlier Factor, Isolation Forest, Minimum Covariance Determinant

TABLE OF CONTENTS

.

1. INTRODUCTION

Discovery of suspicious and abnormal trends and instances in data science has started to receive more and more attention with both recent developments in machine learning and immense stream of data. Currently, amount of stored data steadily grows as a result of huge cloud systems and almost 90% of the stored data is unstructured (Reinsel, Gantz & Rydning, 2018). As a consequence, knowledge discovery and anomaly detection became a more challenging task each passing day.

Detecting anomalies accurately is a very crucial phenomena in such fields as financial fraud detection, public and medical health, image processing and sensor networks. Anomalies are generally perceived as a clue for abnormality which raises attention for further investigations. For example, anomalous instances in financial transactions may be a sign of money laundering or credit card theft (Srivastava et al., 2008). In health area, unusual level of Alanine Aminotransferase in blood may indicate a person has liver disease (Schindhelm et al., 2006).

In this study, three anomaly detection approaches are tested, and different methods are combined to create a more convenient and robust anomaly detection model. This study aims to improve the quality of anomaly detection process of financial institutions, particularly in factoring. Main motivation of this study is the lack of academic studies in the literature that especially focus on developing anomaly detection model for factoring companies. Dataset used in this study is provided by XXX Factoring and the company aims to build an anomaly detection infrastructure for their daily internal audit process and a fast responding anomaly detection model. For the rest of this chapter, anomalies are introduced in detail and related studies in anomaly detection has been presented. Information about XXX Factoring and dataset will be presented in later chapters.

Anomaly detection is the task of finding unusual and abnormal observations, or simply outliers. Outlier is an entry or observation that differs from other instances which could be generated by a different source (Hawkins, 1980). In other words, anomalies or outliers are observations in data that do not fit normal behavior exhibited by other observations (Chandola, Banerjee, & Kumar, 2009).

In data science, there are three types of anomalies (Chandola et al., 2009);

- *Point Anomalies* are single observations in data which labeled as anomaly because the rest of the observations follows same pattern differing from that observation. Point anomalies are very common and simple, most of the studies has been made for point anomalies. If anomalous case is detected by using a single feature, let's say amount spent on credit card, it can be classified as point anomaly.
- *Contextual Anomalies* are different data points in defined pattern. These anomalies could be perceived as anomaly only for given context. If person spends much more than his/her average amount for a week, it would be an example of contextual anomaly.
- *Collective Anomalies* are collection of data points detected perceived as anomaly. These data points are not seen as anomalous cases individually. Cyber-attacks to steal private data from internet pages are perceived as collective anomalies.

Another definition about outliers from clustering perspective, outlier is an observation that is not included in any cluster of given data (Chandola et al., 2009). Clustering algorithms could be useful in outlier detection but very few of them specifically focus on outliers.

1.1. Literature Survey

Literature has extensively analyzed methods of anomaly detection for various cases. In this study, investigated anomaly detection techniques can be classified in three categories; statistics-based approaches, density-based approaches and machine learning approaches.

Early anomaly detection methods were predominantly constituted by statistical approaches. Statistical anomaly detection models are mainly based on statistical distributions of features in dataset. Particularly, most of the statistical techniques are only applicable to normally distributed data. Observations in normally distributed data follow a similar pattern, hence outliers can be defined as observations that follow a different pattern from the rest of other observations. One of the earliest techniques for detecting outliers is application of Grubbs' Test. Grubbs' Test can only be used in single dimensional data and it calculates Z-score of each observation, and it compares Z-scores with critical value of 95% significance level (Grubbs, 1969).

Another simple technique which again can only be applicable to single dimensional data is using boxplots to determine potential outliers visually. Points which placed 1.5 times of interquartile value away from minimum and maximum points can be classified as outliers (Laurikkala et al., 2000). Laurikkala et al. (2000) also suggest using Mahalanobis distance for multivariate datasets in order to get rid of “Curse of Dimensionality” which refers to the problem of having several features in dataset.

Mahalanobis distance is calculated with below formula (Mahalanobis, 1936);

$$\sqrt{(x-\mu)^T C^{-1}(x-\mu)}$$

C is the covariance matrix of features in given multivariate dataset and μ is the mean (can be called as centroid) of features. Centroid is the intersection point of means of all features in multi-dimensional space. Mahalanobis distance calculates how far away an observation from the centroid with respect to covariance matrix of given data. As calculation of Mahalanobis distance takes into account of covariance matrix of given dataset, calculated Mahalanobis distance of an observation is inevitably affected by existing outliers in dataset.

Many researchers worked on the subject of obtaining more robust estimators of covariance. Minimum covariance determinant (MCD) can be the solution for getting robust parameters for Mahalanobis distances (Rousseeuw, 1985). MCD algorithm iteratively calculates estimated μ and C (Covariance Matrix) parameters in Mahalanobis distance formula by taking the closest points of dataset in each iteration. By taking the closest points, impacts of outlying observations in MCD estimation drastically lowered (Hardin and Locke, 2005). Rousseeuw and Van Driessen (1999) also put forward another formula for calculation of MCD. The updated formula named as Fast-MCD is more robust and faster than the previous MCD algorithm and it is much more computationally easier than the previous formula.

Some outliers are not observable via distance functions since they are not considered as outliers in the definition of global outlier. Observations can also be taken as outliers when they are compared with their nearest neighbors or their nearest cluster. These outliers are known as local outliers (Breunig, Kriegel, Ng & Sander, 2000).

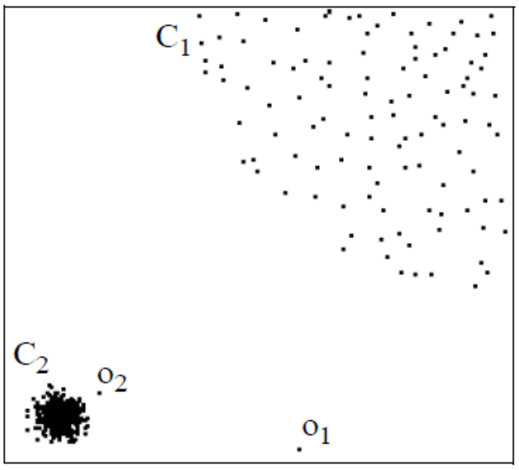


Figure 1: Visualization of Local Outliers (Breunig et al., 2000)

As shown in *Figure 1*, point O_1 is obviously a global outlier. Furthermore, it can be observed that cluster C_2 is much denser than cluster C_1 since points in C_1 are seemed to be very close to each other. With that much density exhibited by cluster C_2, point O_2 can be labeled as local outlier.

The most remarkable addition to the literature about local outliers proposed by Breunig et al. (2000) is being an outlier is not a binary situation. They created a new algorithm, named as Local Outlier Factor (LOF), to detect outliers by assigning a point to each observation to determine whether an observation is an outlier or not. LOF is a density-based anomaly detection approach. In order to measure LOF score of an observation; k-distance, reachability distance and local reachability density of an observation must be calculated respectively.

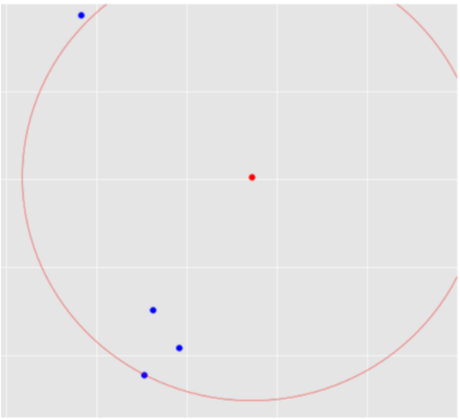


Figure 2: Visualization k-distance of an observation *(k=3)*

LOF score of an observation can be obtained by dividing mean of local reachability density of an observation's k-nearest neighbors to local reachability of an observation itself.

K-distance is a distance between observation and its k-th nearest neighbor points as shown in *Table 2* where k specified as 3.

Isolation Forest (iForest), which is a machine learning based approach proposed by Liu, Thing & Zhou (2008), isolates anomalous observations rather than learning from normal observations. Main difference of iForest is it exploits two specific attributes of outliers: i) outliers are very uncommon among all observations (generally correspond to less than 1% of data), ii) outliers have very different characteristics (when they compared to normal points). iForest is an ensemble machine learning method that creates many trees to specify anomalous observations. Outliers are observations that have small number of average path lengths in created trees (Liu et al., 2008). Particularly, normal observations need much more cut to be labeled as a normal point while outliers demand less partition to be identified as outliers.

1.2. About Factoring Sector and XXX Factoring

Factoring is a financial product that is assigned by a factoring company to lend the receivables of the companies arising from the sales of goods and services. Factoring customers commonly use term checks on their trade receivables as a means of payment. Since factoring is essentially a credit to trade, it is important to check and verify the correctness of the trade both by law and by risk assessment. Thus, detecting anomalous activities precisely in transactions is very vital for factoring companies.

XXX Factoring received the operating permit from the Banking Regulatory and Supervision Agency of Turkey in 2012. With more than 500 employees and 34 branches, XXX Factoring operates in 21 different cities within Turkey. XXX Factoring has a partnership with European Bank for Reconstruction and Development (EBRD) and strong capital structure that XXX Factoring has paid-in capital of more than 116 million TRY. With its technological infrastructure and innovative spirit, XXX Factoring established the first R&D center among non-bank financial institutions. XXX Factoring aims to find quick and effective solutions to the urgent financing needs of small and medium-sized enterprises (SME) by using its extensive branch of network and its alternative channels.

2. ABOUT FACTORING DATASET

Dataset used in this study is provided by XXX Factoring and dataset contains information about daily transactions of XXX Factoring and characteristics of checks, drawers and payees/sellers. Before description of dataset, it is better to present basic definitions about factoring sector. There are three main players in factoring sector; factoring institution, drawer and payee/seller. Drawer is the person or legal entity who writes a check and makes the payment at the expiration date of the check. Payee (or seller) is the person or legal entity who receives the payment. Obviously, quality and trustworthiness of a check mainly depend on drawer's ability to pay it back at the expiration date.

2.1. Description of Dataset and Preprocessing

Factoring dataset initially has total 890.118 rows and 122 features. Dataset contains information about checks (expiration date, amount, assigned color by XXX Factoring, final approval/rejection status, etc.), drawers and payees/sellers (total risk amount, total debt amount, number of completed transactions, etc.). More detailed description about features will be presented at the end of this part.

Many features have huge numbers of missing values. Threshold for missing value ratio is set to 40% and features which include missing values more than this threshold were omitted from dataset. Hence, 29 features were eliminated from dataset and these features are listed in Appendix.

Since provided dataset is the combination of company's several different inner databases, it would be better to check duplicate rows and columns. After the duplication check, all rows are turned out to be unique but there are 6 duplicate features. These features are removed from dataset and listed in Appendix. Furthermore, there are also many near duplicate features which means some features are updated versions of other features and they are very much identical. These features are also eliminated from dataset and explained in Appendix.

In order to find potential anomalous observations in dataset, we don't really need any information about identity number of a person or a company name of legal identity. In addition to that, there are also unnecessary features about system messages such as whether information about check is received from the company's intelligence system or not. Keeping these uninformative features would create confusion in the applications of both statistical and machine learning based methods. To prevent such confusion, features with system and ID information are dropped from dataset and listed in Appendix.

Uninformative features, categorical features that consist of only one unique value and miscoded features (numeric features which must be encoded as categorical) are eliminated from dataset and explained in Appendix.

XXX Factoring uses its own algorithm to assign a color to each received check. Dataset has four features about colors of checks given by company's scoring algorithm. *"CEK_CUTOFF_CEKRENK"* and *"CEK_PRE_CEKRENK"* are nearly same features, they represent the initial colors of checks. Then, colors of checks are updated by company's decision tree algorithms which creates a new feature named as *"CEK_KIOSCEKRENK"*. *"CEK_KIOSCEKRENK"* has four main check colors (Red, Yellow, Green and Black) and three supplementary check colors (AYellow, AGreen and Orange). At the final stage, which corresponds to *"CEK_CEKRENK"* feature, three supplementary check colors are converted into main colors. Thus, it is better to keep feature that represents final four colors of checks (*"CEK_CEKRENK"),* while other three features are dropped.

Categorical features consist of "Yes" and "No" labels are turned into binary data format in order to get better outputs (listed in Appendix). There is also another categorical feature *"E_ANASTATU"* which contains information about check's final situation whether it is rejected or confirmed. Some checks are labeled as "Fiyatlaniyor", "Kesinlesti" and "LimitTahsis" in that feature and all of them are only account for 1031 observations. Rows containing these labels are omitted from dataset since these three labels are quite ambiguous in terms of explaining final situation of the check. Most importantly, this feature will be supplementary checker of developed models since if anomaly detection algorithm spots an anomaly, it can be controlled via this feature whether check is confirmed or rejected at the end.

There were also Turkish letters and white spaces in responses of categorical features. Turkish letters transformed into English letters and all white spaces are removed for better analysis.

2.2. Missing Value Replacement

After brief description of dataset, missing values are checked for both numeric and categorical features.

"E_ISLEMSEGMENT" is a categorical feature which shows segmentation of check sellers and it has 202 missing values which is a drastically small number when it is compared with total number of observations (890.118). Rows that contain missing values for this feature basically dropped from dataset in order to prevent possible distortion since this feature has 20 unique values and such imputation of missing values with mode, median or any other imputation technique may lead to false outcomes in the analysis.

"STC_TAKIPKODU" comprises of 6.5% of missing values and it has 14 unique categorical labels. This feature contains information about legal situations of check sellers. Missing values are not imputed since these sellers could possibly be in a bad legal position or not. Imputing missing values with mode, median or any other technique may also misrepresent the general outlook of data in this case and this feature is omitted from dataset. Similarly, *"KSD_TAKIPKODU"* represents legal situations of drawers but this feature includes more than 25% of missing values. Imputing 25% of missing values with 14 unique labels will absolutely damage the dataset, thus *"KSD_TAKIPKODU"* is dropped from dataset.

"CEK_CEKORTVADEGUNSAYISI" is a numeric feature with missing values and it shows remaining average days until expiration date of a check. As dataset also has another feature which exactly contains same information about check's expiration date. There is no logical reason to keep this feature and it is eliminated from dataset.

"STC_RISK" shows current calculated risk of check sellers and it has 20% of missing values. Replacing with that much missing values could be problematic and replacement may lead false outcome in the analysis. Consequently, *"STC_RISK"* is dropped from dataset.

"ISLEM_TIPI" contains more than 20% of missing values and it shows the type of transaction whether transaction is "new", "current" or "activation". Imputation of missing values will directly impact 20% of total data. Thus, this feature is also dropped from dataset rather than imputation of missing value.

At the end of this chapter, number of features reduced very much with the aim of minimum information lose. Final dataset has total 888.883 rows and 36 features, 13 features are categorical type and 22 features are numeric type.

2.3. Exploratory Data Analysis

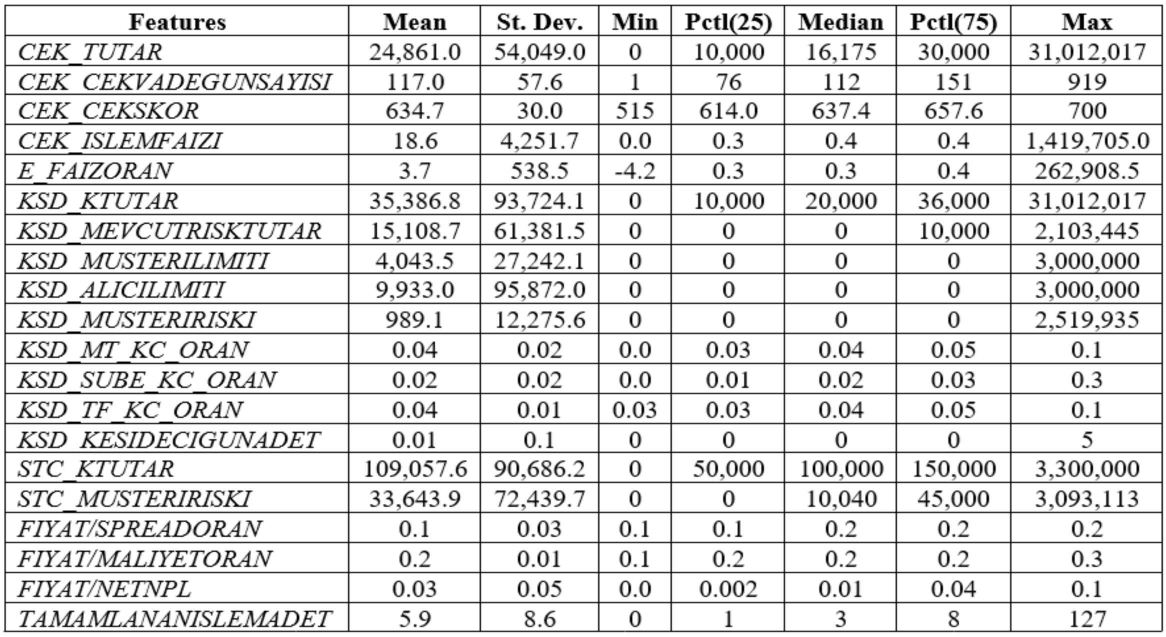
After many preprocessing steps, it is observed that one numeric feature is actually a date information. Hence, there are 21 numeric features left and they are described in below *Table 1*;

Table 1: Description of Numeric Features

Feature	Description
"CEK_TUTAR"	Shows check amount
"CEK_CEKVADEGUNSAYISI"	Shows remaining days until expiration date of a check
"CEK_CEKSKOR"	Shows check score
"CEK_ISLEMFAIZI"	Indicates interest rate of a check weighted by check amount and expiration date of check
"CEK_HESAPLANANFAIZORAN"	Shows interest rate of a check
"E_FAIZORAN"	Indicates the rate at which the check is used
"KSD_KTUTAR"	Shows the total amount of checks belonging to the same drawer
"KSD_MEVCUTRISKTUTAR"	Shows how much risk the drawer has inside the company
"KSD_MUSTERILIMITI"	Shows payee’s limit
"KSD_ALICILIMITI"	Indicates whether the drawer has a recipient limit
"KSD_MUSTERIRISKI"	Shows how much risk the drawer has (if the drawer is also the customer at the same time)
"KSD_MT_KC_ORAN"	*Ratio (ambiguous)*
"KSD_SUBE_KC_ORAN"	*Ratio (ambiguous)*
"KSD_TF_KC_ORAN"	*Ratio (ambiguous)*
"KSD_KESIDECIGUNADET"	The number of days in the contact table for drawers
"STC_KTUTAR"	Shows payee’s current limit
"STC_MUSTERIRISKI"	Shows payee’s current risk at the time of the transaction
"FIYAT/SPREADORAN"	The spread rate determined by payee's segment
"FIYAT/MALIYETORAN"	The cost ratio determined by the expiry date of a check
"FIYAT/NETNPL"	The ratio of NPL determined by the color of a check
"XXXAMLANANISLEMADET"	Shows the number of transactions completed

One feature that shows calculated interest rate of a check is omitted from descriptive statistics since it has very huge mean and maximum value, thus it will be analyzed later. Descriptive statistics of numeric features presented below in *Table 2*;

Table 2: Descriptive Statistics



Features	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
CEK_TUTAR	24,861.0	54,049.0	0	10,000	16,175	30,000	31,012,017
CEK_CEKVADEGUNSAYISI	117.0	57.6	1	76	112	151	919
CEK_CEKSKOR	634.7	30.0	515	614.0	637.4	657.6	700
CEK_ISLEMFAIZI	18.6	4,251.7	0.0	0.3	0.4	0.4	1,419,705.0
E_FAIZORAN	3.7	538.5	-4.2	0.3	0.3	0.4	262,908.5
KSD_KTUTAR	35,386.8	93,724.1	0	10,000	20,000	36,000	31,012,017
KSD_MEVCUTRISKTUTAR	15,108.7	61,381.5	0	0	0	10,000	2,103,445
KSD_MUSTERILIMITI	4,043.5	27,242.1	0	0	0	0	3,000,000
KSD_ALICILIMITI	9,933.0	95,872.0	0	0	0	0	3,000,000
KSD_MUSTERIRISKI	989.1	12,275.6	0	0	0	0	2,519,935
KSD_MT_KC_ORAN	0.04	0.02	0.0	0.03	0.04	0.05	0.1
KSD_SUBE_KC_ORAN	0.02	0.02	0.0	0.01	0.02	0.03	0.3
KSD_TF_KC_ORAN	0.04	0.01	0.03	0.03	0.04	0.05	0.1
KSD_KESIDECIGUNADET	0.01	0.1	0	0	0	0	5
STC_KTUTAR	109,057.6	90,686.2	0	50,000	100,000	150,000	3,300,000
STC_MUSTERIRISKI	33,643.9	72,439.7	0	0	10,040	45,000	3,093,113
FIYAT/SPREADORAN	0.1	0.03	0.1	0.1	0.2	0.2	0.2
FIYAT/MALIYETORAN	0.2	0.01	0.1	0.2	0.2	0.2	0.3
FIYAT/NETNPL	0.03	0.05	0.0	0.002	0.01	0.04	0.1
TAMAMLANANISLEMADET	5.9	8.6	0	1	3	8	127

When we look at the summary statistics in *Table 2*, we observe that check score is quite normally distributed since mean and median values of check score are quite close to each other and 1st and 3rd quantiles are seemed to be symmetric. Price/Spread and Price/Cost features are also presented very similar statistics like check score and they seem to be normally distributed.

When we closely analyze *"KSD_KESIDECIGUNADET"* feature which shows how many days have been passed since drawer is added to the company's database, it is observed that 99.5% of values are equal to zero. Consequently, *"KSD_KESIDECIGUNADET"* is dropped from dataset. Moreover, *"KSD_MUSTERILIMITI"*, *"KSD_ALICILIMITI"* and *"KSD_MUSTERIRISKI"* contain more than 96% of zero values but there are some outliers clearly. Having such big values in these features could be a sign for anomaly, hence these features are not omitted from dataset.

Interest rates of checks are described in three different features, when we closely analyze the checks initial interest rate feature, we see that 99.9% of interest rate values are "0". Another feature about interest rate which shows calculated interest rate of a check by XXX Factoring has three huge outliers and other observations are rated between 0 and 1. Boxplot of calculated interest rates presented and three outliers are not included in below *Figure 3*;

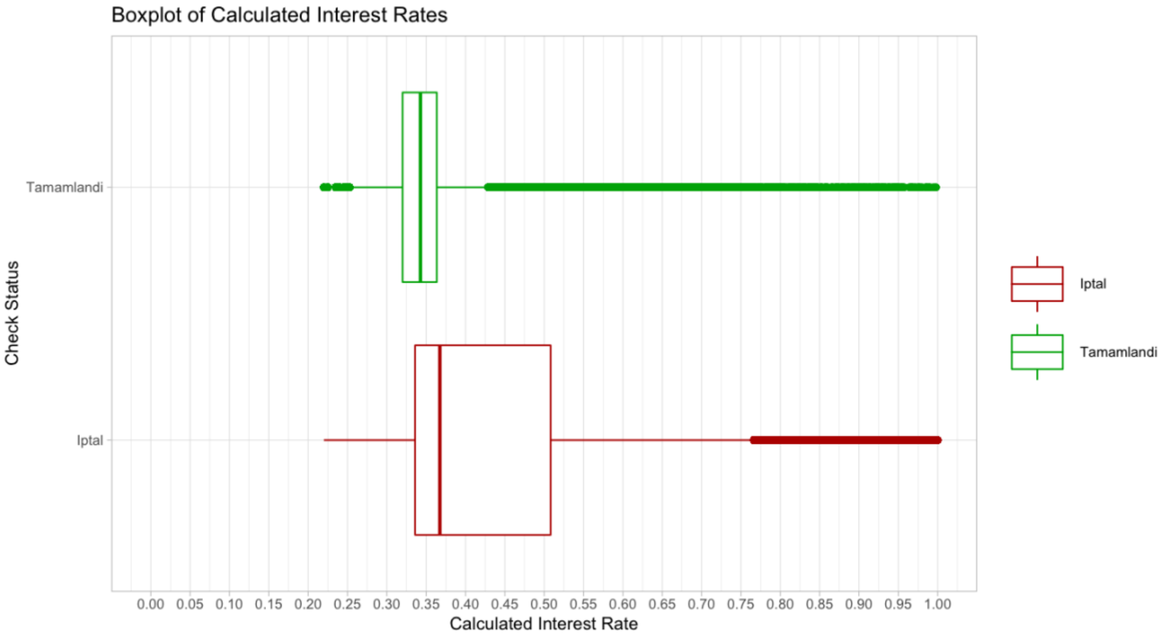


Figure 3: Boxplot of Calculated Interest Rates (approval status is colored)

As above *Figure 3* shows, rejected checks and accepted checks have many outliers. On the other hand, we can say that XXX Factoring mostly accepts checks which have calculated interest rate between 25% and 45%. Interestingly, there is not a single accepted check rated below 20% and XXX Factoring accepted checks with unrealistic calculated interest rates which could be considered as an anomaly.

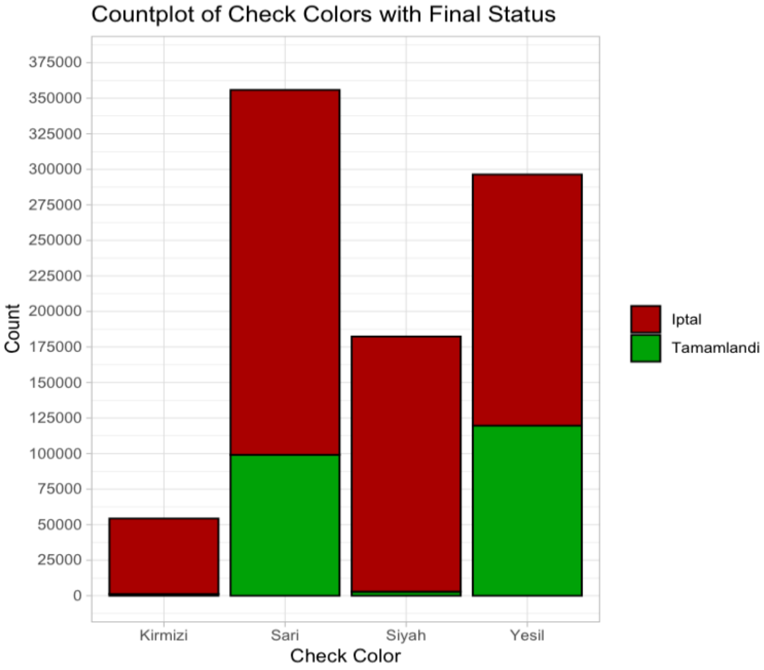


Figure 4: Count Plot of Check Colors with Final Status

Count plot of final colors of checks with final situation of checks whether they are rejected or accepted presented in *Figure 4*. It is so obvious that XXX Factoring generally works with yellow and green colored checks. There is a very limited number of accepted black and red checks but we can say that red and black checks are almost always labeled as rejected.

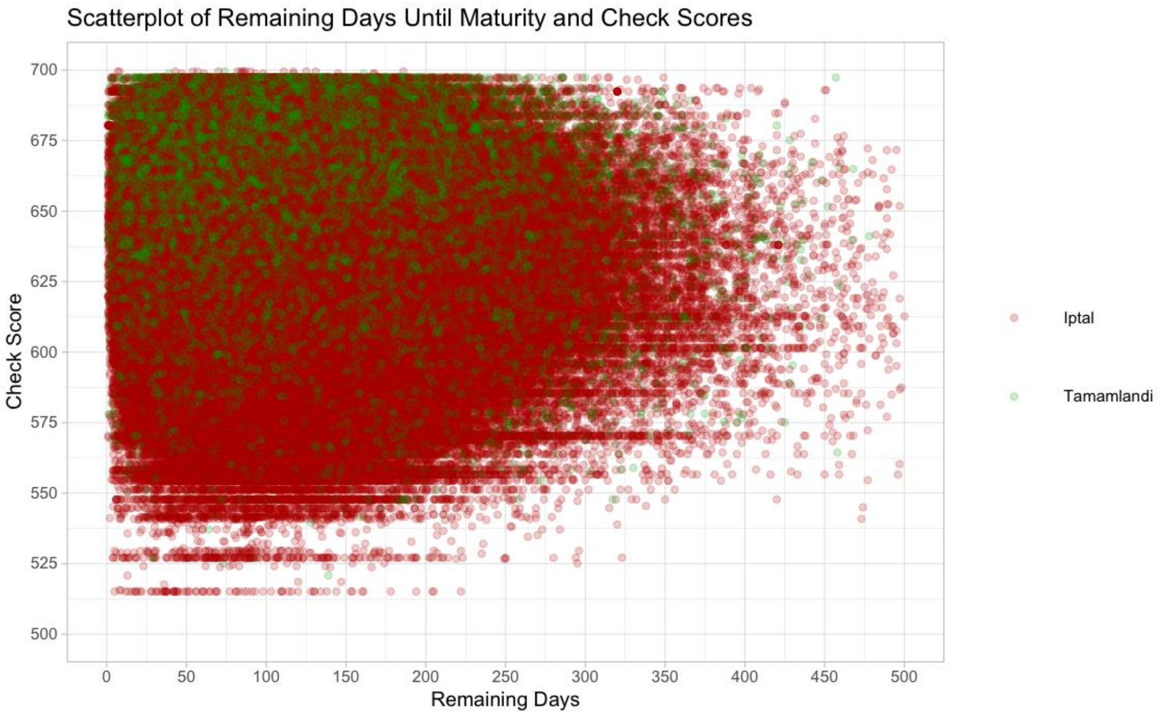


Figure 5: Scatterplot of Remaining Days until Maturity and Check Scores

Remaining days until expiration of checks and scores of checks are visualized in above scatterplot graph *(Figure 5)*. Approved checks labeled as green in the graph, and company's approval criteria tried to be identified. As graph tells, XXX Factoring mostly approves checks that have check score of higher than 600 points and remaining days until expiration between 0 and 300 days. Many checks which can be considered as highly probable anomalous cases are accepted by XXX Factoring. For instance, checks which have check score lower than 600 points and remaining days until expiration longer than 400 days could be labeled as anomaly intuitively.

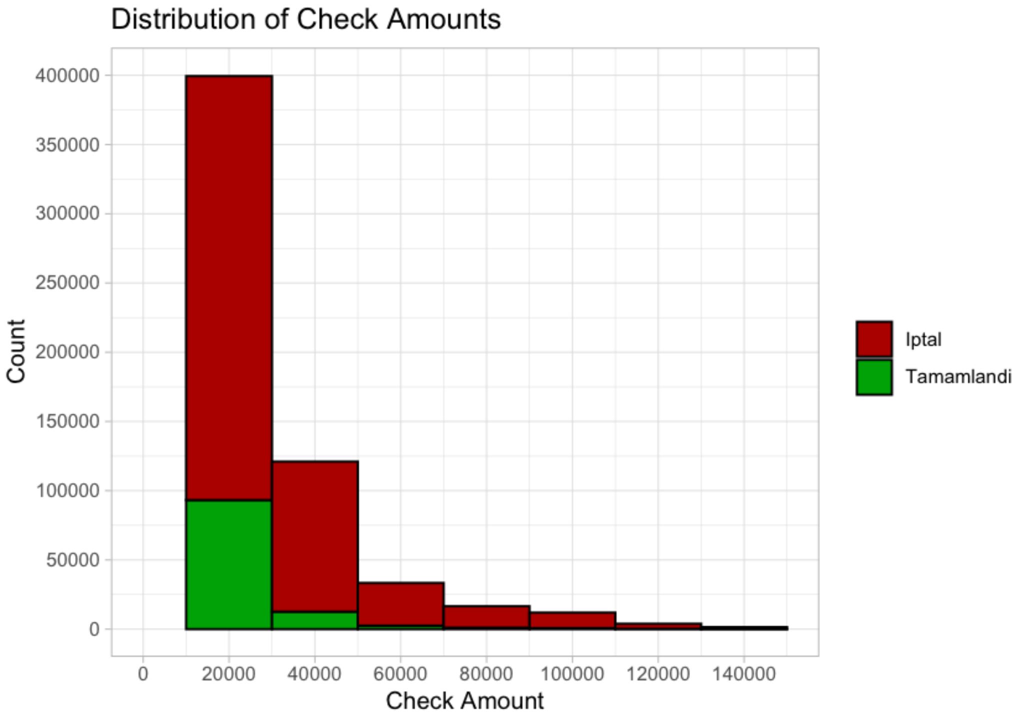


Figure 6: Distribution of Check Amounts

In order to visualize amounts of checks, subset of dataset has been used since some checks have very big written amounts. It is observed that only 6607 observations have more than 150k check amount which corresponds to 0.007% of the whole dataset. Without these outliers, histogram about check amounts has been presented in *Figure 6*. Most of the check amounts are distributed between 20k and 40k which implies that company mostly approves checks that have small amounts.

3. ABOUT PROJECT & METHODOLOGY

In this part, motivation and definition of this study is presented with expected outcomes of XXX Factoring and with three different adopted methodologies. Expectations about project is noted in the meetings held with XXX Factoring.

3.1. Project Definition

XXX Factoring obtains financial and check usage data of the payee (or seller) and the drawer from Credit Bureau. XXX Factoring can see the attributes of all queried checks and records in the database regardless of the realization of the transaction.

Financial and check usage data contains information about; bank limits, whether or not it has previously done factoring, how many companies it has worked with, the status of its debts, payment performance, first and last check date, last payment date, number of checks issued, total amount of issued checks, payment performance, etc.

XXX Factoring conducts its factoring credit risk assessment through analytical techniques using its own scorecard algorithm and decision support workflow application. Although current algorithm measures credit risk properly, it is insufficient in terms of responding internal and external anomalies or outliers.

Since factoring is basically a micro credit given to businesses, it is very crucial to check and verify the correctness of each transaction. As the number of transactions increases, it is practically difficult to carry out all possible controls in each process and it brings a very heavy burden on control units in the company. Consequently, both control points and risk assessment processes against the anomalies were believed to be inadequate by XXX Factoring.

3.2. Expected Outcomes

With ongoing anomaly detection project, it is aimed to establish an early warning mechanism for loans, sales and operation units by detecting internal and external anomalies or outliers in XXX Factoring’s transactions.

Expected project outputs decided by XXX Factoring listed below;

- Instant detection of anomalies in transaction process
- Determination of additional control points to be applied in the transaction process

- Detection of anomalies for post-procedure controls (to contribute to the internal control process)
- Increasing productivity by associating anomalies with performance
- Reducing risk by linking anomalies to fraud

As a result of this project, it is planned to measure commercial anomalies more accurately, reduce commercial and operational risks, increase operational efficiency and increase the efficiency of internal control process.

3.3. Methodology

In order to build a robust anomaly detection mechanism, three different anomaly detection methods (Mahalanobis Distance, Local Outlier Factor and Isolation Forest) have been applied to factoring data and performance of each algorithm is evaluated. Statistic-based, density-based and machine learning based approaches have been used and compared in the study. Statistic-based and density-based approaches are rule-based methods in practice, while machine learning based approach can be considered as model-based method.

After preprocessing step; we have total of 888.883 rows, 20 numeric features and 16 categorical features. Since Local Outlier Factor and Isolation Forest methods are quite computationally expensive, random sampling has been applied to dataset and 100k rows are randomly selected without replacement. As Mahalanobis and LOF algorithms are only applicable to numeric features, all three models, including Isolation Forest, are built with numeric features in order to evaluate each model equally. Models are built in RStudio[1] which is a commonly used open-source programming language. Mahalanobis distance, LOF and Isolation Forest are calculated via R's built-in packages, "dbscan" package (Hahsler et al., 2019) and "solitude" package[2].

In the study, Mahalanobis distance has been used to detect outliers as a statistical approach. Mahalanobis distance and minimum covariance determinant (MCD) based Mahalanobis distance of each observation are calculated and outputs of both methods are compared. Noises are randomly created for MCD-based Mahalanobis distance calculations

[1] R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

[2] Komala Sheshachala Srikanth (2019). solitude: An Implementation of Isolation Forest. R package version 0.2.0. https://CRAN.R-project.org/package=solitude

for covariance matrix via “mvtnorm” package (Genz and Bretz, 2009). The reason behind calculation of regular Mahalanobis distance is to evaluate robustness of MCD-based Mahalanobis distance.

As a density-based approach, LOF score of each observation is calculated and only numeric features are used since LOF is not applicable to categorical features. LOF score of an observation basically shows the density of an observation relative to density of observation’s neighbors. Observation that scored more than 1 should be perceived as an outlier which suggested by Breunig et al. (2000). Observation which has more density than its relatives expected to have LOF score of smaller than 1 implying that observation is not an outlier. “k” value which defines the number of neighbors of an observation for calculation of LOF scores is assigned as *(k=10)* initially which is a minimum viable value for LOF calculations as the authors of the study suggest. However, different k values are tested and impacts of k values presented in the next chapter.

Isolation Forest (iForest) which is a machine learning based algorithm is applied to factoring dataset. Main difference of this method is inclusion of categorical features but categorical features in factoring dataset are not used in the models in order to avoid unequal evaluation of models. iForest method has two parameters which are number of trees and size of sample fraction. For parameter optimization, size of trees iteratively increased but it is observed that running time of algorithm doubled at the same time. Thus, size of trees specified as (n=100). For size of sample fraction, smaller values presented better results so size of sample gradually decreased for better results.

4. EVALUATION OF ANOMALY DETECTION METHODS

In this part, three different anomaly detection methods have been applied to the factoring dataset and all models are evaluated in terms of their anomaly detection performances and robustness. All of the applied anomaly detection approaches have one thing in common which is they assign a score to each observation and the highest scored observations are considered as highly probable anomalous cases. In order to evaluate each model equally, top 10% of the assigned anomaly scores has been taken for each model. Then, these scores are compared with *"E_ANASTATU"* feature which shows whether check is accepted at the end of the transaction or not. This feature can be used as a supplementary target feature because we do not really have any feature for anomalous observations in the factoring dataset.

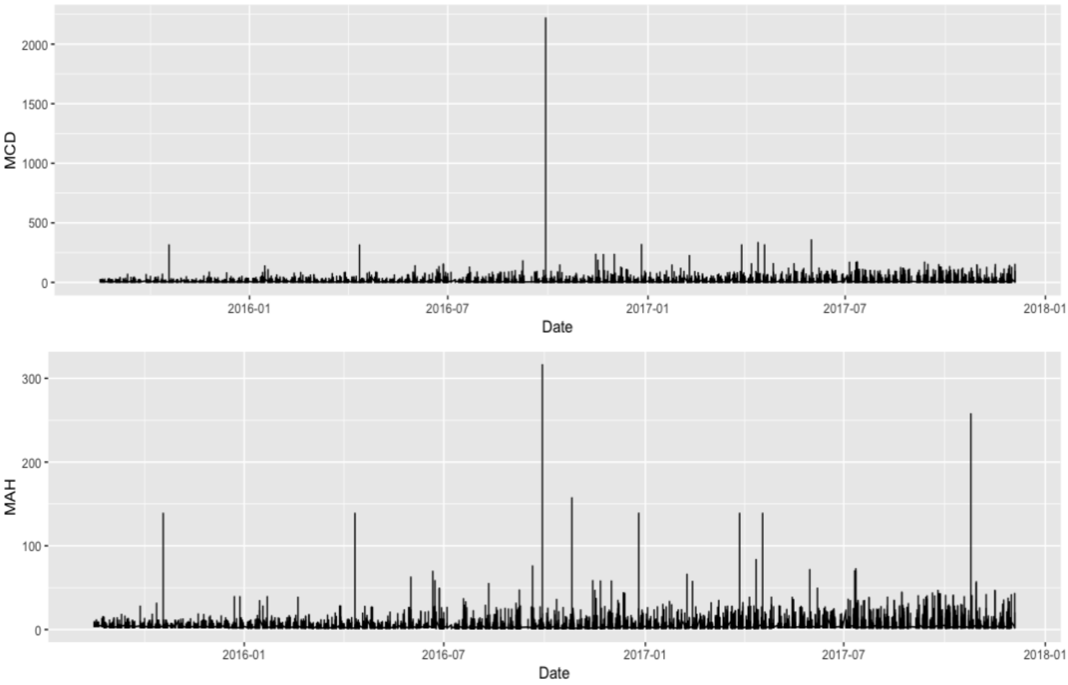
Accuracy scores of MCD-based Mahalanobis distance and regular Mahalanobis distance are presented below in *Table 3*.

Table 3: Accuracy Scores of Mahalanobis Distances

	Mahalanobis Distance	**MCD-based Mahalanobis Distance**
Accuracy Score (in top 10% of scores)	81.30%	84.93%

As a rule-based anomaly detection approach, MCD-based Mahalanobis distances presented very fast and satisfying results. When we closely analyze top 10% of the highest Mahalanobis distances and MCD-based Mahalanobis distances, it is observed that MCD-based Mahalanobis algorithm predicted anomalies better than regular Mahalanobis algorithm. This can be a proof of using robust covariance estimator gives better results since MCD is not effected by existing outliers. MCD-based Mahalanobis distances is the fastest algorithm among other models and it would be very convenient for XXX Factoring since company wants to respond anomalous cases instantly. Performances of both algorithms presented visually in *Figure 7*.

Figure 7: Time-series Plot of MCD-based and Regular Mahalanobis Distances



As *Figure 7* shows both of the algorithms spotted anomalies fairly well. However, regular Mahalanobis distances (MAH) presented very high scores for some of the anomalies observed as sharp hikes in *Figure 7*, while MCD-based Mahalanobis distances (MCD) presented straightforward and much closer scores (except one observation). As covariance matrix used in regular Mahalanobis distance calculations are affected by existing outliers, above graph can also be an evidence for robustness of MCD-based Mahalanobis calculations.

For Local Outlier Factor (LOF) algorithm, top 10% of the observations with the highest anomaly scores has been compared with corresponding supplementary target feature for evaluation instead of assigning a threshold value for being an anomaly. Assigning a threshold value for LOF scores considered as a double-edged sword because LOF scores which are very close to threshold value can be falsely labeled as an outlier or vice versa. Breunig et al. (2000) suggest using different k values ranged between 10 and 50 according to the density of clusters in dataset. Different k values are tested in the analysis and accuracy of LOF algorithm with different k values presented in *Table 4*.

Table 4: Accuracy Scores of LOF

	LOF *(k=10)*	**LOF** *(k=20)*	**LOF** *(k=30)*	**LOF** *(k=40)*	**LOF** *(k=50)*
Accuracy (in top 10% of LOF scores)	82.84%	83.52%	83.94%	84.13%	84.21%

LOF algorithm with *(k=10)* correctly labeled anomalous cases with the accuracy rate of 82.84% but when k value increased accuracy is also increased simultaneously. At final, accuracy score rises up to 84.21% when k value initialized as 50. Obtaining the highest accuracy with *(k=50)* indicates denser clusters in factoring dataset. Nevertheless, LOF algorithm presented pretty well results with bigger k values, MCD-based Mahalanobis algorithm presented better accuracy score and much faster results.

As a machine learning based approach, accuracy scores of Isolation Forest algorithm presented in *Table 5* with different sizes of sample fraction while number of trees kept at 100. It is observed that algorithm presented better and faster results with 100 trees which is suggested tree number by the authors of the study (Liu et al., 2008). At the same time, decreasing sample fraction of dataset fitted into the model increased the accuracy of the algorithm until size of sample fraction is equal to 0.1.

Table 5: Accuracy Scores of Isolation Forest Applications (with different size of sample fraction)

	iForest *(sample fraction=1)*	**iForest** *(sample fraction=0.6)*	**iForest** *(sample fraction=0.3)*	**iForest** *(sample fraction=0.2)*	**iForest** *(sample fraction=0.1)*
Accuracy Score (in top 10% of IF scores)	80.81%	82.96%	84.69%	84.98%	86.63%

When size of sample fraction is set to 0.1, we got accuracy score of 86.63% which is the highest accuracy score among all applied methods but iForest algorithm works very slow when it is compared with the second best scoring algorithm MCD-based Mahalanobis distance.

As XXX Factoring prioritize the fastest algorithm, there is a trade-off between iForest and MCD-based Mahalanobis algorithms. However, there is a slight difference between accuracy scores of both algorithms, MCD-based Mahalanobis presented anomaly scores of observations in a second. Thus, it is decided that MCD-based Mahalanobis algorithm is the most suitable approach for XXX Factoring.

5. CONCLUSION & DELIVERED VALUE

5.1. Conclusion

In this study, statistics-based, density-based and machine learning-based approaches are applied to the factoring dataset with random sampling. After the comprehensive feature elimination stage, number of features drastically lowered and the categorical features are not included in the analysis for equal evaluation of three different models since Mahalanobis and LOF calculations are only applicable to numeric features. All applied algorithms in the study assign an anomaly score to each observation and observations with the highest scores labeled as an anomaly.

Performance of each algorithm is evaluated by taking top 10% of the observations with the highest scores as an anomalous cases and selected observations are compared with supplementary target feature which shows whether a check accepted or rejected at the end. In this perspective, MCD-based Mahalanobis distances presented fairly fast and powerful results with the accuracy of 84.93%. Although Isolation Forest algorithm scored slightly higher (86.63%) than MCD-based Mahalanobis algorithm in terms of accuracy, Isolation Forest was slower in calculations when it is compared with MCD-based Mahalanobis calculations. As factoring dataset does not have a reliable label for anomalous observations, taking top 10% of the observations with the highest scores and labeling them as anomaly considered as an only option for evaluation.

5.2. Delivered Value

As XXX Factoring aims to improve its internal audit process; by taking top 10% of the observations with the highest anomaly scores and labeling these observations as an anomaly for review, internal audit process of XXX Factoring became more effective and concise at the same time. XXX Factoring conducts its audit process daily and specifying suspicious observations for further investigation will be very time efficient for XXX Factoring. In addition to that, the workload on audit teams is also tried to be reduced.

Secondly, XXX Factoring also wanted a fast working algorithm to proactively take precautions against anomalous activities in transactions, MCD-based Mahalanobis algorithm presents instant anomaly scores for all observations which satisfies one of the XXX Factoring's success criteria.

5.3. Further Steps

Dataset of XXX Factoring does not have a feature for observations labeled as anomalies and therefore performances of models are evaluated according to acceptance and rejection status of checks at the end of the day.

However, it has seen that XXX Factoring accepted checks which can be considered as anomalies. Therefore, although the performance results of the models are not completely reliable, close studies should be carried out with XXX Factoring as a next step in order to evaluate presented outputs better. Due to the continuation of the project process with XXX Factoring, it is important to re-examine the observations with the highest scores with XXX Factoring.

APPENDIX

- 29 features containing more than 40% of missing values are listed below;

"KSD_SONUC", "KSD_KARARFIRMASEGMENT", "KSD_KARARSONUC", "STC_DUYUM", "STC_KARARSONUC", "STC_HAMCEKID", "STC_HAMMEMZUCID", "ISTISNA/MAHSUP", "ISTISNA/CEK TUTARI KUCUK", "ISTISNA/GUCLU SATICI", "ISTISNA/BUYUK KESIDECI", "ISTISNA/DIGER", "ISTISNA/GUCLU CIRANTA", "DUZENLEME/KULLANICI TALEBI-RNK", "STC_SUBE", "ISTISNA/GENEL", "ISTISNA/CEK<=2000", "E_CEKHESAPNO", "E_CEKBANKA", "E_CEKSUBE", "E_CEKNO", "E_CEKRECNO", "E_CEKSTATU", "E_VKN", "FB_FIRMAADI", "E_ISTONAY", "E_ONAYLAYANAD", "E_ISTUZMANIAD", "E_ISTSONONAY"

- 6 duplicate features are listed below;

"CEKID2", "STC_KAYNAKNO", "STC_KISIID", "CEK_ID", "STC_STATU", "STC_MUSTERILIMITI"

- 22 features that include ID and name information listed below;

"KSD_SATICI_KISIID","YENI_VKN", "YENI_CEKNO", "CEK_RECNO", "CEK_CEKNO","CEK_CEKHESAPNO","CEK_KARARCEKID", "KSD_KISIID", "KSD_VKN","KSD_FIRMANO", "KSD_FIRMAADI", "KSD_HAMMEMZUCID", "STC_VKN", "STC_FIRMANO", "STC_FIRMAADI", "STC_MT", "STC_ANASEKTOR", "STC_MTKOD", "KSD_HAMCEKID", "KSD_HAMRISKID","KSD_KAYNAKNO", "STC_HAMRISKID"

- Features about system messages, *“E_ISTDURUM”*, *“E_DETAYSTATU”* and *“E_ANASTATU”*, show information about system process. They do not contain any useful information and these 3 features are eliminated from dataset.
- *“E_SUBE”* and *“KSD_SUBE”* show information about location where check is received. They are not fully identical but *“E_SUBE”* is updated version of *“KSD_SUBE”*, so *“KSD_SUBE”* is omitted from dataset.
- *“CEK_CEKSUBE”* and *“CEK_CEKBANKA”* features represent information about where check is received, and which bank is addressed in check respectively. Both features are encoded as numbers, but they are actually referring to places.

These features would distort statistical calculations and machine learning algorithms; hence both features are dropped from dataset.

- *"STC_FIRMASEGMENT"* feature shows assigned segment of check seller, but it is a numeric feature which again should be encoded in categorical format. On the other hand, we have another two features (*"STC_KARARFIRMASEGMENT"* and *"E_ISLEMSEGMENT")* which contain exactly the same information about seller's assigned segment with characters. Both *"STC_KARARFIRMASEGMENT"* and *"E_ISLEMSEGMENT"* features are 99% identical. It is better to drop *"STC_FIRMASEGMENT"* from dataset with *"STC_KARARFIRMASEGMENT"* since they are near duplicate features.
- *"CEK_FAIZORAN"* exhibits interest rate of a check similar to another two features but this feature shows zero interest rate for rejected checks. Thus, keeping other two features about check's interest rate is better while eliminating *"CEK_FAIZORAN"*.
- *"KSD_KAYNAK", "KSD_KISI_TIP", "STC_KAYNAK", "STC_KISI_TIP"* and *"KSD_STATU"* are categorical features and they only have one unique value. Basically, they have no information power and omitted from dataset. In addition to that, *"KSD_CREATEUSR"* feature have two unique values but one of them is only present in 0.005% of the all observations, thus it is dropped from dataset.
- Both *"KSD_CREATEDAY"* and *"KSD_CREATEDATE"* features contain same date information about when check is received from seller or payee. Keeping *"KSD_CREATEDATE"* would be better since it presents exact receive date with hour.
- *"KSD_KESIDECIGUNADET"* shows how many days have been passed after drawer of the check is entered into the system. *"KSD_KESIDECIHAFTAADET"* is weekly representation of the previous feature and it is omitted since we can say that it is near duplicate.
- *"CEK_CEKVADE"* and *"E_ISLEMTARIHI"* represent expiration date of a check and first arrival date of a check respectively. *"CEK_CEKVADEGUNSAYISI"* which is a numeric feature basically shows difference between expiration date and first arrival date of a check in day format. Thus, *"CEK_CEKVADE"* and

“E_ISLEMTARIHI” dropped from dataset, while keeping *“CEK_CEKVADEGUNSAYISI”*.

- *“CEK_PRE_CEKSKOR”* which presents scores of checks assigned by company’s own scoring algorithm found out to be very similar with *“CEK_CEKSKOR”* which is assigned final score of check. *“CEK_PRE_CEKSKOR”* is omitted from dataset.
- *“STC_FRMSGMNTGUNCELLEMETARIH”* shows exact date when working segment of seller/payee last updated, these dates ranges between 2015 and 2018. There are also 6% of missing values in that feature but imputation of date values is a quite challenging task. Since last update date of working segment basically has so little informative power, this feature is omitted from the dataset.
- 4 categorical features which are *“STC_STCILKISLEM”*, *“STC_SONUC”*, *“CEK_KARARSONUC”* and *“CEK_KARARSTATU”* consist of “Yes” and “No” labels. To get better analysis results, all these four features are converted into binary type.

REFERENCES

- Breunig, M. M., Kriegel, H., Ng, R. T., & Sander, J. (2000). LOF: Identifying Density-Based Local Outliers. *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data - SIGMOD 00*. doi:10.1145/342009.335388
- Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly Detection: A Survey. *ACM Computing Surveys,41*(3), 1-58. doi:10.1145/1541880.1541882
- Genz, A., & Bretz, F. (2009). Computation of Multivariate Normal and t Probabilities. *Lecture Notes in Statistics*. doi: 10.1007/978-3-642-01689-9
- Grubbs, F. E. (1969). Procedures for Detecting Outlying Observations in Samples. *Technometrics 11, 1–21.*
- Hahsler, M., Piekenbrock, M., & Doran, D. (2019). dbscan: Fast Density-Based Clustering with R. *Journal of Statistical Software, 91*(1). doi: 10.18637/jss.v091.i01
- Hardin, J., & Rocke, D. M. (2005). The Distribution of Robust Distances. *Journal of Computational and Graphical Statistics,14*(4), 928-946. doi:10.1198/106186005x77685
- Hawkins, D. M. (1980). *Identification of Outliers*. London: Chapman & Hall.
- Hodge, V., & Austin, J. (2004). A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review,22*(2), 85-126. doi:10.1023/b:aire.0000045502.10941.a9
- Laurikkala, J., Juhola, M. & Kentala, E. (2000). Informal Identification of Outliers in Medical Data. *Fifth International Workshop on Intelligent Data Analysis in Medicine and Pharmacology IDAMAP-2000*.
- Liu, F. T., Ting, K. M., & Zhou, Z. (2008). Isolation Forest. *2008 Eighth IEEE International Conference on Data Mining*. doi:10.1109/icdm.2008.17
- Mahalanobis, P. C. (1936) On the Generalised Distance in Statistics. *Proceedings National Institute of Science, 2.* India, 49–55.
- R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

- Komala Sheshachala Srikanth (2019). solitude: An Implementation of Isolation Forest. R package version 0.2.0. https://CRAN.R-project.org/package=solitude
- Reinsel, D., Gantz, J., & Rydning, J. (2018). DataAge 2025 - The Digitization of the World: From Edge to Core. Retrieved from https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf.
- Rousseeuw, P. (1985). Multivariate Estimation with High Breakdown Point. *Mathematical Statistics and Applications*,283-297. doi:10.1007/978-94-009-5438-0_20
- Rousseeuw, P. J., & Driessen, K. V. (1999). A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics,41*(3), 212-223. doi:10.1080/00401706.1999.10485670
- Schindhelm, R. K., Diamant, M., Dekker, J. M., Tushuizen, M. E., Teerlink, T., & Heine, R. J. (2006). Alanine aminotransferase as a marker of non-alcoholic fatty liver disease in relation to type 2 diabetes mellitus and cardiovascular disease. *Diabetes/Metabolism Research and Reviews*, *22*(6), 437–443. doi: 10.1002/dmrr.666
- Srivastava, A., Kundu, A., Sural, S., & Majumdar, A. (2008). Credit Card Fraud Detection Using Hidden Markov Model. *IEEE Transactions on Dependable and Secure Computing*, *5*(1), 37–48. doi: 10.1109/tdsc.2007.70228