UNIVERSITY OF TARTU
Social Sciences
Innovation and Technology Management

Mart Kevin Põlluste

DETECTING CORRUPTION IN PUBLIC PROCUREMENT THROUGH OPEN DATA ANALYSIS

Master Thesis (20 ECS) for applying to Master's degree in Social Sciences

Supervisor: Rajesh Sharma

Tartu 2019

DETECTING CORRUPTION IN PUBLIC PROCUREMENT THROUGH OPEN DATA ANALYSIS

Abstract:

Corruption is present in all aspects of the society and it hinders the progress of various sectors of the economy. In this context, corruption is defined as the act of dishonesty for personal gain by those in power. One of the biggest sectors it influences is public procurement. Previous research has shown that corruption is present in public procurement and it reduces the transparency of the process. Taking into account the monetary value of the public procurement sector, it is clear that this is a problem that must be addressed. Various studies have used qualitative analysis to root out the core of the issue, but as it still thrives, it essential that more accurate and acute measures are used.

In order to tackle this problem, there have also been studies that try to quantify the likelihood of it, rather than only looking at qualitative research and this is where data analytics comes into play – the core of this study. This thesis aims to determine whether using open data resources and data analytics it is possible to classify corruption in the public procurement processes and therefore suggest a suitable set of data to make the detection of corruption easier and quicker. Building on existing work on corruption, it asks: could corruption be predicted on available data and what methods should be used?

Based on a review of the literature on corruption and theories of machine learning, data analytics was used to assess possible corruption in public procurement in Estonia. In the data analytical process the author used machine learning approaches that predict the classification of procurement as corrupt or non-corrupt. The analysis of the results demonstrated that based on available data it is possible to predict corruption in public procurement in Estonia. Furthermore, the results also indicate that some features have a bigger impact on corruption in public procurement. Taking into account the background, related work and the current results, the author suggests that data analytics is vital in the fight against corruption and using machine learning can

yield in good results in predicting corruption. Further research is needed to identify other factors that could strengthen the effectiveness of these approaches.

Keywords:

Corruption, public procurement, machine learning

CERCS: P160

Annotatsioon:

Töö pealkiri on: "Korruptsiooni tuvastamine riigihangetes läbi andmeanalüüsi".

Magistritöö eesmärk oli uurida, kas olemasolevatele andmetele tuginedes ja andmeanalüüsi kasutades on võimalik hinnata korruptsiooni tõenäosust riigihangetes ning tulenevalt eelnevast anda soovitusi, kuidas andmed enda kasuks korruptsiooni vastu võistluses tööle panna.

Selleks, et antud eesmärki saavutada, andis autor ülevaate korruptsioonist üldiselt ja korruptsioonist riigihangetes. Lisaks andis autor ülevaate viimase aja olulisematest arengutest korruptsiooni vastu võitlemises maailmas ning selgitas analüüsis kasutatatud masinõppe algoritmide olemust. Korruptsioonile kui probleemile riigihangetes on viidanud mitmed uuringud nii Eestis kui mujal maailmas. Autor tugines enda töös riighangete registrist pärinenud andmetele ja täiendas neid kohtulahenditest leitud konkreetsete juhtumitega. Kasutades masinõppe lähenemisi hindas autor korruptsiooniriski ja visualiseeris tulemusi.

Andmete analüüsi ja masinõppe algoritmide tulemusena jõudis autor järeldusele, et olemasolevate andmete põhjal on võimalik edukalt hinnata korruptsiooniriski riigihangetes Eestis. Eelneva põhjal saab seega öelda, et andmete automaatse töötlusega on võimalik korruptsiooniriski tuvastamise protsessi muuta efektiivsemaks, mis omakorda hoiab kokku kaasnevat ressurssi. Lähtudes teooriast ja tehtud praktilisest tööst, esitas autor enda poolsed soovitused, milliste andmete kasutamisel ja analüüsil oleks võimalik korruptsiooniriski tuvastada ja seeläbi korruptsiooniriski maandada.

TABLE OF CONTENTS

Table of Figures

Table of Tables

GLOSSARY

AUC	Area Under the Curve
CPI	Corruption Perception Index
Corruption	Dishonest or fraudulent conduct by those in power
FDI	Foreign Direct Investment
Decision Tree	A decision support tool that uses a tree-like model of decisions
DIGIWHIST	Project Digital Whistle-Blower
KONEPS	e-Procurement system in Korea
Logistic Regression	Statistical model used to analyse the probability of a certain class
Machine Learning	An application of artificial intelligence that learn automatically
MAE	Mean Absolute Error
RMSE	Root Mean Square Error
Open Tenders	Platform for public procurement data in the EU
Precision	Percentage of relevant instances among the retrieved
Public procurement	The acquisition of goods on behalf of public authority
Random Forest	An ensemble learning method for classification, regression
Recall	Percentage of retrieved relevant instances over total relevant
Red flag	Information that might indicate corruption

Rent seeking	Attempt to obtain economic rent
Type I error	Rejection of a true null hypothesis
Type II error	Non-rejection of a false null hypothesis
XGBoost	Gradient boosting framework

1.INTRODUCTION

In modern societies public procurement is one of the ways governmental agencies take help from sources outside the public sector to achieve a task [1]. Transparency International (TI) defines public procurement as "the acquisition by a government department or any government-owned institution of goods or services." [2] For example, the need to build public roads is satisfied through public procurement. In the process of public procurement corruption might partially happen because the whole process of public procurement might not be clear.

Corruption can be and has been defined in various ways, but the World Bank puts it simple – corruption is the use of public or private office for personal gain [3], [4]. In particular, public procurement corruption is one of the most common means of using one's position for private gain and it is seen as the most vulnerable to corruption [5]. There have been efforts to manage the problem by implementing laws on institutions to publish open data to provide further transparency. The World Bank makes a distinction between two separate categories of corruption that could be applied to public procurement – state capture and administrative corruption [6].

Corruption in public procurement can have various effects on the country's economy. For example, it can often increase the cost and lower the quality of goods that therefore do not match customers' needs [4]. The OECD has assessed that corruption in public procurement costs between 20 to 25% of the respective country's budget [7]. There are also indirect costs that must be taken into account, like the distortion of competition, limited market access and low interest on foreign investments. Due to this, it is highly important to detect cases, where corruption might be happening in a timely manner and this is the aspect, where data analysis comes into play.

In the past, corruption has been analysed from different angles. For example there have been studies to discover the reasons why corruption takes place [1], [5], [8]–[11]. Also, there have been studies that look at corruption through qualitative research [12]–[18]. Analysing these papers, the author believes that a lot can still be done to make the process of public procurement more transparent by putting the data to work for us. Currently, the indexes like the Transparency International's Corruption Perception Index (CPI) and the World Bank's fight against corruption are all based subjective opinions of experts, who perceive the corruption of each country at

certain levels, based on the economic and political activity. Estonia is ranked number 18 with a score of 73 out of 100 in the 2018 CPI list. Compared to neighbouring countries like Russia with a score of 28 and Latvia with a score of 58, Estonia is doing well [2].

In comparison to the qualitative research, there are also studies that detect corruption through more data based approaches [19]–[23]. Although the number of studies on corruption through data analysis is increasing, there is still less work published on the topic than comparative and qualitative research, due to the fact that corruption is a complex phenomenon. Currently data analytics has been widely used in banking and fraud related papers [24]–[29]. This means that similar approaches can also be used in combating corruption.

In this work, the author has taken data analytical approach to detect corruption in public procurement in Estonia. The data has been collected from Open Tenders website [30]. The total data observations consist of 104 414 observations, where each observation was tagged manually for the corrupt cases. After thorough cleaning and processing of the data, supervised machine learning classification techniques, like Logistic Regression, Decision Trees, Random Forest and XGBoost were applied to estimate the models' capability to predict the likelihood of corruption. The models have shown an accuracy of 68,6 % up to 100% and ROC from 68,4% to 100%, which indicates that the data based approach taken is efficient in this topic.

The main goal of this thesis is to find out, whether using open data resources and data analytics it is possible to classify corruption in the public procurement processes and therefore suggest a suitable set of data to make the detection of corruption easier and quicker.

In order to achieve this, the author will conduct empirical analysis and will investigate, if it is possible to detect corruption using data science. The goal of the thesis will be achieved by using data analytics to build a model to estimate the probability of corruption. In order to validate the results from data analysis, it will be compared against findings in related work on the topic of corruption detection using data analysis.

In Estonia the establishment of e-government and digital public procurement register has helped to make the data available that is needed to assess the bids made during the process. Still it is plausible that the necessary data needed might not be accessible or recent enough to carry out this

research. So, the author will rely on open-data that is up-to-date and available. The author will use various features in the dataset. In addition, most important features will be generated based on feature ranking and applied to the models.

To the best of the authors' knowledge this is the first work that has analysed this dataset in machine learning approach to detect corruption in public procurement in Estonia. The value of this thesis is in finding a useful set of data features that can be used to make corruption detection faster and more efficient. The author believes that with sufficient data a lot can be done to help governmental agencies in the screening of businesses during public procurement. This will lead to an honest market from what the end users and the economy will benefit.

The thesis is constructed into five major sections. In the first the author will introduce the main topics of the thesis. Then the author will give an overall background to the problem of corruption and the effects it has on the economy. In the background section the author will open the main subjects to the reader, in order to enable continuous flow throughout the thesis. The third section of the thesis is dedicated for related work on topics relevant to this thesis. In the fourth section the author will cover the chosen research method and chosen data, as well as data processing. The final part shows the results of the research and compares it with the theoretical background and how it all bounds together. The thesis is ended with conclusions and recommendations for further research.

2. BACKGROUND

In this section, the author provides an overview of the main topics discussed over the course of this thesis. For example, this section discusses the reasons to eradicate corruption and how public procurement fits into the picture of corruption. Also, there is a short introduction to the machine learning methods, which are used for analysis.

2.1. Reasons to eradicate corruption

Corruption influences the flow of the economy and brings dishonest benefits to the people involved. Since it is a problem, there have been various studies conducted on the effects corruption has on the society [8], [10], [15], [31], [32]. It may be also seen as a natural part of the economy and the tool to get things done, but the negative effects of corruption do not justify this behaviour.

For example, Vadi et al. bring out aspects from their empirical study based on qualitative research conducted on company managers in Denmark and Estonia and how these people see the influence of corruption in their everyday work [15]. First, managers in both countries have found that loss of reputation is most probable, although they are more outspoken in Denmark - two thirds of the Danish managers believe that corruption will damage the company's reputation. Second, the Danes appear to be more concerned about the damage that such a case could do to their own career (63% compared to only 38% in Estonia) and finally around 50% of responses acknowledge possible financial losses in both countries.

Then there is the study by Eugeni and Tosato, who bring out aspects that also support the views of Salih on the effects of corruption [8], [33]:

- Corruption causes bureaucratic inefficiency by creating unnecessary tasks that hinder the workflow of the bureaucratic machinery;
- It influences the business and (local) investment climate by making the market an unfair playing field, where tenders with higher bribes will be chosen over price and quality;
- Civil and political rights are not followed as the act of corruption violates open market principles;

- Economic growth stifles due to corruption;
- Foreign direct investment goes down, since foreign investors are not interested in markets that unreasonably favour;
- International trade is influenced by choosing local companies due to bribery over foreign cooperation that might be beneficial in the long-term;
- Political legitimacy is put under pressure, since political campaigns are funded through private investors, which can include hidden agendas;
- Shadow economy grows.

As it can be seen in the list above, the effects of corruption are wide and influence various aspects of the economy. It stifles the economic growth and makes the business climate less attractive for investors. The consequences of corruption in public procurement can come through rent seeking and state capture. For example, rent seeking in public procurement can cause a lot of issues:

- Projects can have worse price to quality ratio, which depends on corruption rather that the actual bid [34];
- Rent seeking increases expenditures and the bigger the project the higher the bribe [35];
- Affects the allocation of resources as projects with easier corruption options are chosen [34];
- Laws can be influenced and made for corruption [5].

State capture is the second aspect of corruption in public procurement. It can be defined as bending the rules for one's good, which weakens the state's capacity or commitment to enhance security of property and contract rights, a weakening that further amplifies the incentives of other firms to influence on state activity by the help of corruption, representing a vicious cycle [1].

As previously brought out in the general issues, corruption in public procurement may also abate the international interest in both trade and foreign direct investment (FDI) [36]. In the paper Wei brings out that the reason why corruption is "higher taxing than tax itself" is that it is arbitrary and uncertain. This uncertainty effects public procurement processes, because in its essence it is a non-public auction site, where the best tender wins, but due to the possibility of corruption the

best quality-to-price ratio may not actually be the contractor. This may lead the economy into a never-ending cycle, ending up in circumstances under which corruption is the standard, where honesty is too costly, with a general disregard of law and a higher level of criminal activity.

There are clear reasons why corruption should become a thing of the past, but it still thrives with deep roots in the society. Luckily, illicit behaviour manifests itself in detectable data and now there are tools to put this data to proper use. If the same data points across projects are tracked over a certain time period, the behavioural patterns that indicate corruption risk can be recognized. This is where data analytics and machine learning become of value.

2.2. Public procurement, data and corruption

Public procurement refers to all of the offers that have been put up by the governmental bodies to satisfy the economy's needs, ranging from bids for bed sheets to construction works. This covers the whole process from start to end, which means that there are a lot of counterparties involved and the risk for corruption is high. Transparency International (TI) defines public procurement as "the acquisition by a government department or any government-owned institution of goods or services." [4] Public procurement corruption is one of the most common means of using one's positions for private gain and it can be seen as the most vulnerable to corruption. The World Bank makes a distinction between two separate categories of corruption that could be applied to public procurement [3]:

1) **State capture**, which refers to processes by individuals, groups, or organizations to influence public policy formation by illegally transferring private benefits to public officials (i.e., efforts by private actors to shape the institutional environment in which they operate); and
2) **Administrative corruption**, which refers to the use of the same type of corruption and bribes by the same actors to interfere with the proper implementation of laws, rules, and regulations.

Public procurement process is complex and exploiting various aspects of the system can cause corruption. For example, the corruption through hidden violations of ordinary procurement rules, which include [1]:

- Power of invitation – the officials can have the leverage to decide, who to invite on the tender;
- Short-listing/pre-qualification – higher weight on one particular criteria of evaluation met by the brining company only;
- The choice of technology;
- Confidential information – knowing beforehand the importance of various parameters of evaluation;
- Modification of the contract.

Although, there are also legitimate ways of deviating from honest public procurement, for example [1]:

- Justification for bilateral negotiation of governmental contracts;
- Major events and states of emergency – speed may help to deviate from the standard procedure;
- "Grease the wheel payments" – corruption without a specific case.

As it can be seen in the text above, there are various ways of manipulating tenders and bids during the public procurement process. This means that is vital to utilise data in the fight against corruption.

Most research that has been conducted on corruption on public procurement has been qualitative and based on various indexes that gave a perception of the level of corruption in a region or country. The first generation of corruption indexes was developed in order to measure general corruption levels in a country through surveys of the public and so-called 'expert witnesses'. An example of this is the Transparency International's Corruption Perceptions Index (CPI) [37]. The index does not understand the experience of corruption, rather it is based on perceptions and due to the fact that corruption has been seen as a sensitive topic during its development processes, it

could be possible that respondents were not be willing to discuss the actual situation. Due to these reasons, there are flaws that let down these indexes. For example, perceptions are often influenced by the media and one-off events like scandals, or by the culturally determined preconceptions of experts [38]. Another problem with the CPI is that few respondents have experience of the full range of public services, which might be affected by corruption and therefore, their views may not be accurate for services, where they have no personal experience. Public procurement is a classic example, where few members of the public will have direct experience and due to this it is one of the most corrupt sectors [7].

Thankfully, there is a lot of data generated in public procurement, which means that data can be seen as a key to understanding the possibilities of fighting corrupt behaviour and rooting out the old mind-set that in order to get things done, one needs to bribe [20]. This thesis presents available open data and empirical evidence on corruption in public procurement—an important issue that imposes political, economic, and environmental costs to societies around the world. Corruption is a phenomenon that involves many different aspects, and due to this it is difficult to assign a precise and comprehensive definition. However, the common aspect of most corruption definitions is the idea that a corrupt act refers to the misuse of entrusted power for private gain [3]. Bribery, favourism, and embezzlement are the most common examples. There are also more subtle or sometimes even legal examples in the public procurement processes, which include lobbying and patronage.

These are some of the common ways to lever the procurements in chosen directions. In order to uncover these tilted offers, this thesis will look into the e-procurement data in Estonia and will use previously established red flags or indicators of corruption. The red flags can be defined as indicators that show a potential risk of corruption. If a red flag is present, it does not automatically mean that an act has been corrupt, but it gives a indication that there might be a threat [20].

Recently there has been an increase of using data in the fight against corruption. For example, Fazekas et al. look in their work at the public procurement data in Hungary and provide new insights on the topic [19], [20]. There are also publications by the European Commission and

other authors that all have aspects in common [39]–[41]. The indicators developed can be divided into three stages of the process: submission, assessment and delivery phase of the procurement.

In the studies on red flags, the most important indicator that was brought out constantly was single bidding. Whether a tender only got a single bid, or during the assessment phase a single bid was left on the table. It was also mentioned that procurement data red flags go together with data on company registries, financial information, ownership information and network analyses that connect buyers and bidders. This is an aspect that should be sought after– building the understanding, what data should go together with public procurement and could support it. Looking at the related work, we can see, what have been the common grounds and build on our dataset by looking at relevant information that might prompt red flags. In order to avoid false flagging, it is important to employ and strategy of triangulation, so that it creates a good base for the accuracy of the actual indicator. Triangulation is the technique of using multiple methods of analysis on the same subject to confirm the validity of the results [42]. It is vital, because the existence of red flags does not mean that the act is automatically corrupt, but it creates a risk index that there is a possibility for that. Red flags can be seen as a tool in the fight against corruption, as indicators to detect, but also to prevent it.

2.3. Data and machine learning as tools against corruption

The available data on public procurement has been there for some time, but it never has been put to the use the way that it could be. Long-run data on corruption is a limited resource, but various cases and current media output suggest that corruption has been around in the societies over a long time period. The complexity and hidden nature of corruption makes it complicated to measure. So far, the data on corruption has usually been coming from either direct observation through law enforcement records and audit reports, or perception surveys like the Corruption Perception Index.

A solution to the data problem could be the use of electronic systems to track public procurement offers. For example, in Estonia the use of online systems to carry out tenders means that more and more data comes available [43]. The turn towards e-procurement is encouraged by the OECD and Korea's KONEPS system has been brought out as a good example [7]. This presents

an opportunity that has to be put to use. The public procurement system in Estonia went 100% electronic from 18.10.2018 [43]. The webpage accepts tenders, publishes offers and announces winners. It has been developed to make the process more honest and transparent. This means that there is plenty of data on public procurement, which can be sorted, analysed and published.

Using machine learning techniques, data can be analysed and the red flags of public tender processes could be developed and then the probability of corruption can be assessed. The author will now explain what are red flags in public procurement and how these can be used to scrape out corruption together with machine learning. Red flags can be also defined as corruption indicators in public procurement. These are developed metrics that can show, whether a tender could have an inclination towards corruption or not. There are studies that have collected and gathered the main red flag indicators, which are most common and seen the most often in corrupt cases [19]–[21], [39], [44]–[47].

Nowadays there is a lot of data available for analysis and the amount of data grows constantly [48]. In order to make use of this data, tools are needed. This is were data mining and machine learning come into play. Lying hidden in all this data is information, potentially useful information, that we rarely make explicit or take advantage of. Data mining is the process of finding common patterns in data with the assistance of methods like machine learning, statistics and artificial intelligence [48]. It is used to predict future values, instead of looking at predictive analysis that looks at historical data. There are benefits to using data mining in decision analysis. First, it automates the decision making, since it is running constantly and is not delayed by human judgment. Second, data mining is used to facilitate planning and will provide managers with reliable forecasts based on past trends and current conditions. This all comes together for a save in costs and effort. Data mining makes daily routine decisions easier to grasp and is useful for planning future plans with it's predictions [49].

In this thesis the focus will be on machine learning, but the analysis involves all of the aspects of data mining. Therefore, machine learning is a subset of data mining and it helps us understand patterns in the data. Machine learning is a mathematical computing algorithm that provides systems the ability to automatically learn and improve from experience without being explicitly programmed [48], [50]. Machine learning algorithms use datasets, where all instances are

represented by the same set of features, which may be continuous, categorical or binary. If instances are given with known labels (the corresponding correct outputs) then the learning is called supervised [51].

In this thesis the author will use supervised machine learning algorithms. Supervised machine learning algorithms can be used to apply past experiences to produce general hypotheses, which then make predictions about future instances. This means that a classifier is created to evaluate the model in terms of predictor values and this resulting classifier is used to assign certain values to the testing class that is unknown [51]. In this thesis there is a training set that is created to build a concise model that will generate future predictions on whether a public procurement process is corrupt or not. The analysis is conducted in R – a free software environment for statistical computing and graphics. In R the author will use machine learning techniques like Logistic Regression, Decision Tree, Random Forest and XGBoost. In order to evaluate the model the author will use data mining metrics for example like recall, precision, F1, ROC curves, AUC and accuracy.

The first machine learning approach the author uses in the thesis is Logistic Regression. Logistic regression (LR) is useful in problems, where fraud is binary and in this dataset and thesis, the instance of corruption is a binary feature. Previous publications in connected fields have estimated logistic models of fraudulent claims in insurance, food stamp programs, and so forth [28], [52]–[54]. It has been brought out that identifying fraud can be seen as a similar problem as real life decisions, like medical and epidemiological [27].

The second technique is Decision Tree. This is a predictive modelling approach that uses a decision tree to move from observational values to target values about an item [48], [55]. Each of the nodes in a decision tree represents a feature in an instance to be classified, and each branch within the tree is a value that the node can assume. Starting at the root node the instances are classified and sorted on the basis of their feature values [51].

Then there is random forest. This method combines the random subspace method with bagging to build an ensemble of decision trees. Random forests need only two easily set parameters, which makes them simple to use, and multiple researchers have found that it has excellent performance,

when it comes to the choice of the ensemble method for decision trees. Furthermore, they are also computationally efficient and robust to noise [56]. It has been found by various studies that random forests perform better, when comparing them to Support Vector Machines and other more modern techniques [57], [58]. The last supervised machine learning approach that the author uses to assess the data is XGBoost. XGBoost stands for extreme gradient boosting and the system is available as an open-source package. XGBoost is an ensemble learning method and it has been widely used after the initial launch in 2014. The key success factor for XGBoost has been the scalability in all scenarios and it is capability of running ten times faster than other methods [59].

The goodness of an algorithm can be shown through metrics. The first common machine learning metrics the author will explain are precision and recall. In order to understand the idea behind precision and recall, the basic principles of Type I and Type II error must be explained. These terms revolve around the rejection or non-rejection of the null hypothesis. The decision made by the classifier in a binary decision problem, as is the labeling of corruption, is represented in a structure known as a confusion matrix or contingency table. The used examples are labeled as negative or positive (Table 1). Type I error, also known as false positive, is the rejection of a true null hypothesis. Type II error, also known as false negative, is the non- rejection of a false null hypothesis. Precision and recall are therefore calculated as shown in Table 1 and it can be said, that precision represents the percentage of the results, which are relevant. Recall depicts the percentage of total relevant results that have been correctly classified by the algorithm [60].

To understand all of these metrics depicted in Table 1, it is important to cover the aspect of Confusion Matrix, which is also used in this thesis. Confusion Matrix is utilized in a classification problem, where the output variable is of two or more classes [61]. In it's essence it is simple, as it depicts the relationships between two dimensions: Actual and Predicted. It classifies the True Positives, True Negatives, False Negatives and False Positives. Confusion Matrix can be seen as a base for all other classification metrics discussed.

Table 1. The calculation of metrics, compiled by author; sources [60]–[62]

Metric	Calculation	Value to thesis
Precision	*True Positive / Actual Results* or *True Positive / (True Positive + False Positive)*	Gives an estimate how precise the model is in classifying procurements as corrupt
Recall	*True Positive /Predicted Results* or *True Positive / (True Positive + False Negative)*	Maximise Recall (minimise false negatives)
Accuracy	*(True Positive + True Negative) / Total*	Accuracy is good to use, when the target variable classes in the data are nearly balanced
F1	*2 * Precision * Recall / (Precision + Recall)*	Harmonic mean of both Precision and Recall
Specificity	*True Negative / (True Negative + False Positive)*	Exact opposite of Recall and Recall has to be as high as possible

ROC (Receiver Operating Characteristics) curve is used for classification problems at multiple thresholds settings. ROC is a probability curve and AUC (Area Under the Curve) represents the degree or measure of separation, as it says how well a model is able to distinguish between classes. When the AUC is higher, it means the model is better at predicting certain values at their set levels. In relation to the thesis, the higher the AUC, the better the model is at distinguishing between corrupt and non-corrupt cases. The ROC curve is plotted with True Positive Rate (TPR) against the False Positive Rate (FPR) where TPR is on y-axis and FPR is on the x-axis [60]. Other metrics used in this thesis are Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). These metrics show how far the predicted values were from the actual output [62].

In the next section of the thesis the author will give an overview of related work that is essential to this topic. The author will explain, how data mining has been used in corruption detection and what have been the biggest discoveries and trends.

3. RELATED WORK

In this section the author will cover the related work that has looked at corruption and its analysis in the past. The author will give an overview of various methods that have been used and the reasoning for using machine learning to predict corruption.

3.1. Corruption in public procurement

It is clearly presented in almost all of the research on corruption that it has effects on the economy and mostly negative. In the background the author gave a brief overview of the effects and causes of corruption, but in this section the author will take a deeper look into the related work on this topic.

3.1.1 Reasons for corruption

As it turns out people can find reasoning to being corruptive. This is indicated by the growing research on the topic and related news articles that are generated at regular intervals. Even if everyone knows that a corrupt act is not only morally wrong, but also has some other implications on the market or the whole economy - there are still people, who are willing to take this step.

The first justification people turn to is the ‘everyone-else-does-it’ argument, which normalizes unethical behaviour by suggesting a socially acceptable norm, even if it is not legal – an idea presented by Ashford and Anand [63], which means that bribery can be justified by the managers through arguing that competitors are also doing it. Second, Anand et al. present in another paper that circumstances beyond their control also work as a rationalisation strategy. This refers to peer pressure or the act of following an order by higher-ranking employee – ways of discrediting yourself from the responsibility [64]. Third and fourth rationalisations are that there are perceived positive effects coming or the act of corruption helps to prevent conflicting goals [65]. This reasoning strategy can take either a collective or an individual form. In the private companies, the managers could therefore reason that bribing took place, because the wellbeing of the company was at stake and/or that it helped keep a job.

Naturally there are not only personal reasons that promote corruption. Tanzi looks into the causes that increase corruption demand and supply [5]. The causes that promote corruption demand are: regulations and authorisations; certain characteristics of the tax systems; certain spending decisions; provision of goods and service at below-market prices. These are brought into the economy both by governmental decisions and individual moral. Tanzi also brings out the causes that promote corruption supply: the bureaucratic tradition; the level of public sector wages; the penalty systems; institutional controls; the transparency of laws; the examples set by the leadership [5].

Therefore there care clear-cut reasons for corrupt behaviour and since it is a criminal act, people try to rationalise their decisions. Still, corruption disrupts the flow of the economy and brings dishonest benefits to the people involved.

3.1.2 Effects of corruption

After the reasons for corruption it is suitable to understand the effects of this behaviour. The effects of corruption can vary amongst regions and countries. There have been multiple studies tackling this question and in this section the author will highlight aspects from these papers.

As previously brought out in the background section, Eugen and Tosato present ideas on the effects of corruption, which coincide with the views of Salih [8], [33]. These papers also shed light to new perceptions of effect like brain drain, fiscal debt and the value of human capital goes down. This means that there is the possiblity of losing valuable talent that do not feel comfortable in this dishonest market and are seeking a way out. Also, in the background section the author covered ideas presented by Vadi et al., who bring out aspects from their empirical study based on Denmark and Estonia [15]. Fortunately, these are not the only two important papers on the effects of corruption.

Transparency International has gathered knowledge from various publications and has compiled it into one table. The table depicts the impact of corruption, if it were to increase by one unit [66]. From Table 3 (refer to Appendix 1) it can be seen that corruption has effect on individuals, society and the economy. The aspects of life most influenced by corruption according to this study are - the income of the poor and student dropout rate. Furthermore, Li et al. find in their

study that corruption can have an inverted U-shaped effect on income distribution [67]. This means that if corruption increases, the income distribution lowers and if the corruption decreases, the income distribution grows. They also indicate that corruption influences the Gini differential across developing and industrial countries to a large extent. Their last takeaway is that corruption seems to hold back economic growth. As it can be read from findings on the table of impact of corruption (see Appendix 1), investments and quality of roads lose a lot from corruption. Similar findings can be seen in Hakkala et.al [31] work, where they use data from Swedish companies to assess corruption effect on foreign direct investment. They conclude that based on their data corruption lowers the likelyhood of investments to a specific country.

Corruption also affects social aspects, as again can be seen from Appendix 1, since it increases child mortality rate, school dropout rate and decreases the ratio of public health investments. These numbers are supported by Akcay [68], where he looks into the relationship between corruption and human development. In his work he compares three different corruption indexes to the human development index in 63 countries. The results clearly indicate that there is statistically significant relationship between corruption indexes and human development.

The litarature is clear that corruption affects everybody and everything, the effect of it may vary from country to country, but it exists in the society [5], [8], [32], [34], [69].

3.1.3 Corruption in public procurement

Public procurement amounted to 2 trillion euros expenditure in 2017 in the EU alone, which makes it a high value option for corruption [70]. There are further aspects, like the complexity of the process, the high number of stakeholders involved and the close interaction between public and private sector, that make public procurement one of the most vulnerable to corruption [7].

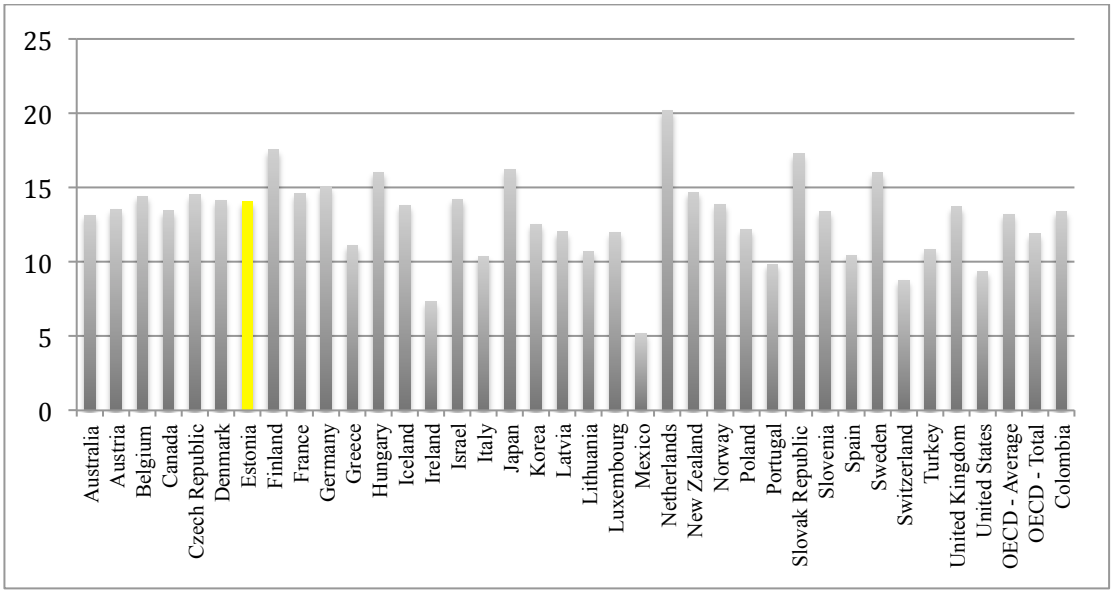

Figure 1. General government procurement as percentage of GDP in OECD countries in 2015, source: [71].

Corruption in public procurement can happen in both country or sub-country levels [7]. This is due to the fact that in lower hierarchies of the nation it may be easier to take advantage of the proximity of agencies and the weakness of the controlling counterparties.

Soreide gives a good overview of the most common ways of performing corruption in public procurement [1]. First, the public officials use their power to invite enterprises of their own choosing to the tender. This category is called the power of invitation. It has many ways of unravelling, either through postponing the release of tender information as long as possible or inviting companies, which are not suitable to win the procurement. Second, the approach of short-listing or pre-qualifying, where companies are evaluated based on some qualitative requirement like previous experience. The same aspect is brought out in other papers as well, for example Fazekas et.al [19], where they have taken a more data analytical approach, but also tender manipulation through previously set qualifications is seen as a red flag indicator of corruption. Thirdly Soreide brings out the choice of technology, where qualifications are specified in order to match the technological capabilities of the bribing company. Then there is the way of keeping confidential information that might be relevant to the decision. For example, execution time and cost of utilisation can be seen as qualitative requirements, rather than evaluating the tender on merely cost basis. Last is the modification of contract, which is highlighted in various papers and also in data analytical research about corruption. In their works

Fazekas et al. [25], Barett et al. [26] and Ferwerda et al. [44] all mark the modification of contract as a possible indicator of corruption.

This section emphasises the real threat that is corruption in public procurement. Due to the fact that it is a high monetary value sector and has technicalities that people can exploit, it can be said to be one of the most prone towards corruption.

3.2. Methods to prevent and detect corruption

3.2.1 Qualitative methods

As brought out in the background section, the most common analysis methods to describe corruption have been indexes, like the corruption perception index (CPI) presented by Transparency International [38]. The author will take a deeper look into the pros and cons of these methods, which have been analysed on various papers. Overall it can be said that indexes work, but they are biased as it is based on emotion and can be influenced by recent events. The first generation of corruption indicators thus sought to measure general corruption levels in a country through surveys of the public and so-called 'expert witnesses'. Transparency International's corruption perceptions index (CPI) is an example of such a measure. The index is based on perceptions rather than experience of corruption because, at the time it was introduced, corruption was thought to be such a sensitive topic that respondents would not be willing to discuss experiences.

As previously stated, there are however some flaws that these indexes carry. Ko and Samajdar evaluate in their paper, whether the CPIs can be trusted or not. In their work they come to conclude that more attention should be paid by the researches in order to minimize the impact of measurement error by conducting thorough data screening and robustness tests based on various data sources and methods [18]. The same perception of the lack of value in indexes is a key point in Corvino et al. International Monetary Fund research paper, where they attempted to address these limitations by leveraging news media coverage of corruption [72].

The design of surveys to measure corruption in public procurement can be viewed along two crosscutting dimensions. On one dimension, the sample of respondents can be drawn from either

the contracting authorities or the contractors, while the survey can seek to measure either perceived corruption or experiences of corruption [14], [38], [73]. As usual there are also contradicting viewpoints that suggest that the current system is working after all and the assumptions of the problems in perceptions or surveys is nothing but exaggerated. In the paper, Charron highlights the fact that the consistency between actual reported corruption, as well as citizen and expert perceptions of corruption, is remarkably high and such perceptions are swayed little by 'outside noise' [74].

There are new researches that want to make the work of qualitative studies on corruption more efficient. For example Ullah and Arthanari have developed a theoretical framework, which could be used to study corruption dynamics by means of system dynamics. The methodology they employed was a case study in which, semi structured interviews with key stakeholders such as: government ministries, donor agencies, judiciary, police departments, non-governmental organizations and the general public were conducted. On the basis of literature and social theory they developed three preliminary Causal Loop Diagram (CLD) models of corruption [14].

Most of these methods require a lot of workforce and take a long time to compile. This way the governmental agencies are always a step behind. This can be seen in the vast list of corruption prevention methods brought forward in the medical industry [75]. This being said, the qualitative research has some valid points in eradicating the corrupt mind-set from the beginning. For example Vadi et al. present ideas to prevent corruption in the first place. These thoughts are valid and should be taken into account. They brought out four recommendations that were based on their study on Estonian and Danish businesses [15]:

- Recommendation 1: Based on discussions between middle-office executives and business, a one-page manual for how to recognize business corruption should be developed and made available for all employees.
- Recommendation 2: Develop a toolkit that helps the firms identify areas, where integrity may be at risk in their processes.
- Recommendation 3: Taking in the examples established in other Nordic countries, like Finland, it could be beneficial to get access to background check during the hiring process.

- Recommendation 4: Companies should make reporting available through multiple channels and it should be specified, encouraged and known as a solution.

The value of qualitative research is there, as it gives in insight to the reasons of corrupt acts and helps root out corruption from the start. In order to prevent corruption it is necessary to understand it and how to help companies or employees deal with corruption.

3.2.2 Data analytics, corruption and machine learning

It appears that since 2000 the literature on corruption has been increasing exponentially and plenty of analysis has been done on rooting out corruption. There are several reviews of the economic literature on corruption in public procurement that are provided, for instance by Søreide [14], Rose-Ackerman [29] and Rose-Ackerman and Palifka [30]. This literature has been mainly dealing with the institutional environments throughout which corruption prospers (e.g. Goel and Nelson [31]–[33]; Mungiu-Pippidi et al. [34]–[36]; Acemoglu and Robinson [37]) on the incentives of officers to demand and take bribes (e.g. Campos and Pradhans [38]; Rose-Ackerman and Søreide [39]; Søreide and Williams [40]) and on the negative welfare consequences of corruption (e.g. Mauro [41]; Wei dynasty [24]; Vinod [42]). This literature reflects the broader efforts of the international community to carry awareness of the harmful effects of corruption and to grasp the corruption development. To highlight, the recommendations converge.

One of these recommendations is that the use of indicators of corruption to separate corrupt from non-corrupt public procurements can be beneficial, when separating it between sectors and countries. This logic continues on. The reason why indicators are beneficial is that that corrupt activities need certain types of economic behaviour (e.g. single bids, very rich public officials, public procurement contracts that need rewriting or are appealed) and this behaviour leaves traces [29]. Consequently, corruption indicators are accumulations of traces, which will purpose to the presence of corrupt activities. Consequently, they are primarily geared towards serving to practitioners, investigators and policy manufacturers in estimating the prospect of corruption of an exact procurement case and lay the groundwork for the inspiration of a replacement evidence-based approach to fighting corruption.

The literature on corruption indicators has been systemically targeted on analysing atypical samples, and consequently may be seen to suffer from a general choice bias drawback. The European Commission conducted an analysis on the red flags of fraud, which included corruption, also in public procurement, built on the known data of corrupt or suspected cases, which had been brought out in the annual reports against fraud [30]. This report supported the list of indicators for corruption in public procurement, which had been composed by the Association of Certified Fraud Examiners. The overlapping of results is not surprising, since both lists are composed on the data of acknowledged fraud case investigations [31]. Similarly, in 2011, the European Anti-Fraud Office revealed an inventory of structural indicators of fraud in countries and the research was supported by the fact-finding experience the OLAF had accumulated over the years. It created use of anonymous cases that are investigated by the OLAF, where elements of fraud had been detected. Knowing the particularities of the cases regarding the investigation and having a deeper understanding of what has gone wrong in that specific case, OLAF was able to conduct a thorough analysis and revealed a number of very important fraud indicators [32].

Moreover, the OECD and Ware, Moss and Campos put forward some of the most common types of corruption (e.g. kickbacks, bid rigging and use of shell companies) and then presented a selection of indicators, which could be used to confirm corruption [33], [34]. What is more, in 2010, the World Bank issued a guide on the top 10 most common indicators of fraud and corruption in procurement for bank-supported projects. This list was, once again, was confirmed by anecdotal proof provided by investigated cases of fraud and corruption within the public and also the private sector. Extra sources of information are theoretical models and expert opinions on the probable symptoms of corruption [29]. Finally, a study of Transparency International on corruption within the sphere of public procurement in Asian countries like Indonesia, Malaysia and Pakistan identified and clustered the indicators of corruption on the general public procurement cycle [12]. Similarly, Ware, Moss and Campos distinguished the subsequent procurement stages and on that corruption indicators may be classified as follows: first, project identification and design; then advertising, prequalification, bid preparation and submission; third bid analysis, post qualification and award of contract; and finally contract performance, administration and superintendence [34]. The corruption indicators that are going to be used in this thesis and are described in the subsequent sections will be organized on these lines of the

overall public procurement technique to assess with higher accuracy the risks of corruption in every stage of the acquisition method.

In the literature on corruption indicators there are common ideas that come through. First is that, the indicators can be divided into three stages of the process: submission, assessment and delivery phase. Secondly, there is an understanding of the most common indicators and the author will highlight the constantly brought out. The most important indicator that had many mentions – single bidding. Whether a tender only got a single bid, or during the assessment phase a single bid was left on the table. The most common indicators of corruption in the literature the author will now highlight and divide them into phases [19]–[21], [39]–[41], [76].

- In the submission phase of public procurement:
    - Single bidder contract;
    - Call for tender not published in official journal;
    - Procedure type;
    - Relative length of eligibility criteria;
    - Length of submission;
    - Relative price of tender documentation;
    - Call or tenders modification;
- In the assessment phase the common indicators were:
    - Exclusion of all but one bid;
    - Weight of non-price evaluation criteria;
    - Annulled procedure re-launched subsequently;
    - Length of decision period;
- In the delivery phase of the public procurement process:
    - Contract modification;
    - Contract lengthening;
    - Contract value increase.

Also, there is research that does not look at machine learning, but using panel data, which presents measurements over time, like Frechette [77] and Elbahnasawy with Revier [78]. The latter estimate in their paper a random effects model that includes the effects of corruption

determinants that change over time and also those that are time-invariant. Elbahnasawy and Revier use a larger panel dataset and a comprehensive set of corruption determinants [78].

As this thesis uses data mining and machine learning, it is vital to seek out related literature on this particular topic. One of the most influential works on data mining has been done by the DIGIWHIST project [19], [79]. They use available open data on public procurement and crawl for data on websites that manage tender offers. This means that they have a lot of data and have established ground indicators to say, whether the business environment in a country is prone to corruption or not. In their work they have brought out 14 new indicators that are classified under: integrity, administrative and transparency. These show how the region is doing in each of these fields. The author of the thesis will use the same data, only on Estonia, which has been brought out in the DIGIWHIST project and is published on Open Tenders website. In order to provide deeper analysis the author will improve the dataset by looking at confirmed corrupt cases through court documents and will suggest new features as indicators that might improve the future projections.

Furthermore, there has been an increase in looking towards data in fighting against corruption and the people at DIGIWHIST are not the only ones. There are many studies that have used data to discover unknown patterns in corruption. For example, Fazekas [19]–[21] looks in their work at the public procurement data in Hungary. There are also publications by the European Commission and other authors that all have aspects in common. Carvalho, Ladeira et. al [80] present a case study of machine learning applied to measure the risk of corruption of civil servants using political party affiliation data. Similarly Olsen, Reynolds and Koltsev [81] discuss how data analytics can be put to use next to qualitative statistics. López-Iturriaga and Sanz [82] use neural networks to predict public corruption on the basis of Spanish provinces. In their work they use a neural network approach to develop an early warning system, specifically self-organizing maps, which is used for predicting public corruption based on economic and political factors. In their findings they present that the aspects that seem to induce public corruption are: the taxation of real estate, economic growth, the increase in real estate prices, the growing number of deposit institutions and nonfinancial firms, and the same political party remaining in power for long periods [82].

Corvino et al. [72] question in their research the validity of corruption perception indexes. In order to overcome these limitations they provide big data on cross-country news flow indices of corruption and anti-corruption. They achieve this by running country-specific search algorithms for over more than 665 million international news articles. They find in their research that news flow indices of corruption shocks appear to negatively impact both financial (e.g., stock market returns and yield spreads) and real variables (e.g., growth), although there is some country heterogeneity [72]. Sales [22] uses credit scoring to predict fulfillment of public procurement tender offers in Brazil.

Published works on the importance of data analytics in corruption detection has been increasing. The main indicators have been developed and now it is necessary to put this research into models and use it for future predictions.

4.METHODOLOGY

In the following section the author will cover the methodology of the thesis. The author will go through the following topics: the goal of the thesis, research questions, data and the implemented research method.

4.1. Research goals

The main goal of this thesis is to find out, whether using open data resources and data analytics it is possible to predict corruption in the public procurement processes and therefore suggest a suitable set of data to make the detection of corruption easier and quicker.

In order to achieve this, the author will conduct empirical analysis and will investigate, if it is possible to detect corruption using data analytics. The goal of the thesis will be achieved by building a model to estimate the probability of corruption. In order to validate the results from data analysis, it will be compared against the findings in published related work.

In order to achieve the set goal of the thesis, the author has established several research questions:

- Relying on the theoretical findings, explain the essence of corruption in public procurement;
- Give an overview of the various implications of corruption in public procurement;
- Present an overview of the state of data in open source data in regards to public procurement in Estonia;
- Develop an applicable solution to analyse the open source data, to assess and visualize the results;
- Validate the results from the analysis by comparing them to the related work;
- Suggest ways to improve the state of public procurement corruption in Estonia.

The established goals are highlighting the available option to use open source data to root out corruption in the public procurement processes. There is lots of data available, which could be

put to use with the help of proper data analysis tools and methods. Using these solutions it could be possible to find indicators of corruption, which could be implied to the prevention of these acts.

4.2. Flow of the thesis

The author structures the process of this thesis into a workflow diagram. The purpose of this is to develop the understanding of the analysis process and actions taken to achieve the goal of the thesis. The steps of the process are as follows:

- Problem statement;
- The collection of proper data;
- Data preparation;
- Machine learning model creation;
- Model training;
- Evaluation;
- Hyperparameter tuning;
- Prediction;
- Results analysis and discussion.

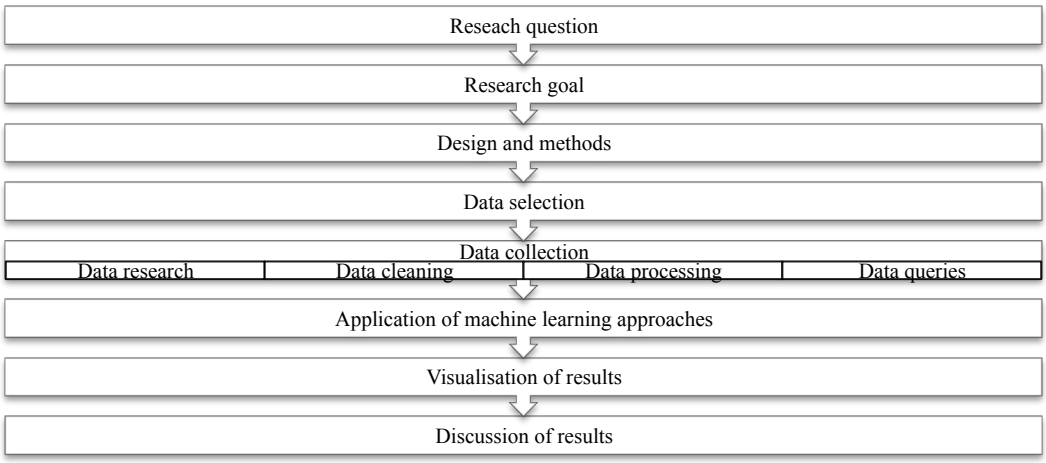

Figure 2. Workflow of the thesis, compiled by author

In the figure (see Figure 2) above, the main tasks have been brought out to highlight the flow of the thesis. The whole process started with identifying the research problem and question. Then it was necessary to set a goal to resolve this issue and in order to do that a plan was designed. After

that, the plan was implemented through various steps like data processing, application of machine learning approaches and the assessment of results. Based on these findings the author discussed possible future research possibilities and how this thesis could be useful in the fight against corruption in public procurement.

4.3. Data description

In the thesis the author used quantitative analysis method. Based on this suitable data collection and analysis methods were chosen. The quantitative analysis methods focus on measuring the objects under observations through datasets [83]. Quantitative methods in corruption research allow us to grasp a wider narrative of what the connections between the factors of interest looks like, understand patterns by looking at large data, and make conclusions on the observable patterns. The data analysed in this thesis gives results that might be applicable to other cases and could be generalizable under certain circumstances to outside the Estonian public procurement data. Quantitative research is a method to test objective theories through the connections of model variables [84].

In order to achieve the main goal of the thesis author used nonreactive data to perform the statistical analysis and data collected from DIGIWHIST project through the Open Tender webpage [19]. Nonreactive data can be characterized by the fact that the subjects under investigation are not aware of that and therefore the process of data collection does not influence the behaviour and thanks to this the open source data is not altered. Nonreactive data collection is said to be an unseen, hidden, natural or non-intrusive action that takes place in the natural process of data gathering. Nonreactive data can be classified into three categories: physical evidence or data, observational and archived data that gather in an non-intrusive way [85]. Combining nonreactive with other kinds of data collected to pursue the same goal is a way to increase the validity of the results. One of the positive aspects of nonreactive data is the fact that their online easy access gives the opportunity to collect and analyse large amounts of data at once. Further supporting aspects of nonreactive data can be their certainty, trustworthiness, cheapness and longevity. On the other hand the lack of control and background information can be a negative side of nonreactive data [85]. Nonreactive data collection methods can be classified into two sections – usual observation and automated. The usual observation in its essence is data

collection without any alterations by the researcher and the data used has already been archived. In the automated methods the data collection happens automatically and in order to provide objectivity, it already gathers large amount of data at fine detail [85].

In this thesis the author uses automated data collected by the DIGIWHIST project team that uses the DIGIWHIST data processing system. This system detects, when an act of public procurement process starts, then it triggers a chain of procedures that leads to the incorporation of the data included in this publication into a final database. In order to achieve this a set of so-called source crawlers were implemented by the project. Each crawler draws upon the combined knowledge of DIGIWHIST public procurement experts and developers, gained through a detailed inspection of the source [79]. The author utilizes the work by DIGIWHIST project and uses gathered data about Estonia.

In this thesis the author uses supervised machine-learning methods to analyse the data about public procurement processes in Estonia, for that a ground truth of the dataset must be established. In machine learning, the term ground truth refers to the accuracy of the training set's classification for supervised learning techniques. Ground truth is used in statistical models to prove or disprove research hypotheses. The term "ground truth" refers to the process of gathering the proper objective (provable) data for this test, it can be compared with gold standard. The data contains variables that describe the public tenders, but it does not say, which of them have been corrupt. In order to have a sense of direction, the author collaborated with various governmental agencies to obtain knowledge of corrupt public tenders that have taken place in Estonia over the analysed time period. Working through hundreds of court documents, the author was able to compile a list of confirmed corruption cases and establish a ground truth for running supervised machine learning [86].

The initial data set comes from the project DIGIWHIST and has been developed by others. In this thesis the author decided to go for this data set, because it coincided with the same data that the author planned to use from the beginning. This is a unique database describing public procurements and buyers' practices across the whole of Europe. It contains detailed information (up to 119 variables) covering the whole life cycle for both above- and below-threshold tenders. It combines all available publications related to a given tender, links them to company and budget

databases using complex algorithms and applies various business rules to create one single representative for each tender and subject. [79]

In the case of Estonia the data source is Public Procurement Registry (PPR) webpage that is an e-governance site for public procurements in Estonia [43]. The data was obtained by DIGIWHIST project and then downloaded by the author. The data comes from an HTML web portal, which contains a search form. The search form allows to search by publication data and each notice appears in a search results for each day, when it was published or modified. The result HTML page contains paged list of publication, as each page allows the user to click on a next page link. If this link had been disabled the end of a list is reached and there is nothing further for the program to crawl. The maximum amount number of search results is 500, which the crawler developed by DIGIWHIST did not reach on a daily basis [79].

From all the data published the DIGIWHIST project processed all publications from the source by type [79]:

- First there is contract notice, which gives general information about the initial publication made;
- Secondly, contract award, which gives information about the initial offers made;
- Third, prior information notice, which gives data about various altrecations to the procurement;
- Last, contract implementation, which gives data about the actual results of the procurement process.

The grouping of all tender publications describing the same public procurement together has been based on the tender ID, which the author used to link the court documents to the dataset to established the ground truth for supervised learning models. A backup strategy was also developed and it was based on matches of URLs of related publications, which was useful on certain occasions.

The data set is from 2009 till May 2019. The years prior to that have not been included, since the format differs and many variables are missing. As the data is constantly updated, it means that the results of this thesis can be implicated in the future. The data consists of 212225 observations

among 119 variables. This is the initial data that was obtained from OpenTenders website. The author added manually an output variable - tender_isCorrupt - in order to perform supervised machine learning.

The actual data used in this thesis ranges from year 2009 till 2015, which consists of 104414 observations of 119 input variables and one output variable – tender_isCorrupt. The choice of this year range has been made, due to the fact that on these years there have been confirmed court cases, which have been added to the dataset by the author. Corrupt acts taking place now will be taken to court in two to three years, so that is why there is the delay in data and the latest procurements are not included in the analysis. The author was able to match 50 observations with court cases from the data [86]. The information about corrupt cases was obtained in collaboration with governmental agencies. These observations present only the cases that were publicly available and in which the procurement numbers were not hidden. Due to this fact the author was not able to link all court cases that were found to the dataset, as the author could not verify the procurement numbers one hundred per cent.

The features of the dataset can be classified as follows:

- 73 features about the initial tender offer;
- 9 features about the buyer;
- 21 features about the lot;
- 9 features about the bid;
- 7 features about the bidder.

In order to get a better understanding of the data, it is important to visualise descriptive statistics (refer to Appendix 4). Therefore, the author will bring out key statistics in the initial data. In the initial data there are 55 character, 35 integer, 21 logical and 9 numeric features. There are 21 features that are totally empty, which include for example features like tender award deadline and tender contract signature date. This is unfortunate, because these could be beneficial to the analysis of the thesis, as the features could say something further about the time aspect of the tender. After removing features with unique ID, empty and duplicate columns, the number of features left in the dataset was 88. So, there are 88 variables that have been derived from the

initial data, which do not have empty values and duplicated. This leaves 52 character, 31 integer and 5 numeric features.

4.4. Research approaches

To achieve the main goal of this thesis, the author uses data mining and machine learning to detect corruption within public procurement in Estonia. Data mining can be defined as the process of discovering common patterns in data. Machine learning is a form of artificial intelligence that enables computers to learn without being explicitly programmed [48]. In this thesis the author will use supervised machine learning to detect corruption. In supervised learning the labels have been manually assigned, so it can predicted, whether an tender will be corrupt or not [87]. It's especially good at recognizing patterns and spotting anomalies in data. Fraudulent and legitimate transactions have different characteristics. Algorithms are created based on those differences to predict the likelihood of fraud. In this thesis, the techniques that have been prioritized allow to explain how decisions are made. The goal of this thesis is to know why and how corruption is happening, so the choice has been made on the following approaches and they have been blended together to provide a probabilistic result that will indicate corruption.

Firstly, there is logistic regression. This statistical technique uses an algorithm to compare already existing procurement processes in order to predict the likelihood of whether a new tender will be either corrupted or not. Secondly, there is decision tree. Decision trees are used in order to automate the creation of rules for classification tasks. They are trained using examples of corruption that have been established in the initial research of the thesis by going through public court records. The tree can be used to understand why the next tender could be corrupt based on following the list of rules triggered by a certain tender offer [48].

Thirdly, there is random forest. This technique uses multiple decision trees to improve classification performance. It is useful in order to allow the smoothing of the error, which might exist in a single tree, increasing our overall performance and accuracy while maintaining our ability to explain the results to users [26]. Lastly, the author will use XGBoost. XGBoost attempts to mimic how the human brain learns, particularly in pattern recognition. They are

trained on legitimate patterns and are therefore adept at flagging fraudulent ones. It complements other techniques and improves with exposure to data [28], [87].

Machine learning is not a silver bullet for corruption detection and prevention. For machine learning models to become accurate, it takes a significant amount of data. For some datasets, it is useful to apply a basic set of initial rules and allow the models to 'warm up' with more data.

Machines have become much better at dealing with large datasets than we are, although there are limitations to machine learning. They can recognize thousands of features from tender's initial publication to the end of the procurement. Machine learning can see deep into the data and make concrete predictions. In large datasets, machine-learning approaches get even better. [23], [48], [52], [88]

4.5. Data cleaning and processing

Data cleaning is vital, when it comes to the accuracy of predictions. Understanding and pre-processing the data into a state, where it is possible to analyse and predict accurately is a thorough process [89]. Filtering and cleaning the data of the parts that are not beneficial towards the goal of the analysis is important, as it helps to train the model easily and will give more accurate results [90], [91].

In this thesis the author started the data pre-processing with removing missing values by taking out totally empty columns. Then, the author removed unique variables, like row number, tender id, title etc., as these variables do not give any value to the modelling. However bidder IDs, physical locations, addresses are kept as factorials because a possible pattern might be revealed in a certain geographical location, bidder or otherwise 'unique' factorial. Secondly, the author removed all columns from analysis with less than 5 per cent filled data (all string variables, leaving us with 65 variables of which 40 are numerical). Thirdly, variables, which are dates, were removed, since the dataset already consists of variables like tender_estimatedDurationInMonths and tender_estimatedDurationInDays. So, it was not necessary to duplicate these. Likewise to variables with dates, the author removed variables for all prices brought out separate in euros, for

example tender_priceEUR. The reason for that is that glimpsing at the data, it is clear the data is duplicated across the dataset.

After the cleaning process, the NAs were counted. The highest number of NAs was in variable tender_estimatedPrice with 97 787 NAs, which means that only 6,35 % of data was filled. As this amount of data is not reasonable to impute, because the number of NAs is higher than the amount filled data, the author decided to remove variables, which have more than seventy thousand missing values. This left a dataset of 104 141 observations for 21 input variables and 1 output variable. Then the data was imputed using MICE package, where the missing values are replaced with predicted values. For example, if there is a number of variables X1, X2 to Xn. Then lets assume that X1 has missing values, so it will be computed on variables X2 to Xn. The NA values of X1 will be replaced by predictive values gathered.

Then after imputing the dataset was balanced. The author used ROSE package and combination of over- and under-sampling, respectively. The reason is that using both over- and under-sampling increases the recall of the prediction model and this is the most important metric in this thesis, as it is necessary to minimize the number of false negatives [92]. Before starting with the training of the model, it is important to make sure that all of the input variables are either numeric or factors [93]. This is necessary, because the approaches used in this thesis require it [51]. The last step in the data pre-processing before modelling was scaling of the dataset, also known as data normalisation. Data normalisation is important, because data can consist of observations that vary highly in magnitudes. Since most of machine learning approaches use Eucledian distance between two data points for calculation, non-scaled data is a problem [94], [95].

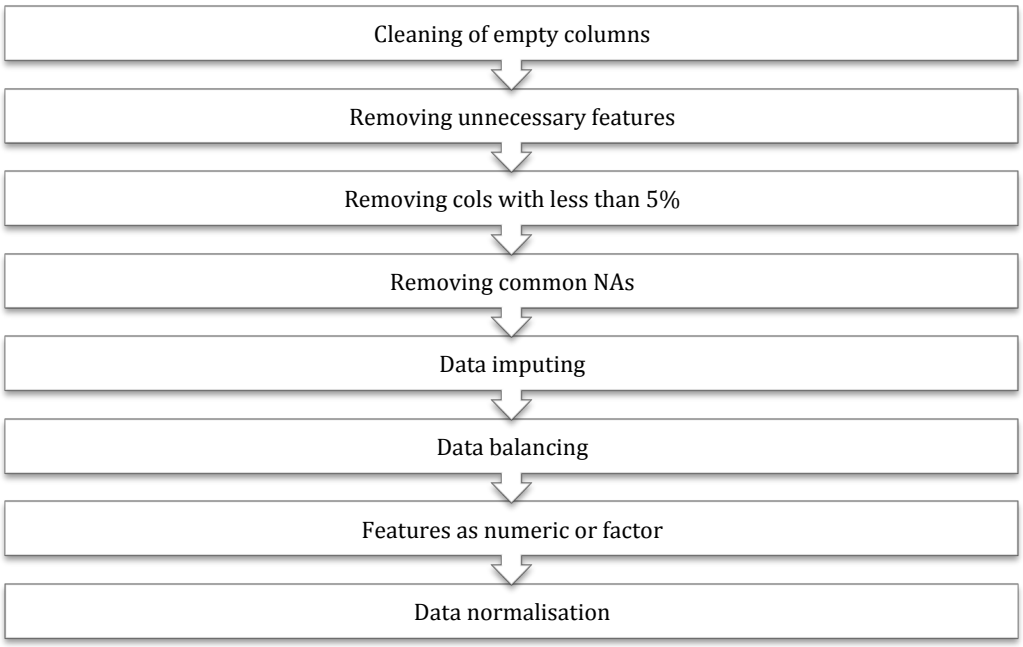

Figure 3. Data pre-processing steps, compiled by author

In the figure above the steps are presented graphically. After these steps the author was able to start with the modelling of the machine learning approaches.

4.6. Model building

In machine learning it is necessary to train the model for prediction. As described in section 4.4 the author uses logistic regression, decision tree, random forest and XGBoost machine learning approaches to estimate the probability of corruption in public procurement in Estonia. The reasoning behind the choice of approaches is that aforementioned techniques are classic approaches that have been found to give good results and have quality [26], [55], [56], [59], [48]. For example, in their study Lessman et. al benchmark commonly-known approaches against newer methods and it is found that these approaches gave very good performances, which can note the fact that most credit scoring data sets are only weakly non-linear. Looking at the results of their study, it is clear that many classification techniques have shown performances, which are quite similar to each other [88]. These methods will be assessed separetely and compared to one another, in order to give a better overview of the validity of the results.

The author uses cross-validation in the model building process, specifically stratified k-fold cross-validation. Kohavi expresses in his study of cross-validation for accuracy estimation that stratifiied k-fold provides results with the smallest bias and variance, because using this method makes it possible to predict on all of the data, rather than doing a simple spilt of 80-20 [96]. Also, by using cross-validation it will be possible to get more metrics and understand more about the model and the data. For the estimation of predictions the author uses metrics like accuracy, precision, recall, F1, Confusion Matrix, Mean Absolute Error (MAE), Mean Square Error (MSE), and ROC (refer to section 2.3 and Table 1). These metrics were chosen, as they suit the analysis of predicting corruption in public procurement [61], [62].

The coding of the models (refer to Appendix 2) was based on authors previous knowledge obtained in courses regarding R, but the final decisions were based on published materials. For example, the choice of number of trees built in Random Forest model was selected at 500 and the mtry was set at 5, which is the square root of total features available. These decisions were based on studies conducted on Random Forest [97], [98]. The same applies for the choice of threshold. The threshold set in the code is 0.5 and the reason is that this way it makes the probabilities into classifications [99]. Guidelines were also used to build Decision Tree and XGBoost models [59], [100]. In order to assess the importance of features the author also built models with only top ten most important features. This gives indication on the valuable features that should become the focus points of future analysis.

In the upcoming section the author will discuss the final data that was used for analysis and will go over the main results.

5. RESULTS

In this section the author will cover the main results of the analysis and will present possible suggestions for imporvement in the future on this particular topic.

5.1. Final data

The initial data of 104414 observations, 119 input and one output variable were cleaned from empty columns, values. Then it was pre-processed for analysis. After the application of these methods the data that the machine learning models were built on consisted of:

- 104 414 observations;
- 21 input variables;
- 1 output variable.

The data was imputed and scaled. In order to understand the final dataset, the author will bring out key descriptive statistics (refer to Appendix 3).

5.2. Visualisations of results

The data was analysed through R software and after cleaning, pre-processing, model building the results were gathered and will be presented in this section. The following Table 2 presents the main results of the machine learning approaches.

Table 2. Results of the analysis, compiled by author

Metrics	CONFUSION MATRIX			OTHER METRICS						
Models	Accuracy	Sensitivity	Specificity	R2	MAE	RMSE	Recall	Precision	F1	AUC
GLM	0.7457	0.7713	0.7195	0.2417001	0.2543217	0.5043032	0.7713	0.7370	0.7537	0.745
GLM_CV*	0.7457	0.7713	0.7195	0.2417001	0.2543217	0.5043032	0.7713	0.7370	0.7537	0.745

GLM _F**	0.7457	0.7713	0.7195	0.2417001	0.2543217	0.5043032	0.7713	0.7370	0.7537	0.745
GLM CV_F	0.6866	0.9909	0.3765	0.2182904	0.3134128	0.5598329	0.9908	0.6182	0.7614	0.684
DT	0.9675	0.9357	1.0000	0.878112	0.0324666	0.1801849	0.9356	1	0.9667	0.985
DT_CV	0.7544	0.9201	0.5855	0.2889138	0.2456065	0.495587	0.9201	0.6934	0.7908	0.768
DT_F	0.949	0.8989	1.0000	0.8148764	0.05104631	0.2259343	0.8988	1	0.9467	0.981
DT_CV_F	0.6912	1.0000	0.3765	0.2335728	0.3088158	0.555712	1	0.6204	0.7657	0.751
RF	**1**	**1**	**1**	**1**	**0**	**0**	**1**	**1**	**1**	**1**
RF_F	**1**	**1**	**1**	**1**	**0**	**0**	**1**	**1**	**1**	**1**
XGB	0.9997	0.9994	1.0000	-	-	-	0.9997	1	0.9994	1
XGB_F	0.9994	0.9988	1.0000	-	-	-	0.9987	1	0.99938	1

* CV refers to Cross-Validation
** F refers to model with only top10 features

Table 2 contains abbreviations of machine learning models:

- GLM refers to Logistic Regression;
- DT refers to Decision Tree;
- RF refers to Random Forest;
- And XGB refers to XGBoost.

As it can be seen from Table 2, the author has not used Cross-Validation for Random Forest and XGBoost. The reason behind not using Cross-Validation in Random Forest, is that multiple

bagging in the process of training Random Forest should prevent the model from over-fitting [56]. For XGBoost the initial model already uses Cross-Validation (refer to Appendix 2).

The results show that the best model for public procurement corruption prediction in Estonia is Random Forest, as the model has accuracy rating of 1. Great results apply to the other metrics as well, which makes Random Forest the best model in current dataset for predicting corruption in public procurement in Estonia. However, it seems that this model might be overfitting the data. This means that the results of the analysis is too close or in this case exact to the data, and therefore might fail to predict future observations reliably. Even though, research has shown that Random Forest models do not overfit, the author has reservations and this would require further analysis [56], [97].

The effect of cross-validation is also clear, but since the models are still predicting well, it can be said that the predictions are valid. The results vary for both Logistic Regression and Decision Tree. Cross-validation provides certainty of stability that this data is useful in predicting future values on chosen models [96]. The results of Logistic Regression (LR) and Decision Tree (DT) models are also good, with accuracy ratings of 0.7457 for LR and 0.9675 for DT. In the case of LR cross-validation does not alter the outcome, but for DT the accuracy decreases to 0.7544, which is also not a bad result.

The worst results come from Logistic Regression model with cross-validation and only top ten important metrics. In this case the accuracy is just 0.6866. On the other hand, the model comes with high recall of 0.9908 and lower precision of 0.6182. It was established earlier in the thesis that achieving high recall in the predictions is important, as it minimizes false negatives. This means that there is lower possibility of not discovering actually corrupt cases. In Figure 4 the AUC (Area Under the Curve) of receiver operating characteristic curve (ROC) is presented. This plot shows the ability of a binary classifier system to predict at various threshold settings [60]. The best in this metric are Random Forest and XGBoost with AUC rating of 1. This means that the model is very good at distinguishing between positive and negative classes.

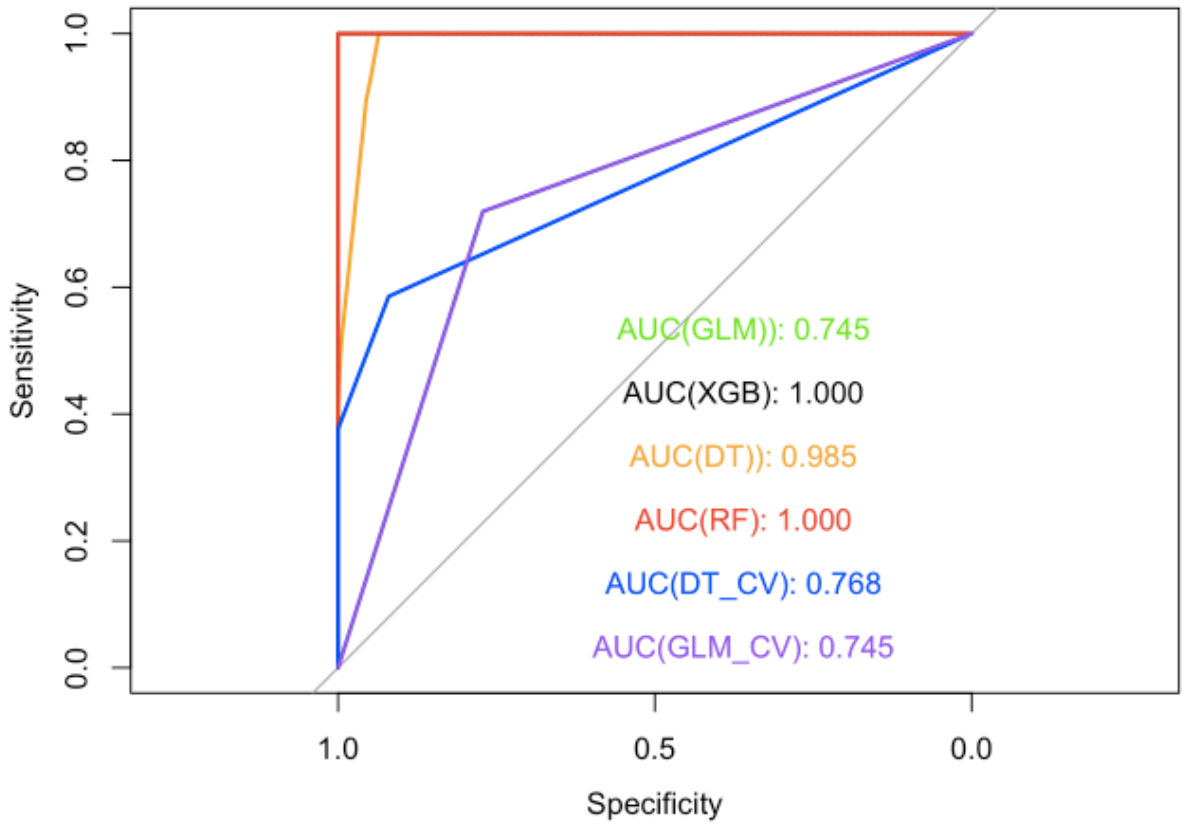

Figure 4. AUC for all models with all features, compiled by author, source: Appendix 2

In order to understand the value of various features, the author plotted their ranking according to importance. These results can be seen in Figure 5, where it depicts that tender_economicRequirements_length has the highest importance with a rating of over 0.2 and the least important feature is tender_indicator_ INTEGRITY_DECISION PERIOD.

The author also compiled models based on the ten most important input features:

- tender_finalPrice;
- bid_price;
- lot_bidsCount;
- tender_economicRequirements_length;
- tender_description_length;
- tender_personalRequirements_length;

- tender_technicalRequirements_length;
- tender_lots_count;
- tender_indicator_INTEGRITY_SINGLE_BID;
- tender_indicator_INTEGRITY_PROCEDURE_TYPE.

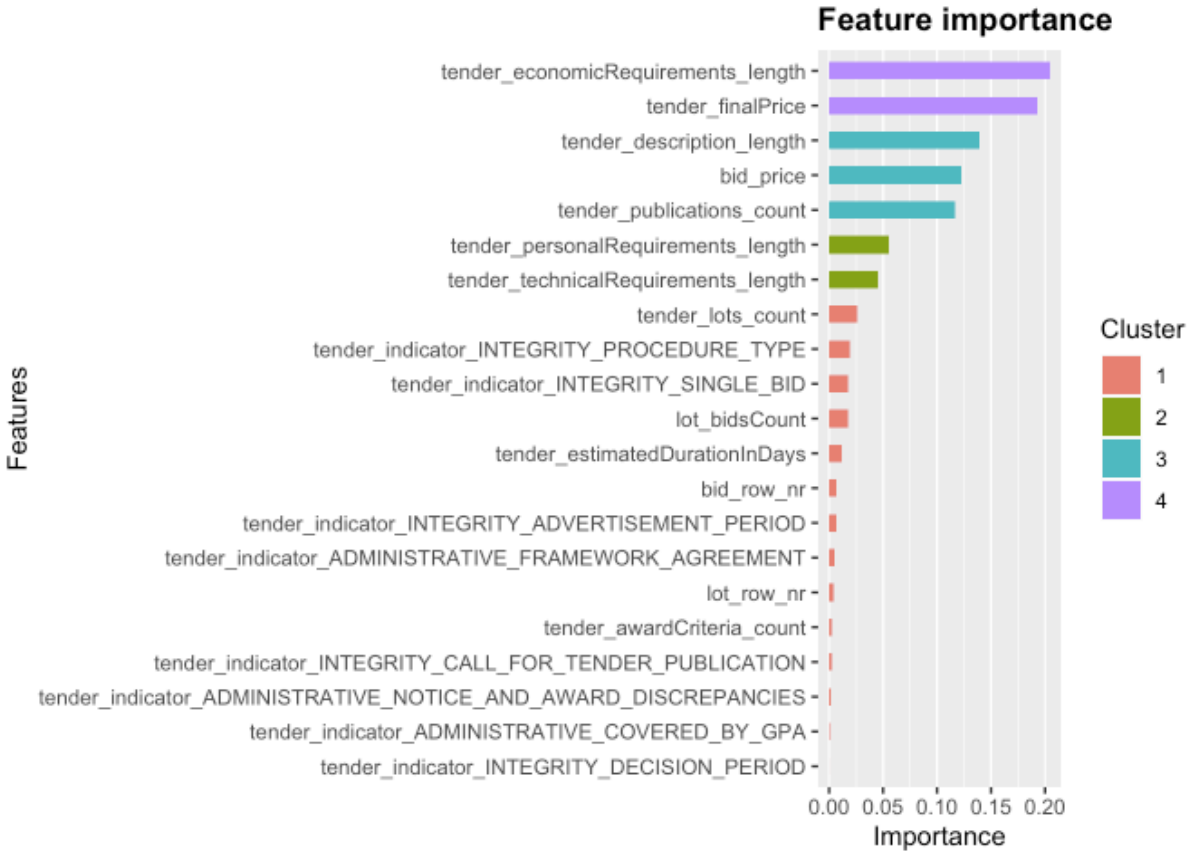

Figure 5. Features ranked by importance, compiled by author, source: Appendix 2

This feature-based analysis was conducted in order understand, how much do certain features matter in the prediction results. The models with only ten most important features were analysed and the results can be see from Table 2. Once again the best performing machine learning approach on this dataset is Random Forest, with same metric ratings as the prediction of all the features. In this part of the analysis the worst ranking approach according to accuracy is Logistic Regression with cross-validation. The accuracy of this model is 0.6866, but it must be pointed out that recall metrics is as high as 0.9908. This could mean that the model is still good to predict the

outcomes and it minimizes the likelihood of false negatives. Figure 6 shows the AUC of all the analysed approaches with ten most important features selected in the dataset.

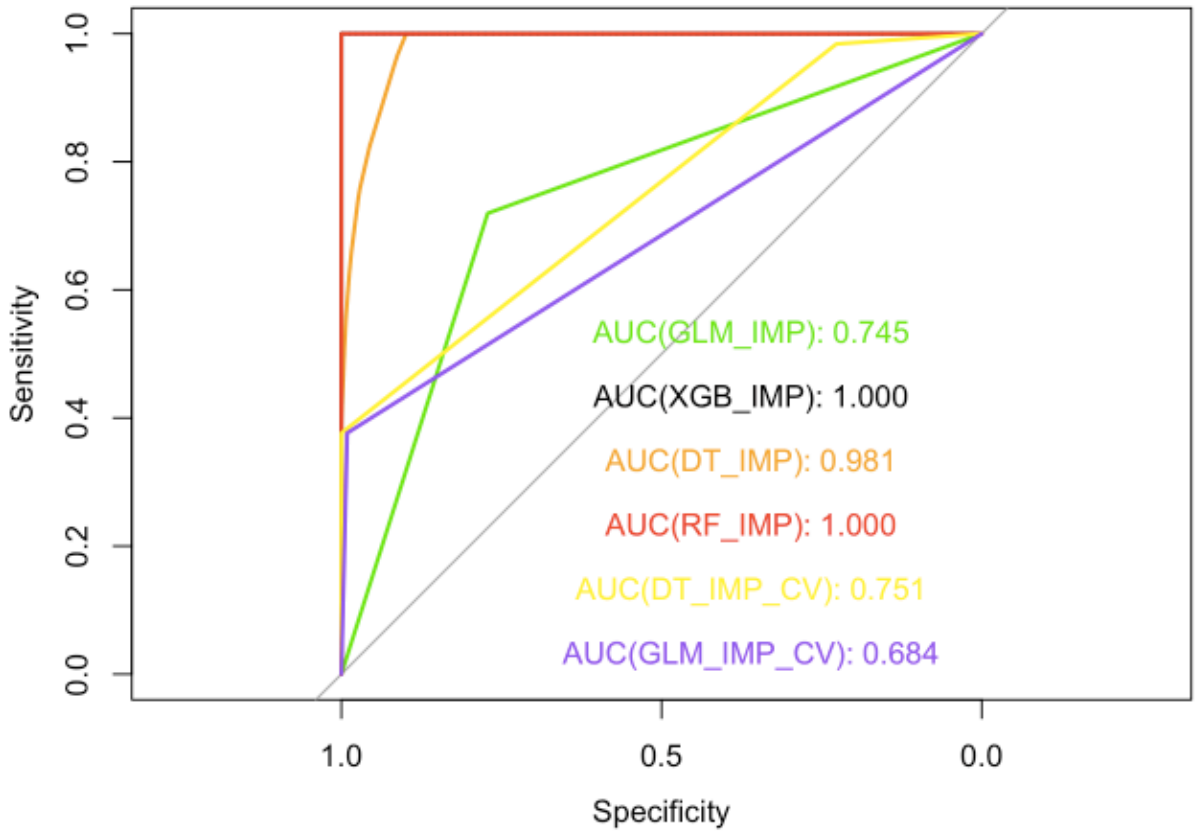

Figure 6. AUC of models with Top10 features, compiled by author, source: Appendix 2

The analysis shows that these models are good at predicting the outcome of classifying corruption in public procurement. The modelling with top ten features shows that even using a smaller sample of features it is possible to predict corruption on this data with relative accuracy and good results.

In the following sub-section the author will discuss and evaluate the results further and how they could be put to use in the future.

5.3. Evaluation

Data is influencing our everyday actions and it has been put to great use in private and public enterprises [101]. At first it was seen as an extra cost that the companies had to endure, but it has transformed how businesses are run nowadays [102].

As the related work suggests, it might be possible to predict corruption with the right features and indicators [19]–[21], [80]. This thesis shows that it is true. The results of the analysis depict that machine learning approaches are capable of classifying non- and corrupt acts based on public procurement data. Using a data analytics it is possible to predict corruption on the given dataset. The results show that the approaches are capable of classifying a tender as non – or corrupt with an accuracy of 74,44 % to 100 %. The possibility of treating a corrupt act as non-corrupt in the predictions is minimised, as the recall of the analysis for various approaches ranges from 78,01 % to 100%. The RMSE and MAE in Table 2 show that the error rate is low in the analysis and MAE is mostly lower with the exception of Random Forest, where both metrics are equal to zero.

The results show that there are certain indicators that matter most, when there is a risk of corruption, for example; the final price of the tender, the actual price of the bid, the indicator of whether it was a single bid or not and the estimated time of tender duration. The important features that have turned up in the analysis, have been brought out in related work as potential indicators of corruption [20], [39], [44]–[47], [103]. This certifies the validity of the analysis, which comes to similar conclusions and presents features that have been also highlighted in other papers. The fact that the length of requirements (personal, economical or technical) is important is another aspect that has been previously brought out in multiple studies and validates the analysis of this thesis. This refers to the problem that tender offers are sometimes orchestrated in order to fulfil certain requirements of a particular buyer. It is clear that this thesis present valuable results that can be taken forward. Previously worked out indicators have proven to be important and based on public procurement data it is possible to predict corruption.

However, this thesis has some drawbacks. First is the completeness and imbalance of the data. After conducting the analysis it is clear that a lot of initial data is missing from the database. On the other hand, it is vital that this database exists, as gathering information about Estonia is not

possible from anywhere else as a set. Having worked together with governmental agencies throughout the process of this thesis, it is evident that getting wholesome data is not possible otherwise. The data used is from the Estonian business registry can become better. Still, the fact that it has become more open over the years is crucial, but steps need to be taken for it to be more useful.

Second drawback of this thesis is the overfitting of some machine learning approaches. In the author's opinion Random Forest and XGBoost models might be overfitting the data, which creates very accurate, but not usable results. This could come from the high imputation level of the data or from the necessity of removing many features. On the other hand, the author provides cross-validated approaches like Logistic Regression and Decision Tree that also show that based on this data predicting corruption is fairly accurate and with low recall, which suits the context of this thesis.

In order to validate further, this research needs to be applied to future datasets with certified controls to give in a fair assessment.

5.4. Suggestions

In this sub-section the author will give recommendations on the subjects of machine learning approaches, features and data quality for more efficient corruption detection and prediction.

5.4.1 Feature selection

In the previous sub-section the author cover the main results of the analysis. These results show that there are features that are more important than others, which is supported by related work. In addition to the most important features the author will bring out relevant indicators could be used in the analysis of predicting corruption in Estonia, since the data is available. Looking at the information available in tender documents and on the Public Procurement Registry (PPR) website [43], the author makes the following suggestions:

- Add a feature about tender award criteria weights;
- Add a feature about automatic and non-automatic evaluation;

- Add a feature that evaluates the length of criteria on suitability;
- Add a feature that evaluates the length of grounds of exclusion criteria;
- Focus on features that have been highlighted as important in section 5.3.

All of these aforementioned features are available on the PPR website, but due to some reason they were not crawled into this dataset. These features could turn out to be very important, since the weights assigned to each criteria is a key decision point in awarding public procurements. The potential of this feature has also been brought out in related work, which supports this idea further. Also, it could be interesting to see, if the automation of evaluation plays a role in corruption. Furthermore, the author believes that the focus should be on the most important features and perhaps these could even be highlighted on the PPR website to provide further certainty and transparency. Bringing out certain data to the companies, public offices using this website, might make them more aware of the problem, which could help eradicate the problem in the first place.

5.4.2 Data quality

Data quality is always a question that must be addressed. As it was visible from the section of data cleaning and pre-processing, a lot of effort went into making this dataset suitable for data analysis. The number of missing data was very high and although some of the data comes from almost a decade ago, there should be efforts going into keeping a certain standard of data completeness. Based on this, the author makes following suggestions:

- Setting a standard to make sure that necessary information is always present in the database;

- Create a plan for utilising the possibilities of open data.

5.4.3 Machine learning approaches

In this thesis the author has used four machine learning approaches to evaluate the data. These were: Logistic Regression, Random Forest, Decison Tree and XGBoost. All of these have their

strengths and weaknesses, which must be taken into account. Based on past and current experience the author make the following suggestions:

- Use cross-validation to provide stability to the results;
- Apply each model separately and then combine to give a deeper understanding of the results;
- Use k-nn, SVM or other methods also to build the predictions.

The reason, why the author suggests using also other methods, is that in this thesis they were not sought after and it could be interesting to see in further research how these approaches compare to the ones currently used.

CONCLUSION

The main goal of this thesis was to assess, whether using open data resources and data analytics it is possible to predict corruption in the public procurement processes and therefore suggest a suitable set of data to make the detection of corruption easier and quicker.

In the theoretical part of the thesis the author gave an overview of the causes, effects of corruption and how public procurement is influenced by corruption. The high risk of corruption in public procurement has been brought out in various studies in Europe and all over the world. The qualitative aspects of these papers have developed an understanding why corruption takes place and how does it influence the society. However, there has been little research conducted on actually quantifying the possibility of corruption.

In the practical part of the thesis the author conducted data analytics on public procurement data in Estonia. The data comes from open sources and this means that the results could be implicated on other regions, as already this particular source has information on 32 other countries. In order to accomplish the main goal of the thesis the author applied machine learning approaches to predict corruption. The author believes that there are suitable data analytical approaches and capabilities available to quantify the problem of corruption, but the public offices need to take matters into their own hands and either use the data or pawn it off to the private sector, so that it turned into value. This means that in the fight against corruption the importance of open data has been undervalued.

Based on theoretical background and practical research the author was able to accomplish the set goal of the thesis.

The main results of this thesis are:

- The author has provided an overview of the state of corruption in public procurement;
- Through analysis the author has shown the potential of data analytics in corruption detection;
- The author has provided suggestions on how to improve;

- The author has developed an approach that can predict corruption well.

Data can be the game changer, when it comes to fighting corruption, as it will give a deeper insight into the complex paradigm that it is. Corruption can coexist with strong economic performance, but the research suggests that in the long run it is harmful for development, as it can be found in bureaucratic and political institutions. This harm can be eased using data analysis and understanding the weak points in the governance or laws that have been exploited by businessmen and politicians. The main problem is systemic corruption that has been ingrained in institutions, behaviours, and the habits of elites against the common good.

Using data analytics, the assessment of corruption can be made easier and faster. Next to qualitative research and work done to prevent corruption in the first place, data analytics can be of help, as people are always running plots against the system and there are various ways of hiding it. The author believes that there is huge potential in using data analytics in corruption detection. It is wrong to assume that these applications will start convicting people of corruption, but it can be assumed that using, analysing and presenting valuable data will help the society in assessing corruption and moulding the public opinion.

In the authors' opinion there are two possible development paths for the thesis. First, it is possible to apply current work it in the context of other countries, if there is necessary data. Following this idea, it could be researched how corruption in various countries differs and what are the common issues. The second development could be the use of more data, metrics and machine learning approaches to develop the analysis that has been done here further.

REFERENCES

[1] T. Soreide, *Corruption in public procurement Causes, consequences and cures*. 2002.

[2] T. International, "Corruption Perception Index 2018," 2018. .

[3] V. Bhargava, "World Bank Global Issues Seminar Series - The Cancer of Corruption," pp. 2–10, 2005.

[4] T.I., *Curbing Corruption in Public Procurement in Asia and the Pacific*. 2014.

[5] V. Tanzi, "Corruption Around the World," *Imf Staff Pap.*, vol. 45, no. 4, pp. 559–594, 1998.

[6] J. S. Hellman, G. Jones, and D. Kaufmann, "Seize the State, Seize the Day An empirical analysis of State Capture and Corruption in Transition Paper prepared for the ABCDE 2000 Conference Seize the State, Seize the Day An empirical analysis of State Capture and Corruption in Transition," 2000.

[7] A. Graycar and T. Prenzler, "Preventing Corruption in Public Sector Procurement," in *Understanding and Preventing Corruption*, 2016, pp. 100–113.

[8] E. Dimant and G. Tosato, "Causes and Effects of Corruption: What Has Past Decade'S Empirical Research Taught Us? a Survey," *J. Econ. Surv.*, vol. 32, no. 2, pp. 335–356, 2018.

[9] R. K. Goel and M. A. Nelson, "Causes of corruption: History, geography and government," *J. Policy Model.*, 2010.

[10] A. R. Menoca, "Why corruption matters: understanding causes, effects and how to address them," *UK Dep. Int. Dev.*, no. January, p. 18,19, 2015.

[11] B. J. Palifka and S. Rose-Ackerman, *Corruption and Government: causes, consequences, and reform*, 2nd ed. Cambridge University Press, 2016.

[12] J. Jensen and T. Anderson, "A Qualitative Comparison of Anti-Corruption Measures in Guatemala and Brazil," *J. Polit. Democr.*, vol. 1, no. 2, pp. 1–20, 2016.

[13] Ipadeola O., "Qualitative Study on the Patterns , Experiences and Manifestations of Coruption in Nigeria," no. March, 2016.

[14] M. A. Ullah and T. Arthanari, "Using a Qualitative System Dynamics Approach to Investigate Perceptions of Corruption," *Int. Conf. Syst. Dyn. Soc.*, 2011.

[15] A. R. Las Johannsen, Karin Hilmer Pedersen, Maaja Vadi, “Private-to-Private Corruption environment,” 2016.

[16] J. You and S. Khagram, “A Comparative Study of Inequality and Corruption,” *Ssrn*, 2004.

[17] G. Ariely and E. M. Uslaner, “Corruption, fairness, and inequality,” *Int. Polit. Sci. Rev.*, vol. 38, no. 3, pp. 349–362, 2017.

[18] K. Ko and A. Samajdar, “Evaluation of international corruption indexes: Should we believe them or not?,” *Soc. Sci. J.*, vol. 47, no. 3, pp. 508–540, 2010.

[19] M. Fazekas and G. Kocsis, “Uncovering High-Level Corruption: Cross-National Objective Corruption Risk Indicators Using Public Procurement Data,” *Br. J. Polit. Sci.*, pp. 1–10, 2017.

[20] M. Fazekas, I. J. Tóth, and L. P. King, “An Objective Corruption Risk Index Using Public Procurement Data,” *Eur. J. Crim. Policy Res.*, vol. 22, no. 3, pp. 369–397, 2016.

[21] E. Dávid-Barrett, M. Fazekas, O. Hellmann, L. Márk, and C. McCorley, “Controlling Corruption in Development Aid: New Evidence from Contract-Level Data,” 2018.

[22] L. Sales, “Risk prevention of public procurement in the brazilian government using credit scoring,” *Work. Pap. - OBEGEF*, vol. 2013, p. 27, 2013.

[23] R. S. Carvalho, R. N. Carvalho, M. Ladeira, F. M. Monteiro, and G. L. De Oliveira Mendes, “Using political party affiliation data to measure civil servants’ risk of corruption,” *Proc. - 2014 Brazilian Conf. Intell. Syst. BRACIS 2014*, pp. 166–171, 2014.

[24] R. Wheeler and S. Aitken, “for Fraud Detection,” 2000.

[25] L. Šubelj, Š. Furlan, and M. Bajec, “An expert system for detecting automobile insurance fraud using social network analysis,” *Expert Syst. Appl.*, vol. 38, no. 1, pp. 1039–1052, 2011.

[26] S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, “Data mining for credit card fraud: A comparative study,” *Decis. Support Syst.*, vol. 50, no. 3, pp. 602–613, 2011.

[27] S. B. Caudill, M. Ayuso, and M. Guillen, “FRAUD DETECTION USING A MULTINOMIAL LOGIT MODEL WITH MISSING INFORMATION,” vol. 1, no. 5, pp. 1–12, 2012.

[28] S. B. E. Raj, A. A. Portia, and A. Sg, “Analysis on Credit Card Fraud Detection Methods,” *2011 Int. Conf. Comput. Commun. Electr. Technol.*, pp. 152–156, 2011.

[29] V. Van Vlasselaer, L. Akoglu, T. Eliassi-Rad, M. Snoeck, and B. Baesens, "Guilt-by-constellation: Fraud detection by suspicious clique memberships," *Proc. Annu. Hawaii Int. Conf. Syst. Sci.*, vol. 2015-March, pp. 918–927, 2015.

[30] DIGIWHIST, "Open Tender EU," 2019. [Online]. Available: https://opentender.eu/ee/.

[31] K. N. Hakkala, P.-J. Norbäck, and H. Svaleryd, "Asymmetric Effects of Corruption on FDI: Evidence from Swedish Multinational Firms," *Rev. Econ. Stat.*, vol. 90, no. 4, pp. 627–642, 2008.

[32] D. della Porta and A. Vannucci, "The 'Perverse Effects'' of Political Corruption,'" *Polit. Stud.*, vol. 45, no. 3, pp. 516–538, 1997.

[33] M. A. R. Salih, "The Determinants of Economic Corruption: A Probabilistic Approach," *Adv. Manag. Appl. Econ.*, vol. 3, no. 3, pp. 155–169, 2013.

[34] V. Tanzi and H. Davoodi, "The Welfare State, Public Investment, and Growth: Selected Papers from the 53rd Congress of the International Institute of Public Finance," pp. 41–60, 1998.

[35] J. G. Lambsdorff and P. Cornelius, "Corruption, Foreign Investment and Growth," pp. 70–78, 2000.

[36] S. Wei, "Why is Corruption So Much More Taxing than Tax? Arbitrariness Kills," 1997.

[37] C. W. Abramo, "How Much Do Perceptions of Corruption Really Tell Us?," *Ssrn*, vol. 2, 2010.

[38] C. Sampford, A. Shacklock, C. Connors, and F. Galtung, *Measuring Corruption - The Validity and Precision of Subjective Indicators (CPI)*. 2006.

[39] European Commision, "Fraud in Public Procurement A collection of Red Flags and Best Practices," 2017.

[40] M. A. Golden and L. Picci, "Proposal for a New Measure of Corruption, and Tests using Italian Data," *Econ. Polit.*, vol. 17, no. 3, pp. 37–75, 2005.

[41] A. Hyytinen, S. Lundberg, and O. Toivanen, *Politics and Procurement: Evidence from Cleaning Contracts*, vol. 17, no. 277. 2008.

[42] P. Ayoub, S. Wallace, and C. Zepeda-Millan, "Triangulation in Social Movement Research," 2014.

[43] “E-procurement web page in Estonia.” [Online]. Available: https://riigihanked.riik.ee/rhr-web/#/.

[44] J. Ferwerda, I. Deleanu, and B. Unger, “Corruption in Public Procurement: Finding the Right Indicators,” *Eur. J. Crim. Policy Res.*, vol. 23, no. 2, pp. 245–267, 2017.

[45] C. Kenny and M. Musatova, “‘Red Flags of Corruption’ in World Bank Projects: An Analysis of Infrastructure Contracts,” *World Bank, Policy Res. Work. Pap. Ser.*, 2010.

[46] IBAC, “The red flags of corruption : Procurement,” pp. 1–3, 2015.

[47] T. U. S. Foreign, C. Practices, A. Fcpa, U. K. Bribery, A. Ukba, and S. Ii, “Bribery and Corruption Red Flags ‘ How to Respond to Corruption Indicators ,’” vol. 44, no. 0.

[48] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier Science, 2016.

[49] X. Sun, Y. Hu, Y. Chen, Y. H. Wong, and E. W. T. Ngai, “The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature,” *Decis. Support Syst.*, vol. 50, no. 3, pp. 559–569, 2010.

[50] C. M. Bishop, *Pattern recognition and machine learning*, First. New York: Springer US, 2006.

[51] S. B. Kotsiantis, “Supervised Machine Learning : A Review of Classification Techniques,” vol. 31, pp. 249–268, 2007.

[52] C. Phua, D. Alahakoon, and V. Lee, “Minority report in fraud detection,” *ACM SIGKDD Explor. Newsl.*, vol. 6, no. 1, p. 50, 2007.

[53] S. Wang, “A comprehensive survey of data mining-based accounting-fraud detection research,” *2010 Int. Conf. Intell. Comput. Technol. Autom. ICICTA 2010*, vol. 1, pp. 50–53, 2010.

[54] S. Yuan, X. Wu, J. Li, and A. Lu, “Spectrum-based deep neural networks for fraud detection,” pp. 1–5, 2017.

[55] S. K. Murthy, “Automatic construction of decision trees from data: A multi-disciplinary survey,” *Data Min. Knowl. Discov.*, vol. 2, no. 4, pp. 345–389, 1998.

[56] L. Breiman, “Random Forests.,” pp. 1–33, 2001.

[57] R. Caruana and A. Niculescu-Mizil, “An Empirical Comparison of Supervised Learning

Algorithms," pp. 161–168, 2006.

[58] D. Meyer, F. Leisch, and K. Hornik, "The support vector machine under test," *Neurocomputing*, vol. 55, no. 1–2, pp. 169–186, 2003.

[59] T. Chen and C. Guestrin, "XGBoost," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, 2016, pp. 785–794.

[60] J. Davis and M. Goadrich, "The Relationship Between Precision-Recall and ROC Curves," *Planning*, vol. 73, no. 10, p. 55, 2007.

[61] M. Sunasra, "Performance Metrics for Classification problems in Machine Learning," 2017. [Online]. Available: https://medium.com/thalus-ai/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b.

[62] A. Mishra, "Metrics to Evaluate your Machine Learning Algorithm," *Towards Data Science*, 2018. [Online]. Available: https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234.

[63] B. E. Ashforth and V. Anand, "THE NORMALIZATION OF CORRUPTION IN ORGANIZATIONS," *Res. Organ. Behav.*, vol. 25, pp. 1–52, Jan. 2003.

[64] P. Anand, G. Hunter, and R. Smith, "Capabilities and Well-Being: Evidence Based on the Sen--Nussbaum Approach to Welfare," *Soc. Indic. Res.*, vol. 74, no. 1, pp. 9–55, Oct. 2005.

[65] T. Rabl and T. Kühlmann, "Why or why not? Rationalizing Corruption in Organizations," *Cross Cult. Manag. An Int. J.*, vol. 16, pp. 268–286, 2009.

[66] T. International, *Global corruption report, 2001*, vol. 42, no. 04. 2001.

[67] H. Li, L. C. Xu, and H. Zou, "Corruption, Income Distribution, and Growth," *Econ. Polit.*, vol. 12, no. 2, pp. 155–182, 2000.

[68] S. Akcay, "Corruption and Human Development ," *Cato J.*, no. 1, pp. 29–48.

[69] D. Kaufmann, "Corruption: The Facts," *Foreign Policy*, no. 107, p. 114, 1997.

[70] Z. Kutlina-Dimitrova and L. Cernat, "Government Procurement: Data, Trends and Protectionist Tendencies," *Trade*, no. 3, pp. 1–27, 2018.

[71] OECD, "OECD Statistics," *Public Procurement*, 2015. .

[72] D. Corvino, A. Oeking, L. Leigh, A. Shukla, S. Hlatshwayo, and M. Ghazanchyan, “The Measurement and Macro-Relevance of Corruption: A Big Data Approach,” *IMF Work. Pap.*, vol. 18, no. 195, p. 1, 2018.

[73] R. Rose and W. Mishler, “Experience versus perception of corruption: Russia as a test case,” *Glob. Crime*, vol. 11, no. 2, pp. 145–163, 2010.

[74] N. Charron, “Do corruption measures have a perception problem? Assessing the relationship between experiences and perceptions of corruption among citizens and experts,” *Eur. Polit. Sci. Rev.*, vol. 8, no. 1, pp. 147–171, 2014.

[75] D. Peltier-Rivest, “The prevention and detection of corruption in pharmaceutical companies,” *Pharm. Policy Law*, vol. 19, no. 1–2, pp. 17–31, 2017.

[76] S. Zimmermann, “Using Data and Transparency to Fight Corruption in Public Procurement.”

[77] G. Fréchette, “Panel Data Analysis of the Time-Varying Determinants of Corruption,” *CIRANO Work. Pap.*, 2006.

[78] N. G. Elbahnasawy and C. F. Revier, “The Determinants of Corruption: Cross-Country-Panel-Data Analysis,” *Dev. Econ.*, vol. 50, no. 4, pp. 311–333, 2012.

[79] J. Krafka, J. Hrubý, T. Pošepný, J. Krafka, B. Toth, and J. Skuhrovec, “EU Grant Agreement number : Project title : The Digital Whistleblower : Fiscal,” 2019.

[80] R. S. Carvalho, R. N. Carvalho, M. Ladeira, F. M. Monteiro, and G. L. De Oliveira Mendes, “Using political party affiliation data to measure civil servants’ risk of corruption,” *Proc. - 2014 Brazilian Conf. Intell. Syst. BRACIS 2014*, pp. 166–171, 2014.

[81] B. B. Olsen, D. Reynolds, and A. Koltsov, “Using Data Analytics to Meet the Government ’ s Anti-Corruption Compliance Expectations,” vol. 5, no. 9, pp. 1–4, 2016.

[82] F. J. López-Iturriaga and I. P. Sanz, “Predicting Public Corruption with Neural Networks: An Analysis of Spanish Provinces,” *Soc. Indic. Res.*, vol. 140, no. 3, pp. 975–998, 2018.

[83] United Nations Office on Drugs and Crime, “Quantitative approaches to assess and describe corruption and the role of UNODC in supporting countries in performing such assessments Background paper prepared by the Secretariat,” vol. V.09-87564, no. November, pp. 9–13, 2009.

[84] L.-L. S. Anna K. Schwickerath, Aiysha Varraich, “How to research corruption?,” in *Interdisciplinary Corruption Research Forum*, 2016, no. June.

[85] N. Fielding, R. M. Lee, and G. Blank, “Nonreactive Data Collection on the Internet,” pp. 161–175, 2019.

[86] Riigiteataja, *Court documents, cases:1-16-7539, 1-18-2357, 1-18-2452, 1-18-2692, 1-18-2529-1-17-2185,1-18-4321, 1-18-7157, 1-12-9126, 2-16-116135/12, 1-15-9268, 1-12-12478/405*. .

[87] A. Khashman, “Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes,” *Expert Syst. Appl.*, vol. 37, no. 9, pp. 6233–6239, 2010.

[88] S. Lessmann, B. Baesens, H. V. Seow, and L. C. Thomas, “Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research,” *Eur. J. Oper. Res.*, vol. 247, no. 1, pp. 124–136, 2015.

[89] G. Seif, “The Art of Cleaning Your Data,” 2018. [Online]. Available: https://towardsdatascience.com/the-art-of-cleaning-your-data-b713dbd49726.

[90] T. Smishad, “Data Cleaning in Machine Learning: Best Practices and Methods,” 2018. [Online]. Available: https://www.einfochips.com/blog/data-cleaning-in-machine-learning-best-practices-and-methods/.

[91] O. Elgabry, “The Ultimate Guide to Data Cleaning,” 2019. [Online]. Available: https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4.

[92] R. Pierre, “Detecting Financial Fraud Using Machine Learning: Winning the War Against Imbalanced Data,” 2018. [Online]. Available: https://towardsdatascience.com/detecting-financial-fraud-using-machine-learning-three-ways-of-winning-the-war-against-imbalanced-a03f8815cce9.

[93] J. Hale, “7 Data Types: A Better Way to Think about Data Types for Machine Learning,” 2018. [Online]. Available: https://towardsdatascience.com/7-data-types-a-better-way-to-think-about-data-types-for-machine-learning-939fae99a689.

[94] A. Sudharsan, “Why, How and When to Scale your Features,” 2018. [Online]. Available: https://medium.com/greyatom/why-how-and-when-to-scale-your-features-4b30ab09db5e.

[95] P. Evans, “Scaling and assessment of data quality,” *Acta Crystallogr. Sect. D*, vol. 62, no. 1, pp. 72–82, Jan. 2006.

[96] R. Kohavi, “A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection,” vol. 14, 2001.

[97] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, “Bias in random forest variable

importance measures: Illustrations, sources and a solution," *BMC Bioinformatics*, vol. 8, no. 1, p. 25, 2007.

[98] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional variable importance for random forests," *BMC Bioinformatics*, vol. 9, no. 1, p. 307, 2008.

[99] J. Le, "Logistic Regression in R Tutorial," 2018. [Online]. Available: https://www.datacamp.com/community/tutorials/logistic-regression-R.

[100] J. Le, "Decision Trees in R," 2018. [Online]. Available: https://www.datacamp.com/community/tutorials/decision-trees-R.

[101] H. Chen, R. H. L. Chiang, and V. C. Storey, "B Usiness I Ntelligence and a Nalytics," *MIS Q.*, vol. 36, no. 4, pp. 1165–1188, 2012.

[102] C. Loebbecke and A. Picot, "Reflections on societal and business model transformation arising from digitization and big data analytics: A research agenda," *J. Strateg. Inf. Syst.*, vol. 24, no. 3, pp. 149–157, 2015.

[103] Development Gateway, *RED FLAGS for integrity : Giving the green light to open data solutions*. .

APPENDICES

Appendix 1

Table 3. Corruption effects, compiled by author, source: [66]

Author(s)	**Impact on**	**Finding**
Mauro (1996)	Real per capita GDP growth	–0.3 to –1.8 percentage points
Leite and Weidmann (1999)	Real per capita GDP growth	–0.7 to –1.2 percentage points
Tanzi and Davoodi (2000)	Real per capita GDP growth	–0.6 percentage points
Abed and Davoodi (2000)	Real per capita GDP growth	–1 to –1.3 percentage points
Mauro (1996)	Ratio of investment to GDP	–1 to –2.8 percentage points
Mauro (1998)	Ratio of public education spending to GDP	–0.7 to –0.9 percentage points
Mauro (1998)	Ratio of public health spending to GDP	–0.6 to –1.7 percentage points
Gupta, Davoodi and Alonso-Terme (1998)	Income inequality (Gini coefficient)	+0.9 to +2.1 Gini points
Gupta, Davoodi and Alonso-Terme (1998)	Income growth of the poor	–2 to –10 percentage points
Ghura (1998)	Ratio of tax revenues to GDP	–1 to –2.9 percentage points
Tanzi and Davoodi (2000)	Measures of government revenues to GDP ratio	–0.1 to –4.5 percentage points
Gupta, de Mello and Sharan (2000)	Ratio of military spending to GDP	+1 percentage point

Gupta, Davoodi and Tiongson (2000)	Child mortality rate	+1.1 to 2.7 deaths per 1,000 live births
Gupta, Davoodi and Tiongson (2000)	Primary student dropout rate	+1.4 to 4.8 percentage points
Tanzi and Davoodi (1997)	Ratio of public investment to GDP	+0.5 percentage points
Tanzi and Davoodi (1997)	Per cent of paved roads in good condition	–2.2 to –3.9 percentage points

Appendix 2 – R Code

```
---
output:
  html_document: default
  pdf_document: default
  word_document: default
---
```r
library(dplyr)
library(ggplot2)
library(ROSE) # Balancing
library(pROC) # AUC
library(reshape) #melt
library(stringr) # work with strings
library(rpart) # decision tree
library(arsenal) #descriptive statistics

library(xgboost) # for xgboost
library(readr)
andmed09 <- read.csv("data2009.csv",sep=";", stringsAsFactors=FALSE, fileEncoding="latin1")
andmed10 <- read.csv("data2010.csv",sep=";", stringsAsFactors=FALSE, fileEncoding="latin1")
andmed11 <- read.csv("data2011.csv",sep=";", stringsAsFactors=FALSE, fileEncoding="latin1")
andmed12 <- read.csv("data2012.csv",sep=";", stringsAsFactors=FALSE, fileEncoding="latin1")
andmed13 <- read.csv("data2013.csv",sep=";", stringsAsFactors=FALSE, fileEncoding="latin1")
andmed14 <- read.csv("data2014.csv",sep=";", stringsAsFactors=FALSE, fileEncoding="latin1")
andmed15 <- read.csv("data2015.csv",sep=";", stringsAsFactors=FALSE, fileEncoding="latin1")


summary(andmed13)
```
```r
table(sapply(andmed, class))
sapply(andmed, function(x) sum(is.na(x)))
```

Not adding data from 2013,as it does not have any corrupt cases

```r
andmed <- rbind(andmed09, andmed10, andmed11, andmed12,andmed14,andmed15)
```

```r
# package required printing important features for XGBoost
#install.packages("Ckmeans.1d.dp")
library(Ckmeans.1d.dp)
```
```

```
# for confusion matrix
#install.packages("e1071")
#install.packages("Rcpp")
library(e1071)
library(Rcpp)
```

```{r}
#Then install caret
#install.packages('caret', dependencies=TRUE)
library(caret)
```

```{r}
emptycols <- sapply(andmed, function (k) all(is.na(k)))
andmed <- andmed[!emptycols]
```
Removing from the dataset completely unique variables (row nr, id, title, url) because these variables do not improve any modelling. However bidder IDs, physical locations, addresses are kept as factorials because a possible pattern might be revealed in a certain geographical location, bidder or otherwise 'unique' factorial.
Secondly remove all columns from analysis with less than 5 percent filled data (all string variables, leaving us with 65 variables of which 40 are numerical)

```{r}
#glimpse(andmed)

# Unique IDs and duplicates removed
andmed.stripped <- andmed %>% select(-c(tender_year, tender_estimatedPrice_EUR, tender_finalPrice_EUR, bid_price_EUR,
lot_estimatedPrice_EUR,tender_row_nr, tender_id, tender_title, tender_publications_lastContractAwardUrl, bidder_row_nr,
tender_indicator_ADMINISTRATIVE_ELECTRONIC_AUCTION))
# Dates
andmed.nodate <- andmed.stripped %>% select(-contains("Date"))
# Poorly filled data (<5%)
removed.cols <- andmed.nodate[, -which(colMeans(is.na(andmed.nodate)) > 0.95)]
#View(removed.cols)

andmed.nodate <- andmed.nodate[, -which(colMeans(is.na(andmed.nodate)) > 0.95)]
filter(andmed.nodate, rowSums(!is.na(andmed)) > 45 | tender_isCorrupt == 'yes') %>%# visualise positive controls, filtering for at least 50% of
columns filled in a single row and all positive  rows regardless of % of NA values.
  count(tender_isCorrupt == 'yes') %>%
  glimpse # count number of positive controls

# Mutate the isCorrupt data to a boolean
andmed.final <- andmed.nodate %>%
  mutate(tender_isCorrupt = ifelse(tender_isCorrupt == "no",0,1))

andmed.final$tender_mainCpv = as.factor(andmed.final$tender_mainCpv)
andmed.final$buyer_row_nr = as.factor(andmed.final$buyer_row_nr)
andmed.final$buyer_postcode = as.factor(andmed.final$buyer_postcode)
andmed.final <- andmed.final[,sapply(andmed.final, is.numeric)]
```

```{r}
table(sapply(andmed.stripped, class))
```

```{r}
count(andmed.final, is.na(andmed.final$tender_finalPrice))
summary(andmed.final)
```

```{r}
count(andmed.final, is.na(andmed.final$tender_indicator_INTEGRITY_SINGLE_BID))
```
Filtering out the rows with more than 70 000 NAs
```{r}
dt <- andmed.final %>%
  select(-c(tender_estimatedDurationInMonths, tender_estimatedPrice,tender_indicator_INTEGRITY_TAX_HAVEN,
tender_indicator_ADMINISTRATIVE_ENGLISH_AS_FOREIGN_LANGUAGE, lot_estimatedPrice,lot_description_length))
```

```
count(dt, is.na(dt$tender_finalPrice))
summary(dt)
```

```{r}
library(Hmisc)
Hmisc::describe(dt)
```

```{r}
library(mice)
library(VIM)
aggr_plot <- aggr(dt, col=c('navyblue','red'), numbers=TRUE, sortVars=TRUE, labels=names(dt), cex.axis=.7, gap=3, ylab=c("Histogram of missing data","Pattern"))
dt_imputed<- mice(dt, m=5, method='cart', printFlag=FALSE)
```

```{r}
dt <- complete(dt_imputed, 1)
```

```{r}
summary(dt)
library(Hmisc)
Hmisc::describe(dt)
```

```{r}
dt_b <- ovun.sample(tender_isCorrupt~. , data = dt, method = "both",
                            p=0.5, seed = 1000)$data

ggplot(dt_b, aes(x = tender_isCorrupt, fill = as.factor(tender_isCorrupt))) +
  geom_bar()
```

```{r}
dt_b$tender_finalPrice <- as.numeric(dt_b$tender_finalPrice)
dt_b$tender_estimatedDurationInDays <- as.numeric(dt_b$tender_estimatedDurationInDays)
dt_b$tender_description_length <- as.numeric(dt_b$tender_description_length)
dt_b$tender_economicRequirements_length <- as.numeric(dt_b$tender_economicRequirements_length)
dt_b$tender_technicalRequirements_length <- as.numeric(dt_b$tender_technicalRequirements_length)
dt_b$tender_awardCriteria_count <- as.numeric(dt_b$tender_awardCriteria_count)
dt_b$tender_lots_count <- as.numeric(dt_b$tender_lots_count)
dt_b$tender_publications_count <- as.numeric(dt_b$tender_publications_count)
dt_b$lot_row_nr <- as.numeric(dt_b$lot_row_nr)
dt_b$lot_bidsCount <- as.numeric(dt_b$lot_bidsCount)
dt_b$bid_row_nr <- as.numeric(dt_b$bid_row_nr)
dt_b$tender_personalRequirements_length <- as.numeric(dt_b$tender_personalRequirements_length)
dt_b$tender_technicalRequirements_length <- as.numeric(dt_b$tender_technicalRequirements_length)
dt_b$lot_bidsCount <- as.numeric(dt_b$lot_bidsCount)


dt_b$tender_indicator_INTEGRITY_SINGLE_BID <- as.numeric(dt_b$tender_indicator_INTEGRITY_SINGLE_BID)
dt_b$tender_isCorrupt <- as.numeric(dt_b$tender_isCorrupt )
dt_b$tender_indicator_INTEGRITY_CALL_FOR_TENDER_PUBLICATION <-
as.numeric(dt_b$tender_indicator_INTEGRITY_CALL_FOR_TENDER_PUBLICATION )
dt_b$tender_indicator_INTEGRITY_ADVERTISEMENT_PERIOD <-
as.numeric(dt_b$tender_indicator_INTEGRITY_ADVERTISEMENT_PERIOD )
dt_b$tender_indicator_INTEGRITY_PROCEDURE_TYPE <- as.numeric(dt_b$tender_indicator_INTEGRITY_PROCEDURE_TYPE )
dt_b$tender_indicator_INTEGRITY_DECISION_PERIOD <- as.numeric(dt_b$tender_indicator_INTEGRITY_DECISION_PERIOD )
dt_b$tender_indicator_ADMINISTRATIVE_COVERED_BY_GPA <-
as.numeric(dt_b$tender_indicator_ADMINISTRATIVE_COVERED_BY_GPA )
dt_b$tender_indicator_ADMINISTRATIVE_FRAMEWORK_AGREEMENT <-
as.numeric(dt_b$tender_indicator_ADMINISTRATIVE_FRAMEWORK_AGREEMENT )
dt_b$tender_indicator_ADMINISTRATIVE_NOTICE_AND_AWARD_DISCREPANCIES <-
as.numeric(dt_b$tender_indicator_ADMINISTRATIVE_NOTICE_AND_AWARD_DISCREPANCIES )
dt_b$tender_indicator_INTEGRITY_DECISION_PERIOD <- as.numeric(dt_b$tender_indicator_INTEGRITY_DECISION_PERIOD )

```

```{r}
scaled_df <- dt_b %>%
   mutate(tender_finalPrice = scale(tender_finalPrice),
       tender_estimatedDurationInDays = scale(tender_estimatedDurationInDays),
       tender_description_length = scale(tender_description_length),
       tender_economicRequirements_length = scale(tender_economicRequirements_length),
       tender_awardCriteria_count = scale(tender_awardCriteria_count),
       tender_lots_count = scale(tender_lots_count),
       tender_publications_count = scale(tender_publications_count),
       lot_row_nr = scale(lot_row_nr),
       bid_row_nr = scale(bid_row_nr),
       bid_price = scale(bid_price),
       lot_row_nr = scale(lot_row_nr),
       tender_personalRequirements_length = scale(tender_personalRequirements_length),
       tender_technicalRequirements_length = scale(tender_technicalRequirements_length),
       lot_bidsCount = scale(lot_bidsCount),
       tender_indicator_ADMINISTRATIVE_NOTICE_AND_AWARD_DISCREPANCIES =
scale(tender_indicator_ADMINISTRATIVE_NOTICE_AND_AWARD_DISCREPANCIES),
       tender_indicator_INTEGRITY_SINGLE_BID = scale(tender_indicator_INTEGRITY_SINGLE_BID),
       tender_indicator_INTEGRITY_CALL_FOR_TENDER_PUBLICATION =
scale(tender_indicator_INTEGRITY_CALL_FOR_TENDER_PUBLICATION),
       tender_indicator_INTEGRITY_ADVERTISEMENT_PERIOD = scale(tender_indicator_INTEGRITY_ADVERTISEMENT_PERIOD),
       tender_indicator_INTEGRITY_PROCEDURE_TYPE = scale(tender_indicator_INTEGRITY_PROCEDURE_TYPE),
       tender_indicator_ADMINISTRATIVE_COVERED_BY_GPA = scale(tender_indicator_ADMINISTRATIVE_COVERED_BY_GPA),
       tender_indicator_ADMINISTRATIVE_FRAMEWORK_AGREEMENT =
scale(tender_indicator_ADMINISTRATIVE_FRAMEWORK_AGREEMENT),
       tender_indicator_INTEGRITY_DECISION_PERIOD = scale(tender_indicator_INTEGRITY_DECISION_PERIOD),
       )
```

```{r}
glimpse(scaled_df)
summary(scaled_df)
```

```{r}
set.seed(101)
samples <- sample.int(n = nrow(scaled_df), size = floor(0.8*nrow(scaled_df)), replace = F)
train <- scaled_df[samples,]
test <- scaled_df[-samples,]
#Logistic Regression
glm <- glm(tender_isCorrupt ~ ., train, family=gaussian)
pred_glm_scores <- predict(glm, test)
threshold <-  0.5 # We set threshold for obtained scores
glm_predict_boolean <- ifelse(pred_glm_scores >= threshold,1,0)

glm_cm <- confusionMatrix(as.factor(glm_predict_boolean), as.factor(test$tender_isCorrupt))
glm_cm

data.frame( R2 = R2(glm_predict_boolean, test$tender_isCorrupt),
          RMSE = RMSE(glm_predict_boolean, test$tender_isCorrupt),
          MAE = MAE(glm_predict_boolean, test$tender_isCorrupt))

print(paste("Recall of Logistic Regression is:", glm_cm$byClass["Recall"]))
print(paste("Precision of Logistic Regression is:", glm_cm$byClass["Precision"]))
print(paste("F1 of Logistic Regression is:", glm_cm$byClass["F1"]))

roc_glm <- roc(test$tender_isCorrupt, glm_predict_boolean)
plot.roc(roc_glm, col="orange")
text(0.36, 0.53, labels=sprintf("AUC(GLM)): %0.3f", auc(roc_glm)), col="orange")

```

```{r}
#GLM with Cross Validation
set.seed(101)
# Define train control for k fold cross validation
train_control <- trainControl(method="cv", number=10, savePredictions = TRUE, classProbs = TRUE)
```

```
# Fit Linear Regression Model
modelGLM <- train(tender_isCorrupt~., data=train, trControl=train_control, method="glm")
# Summarise Results
print(modelGLM)
#GLM with Cross Validation
pred_glm_scores2 <- predict(modelGLM, test)
threshold <-  0.5 # We set threshold for obtained scores
glm_predict_boolean2 <- ifelse(pred_glm_scores2 >= threshold,1,0)

glm_cm2 <- confusionMatrix(as.factor(glm_predict_boolean2), as.factor(test$tender_isCorrupt))
glm_cm2

data.frame( R2 = R2(glm_predict_boolean2, test$tender_isCorrupt),
          RMSE = RMSE(glm_predict_boolean2, test$tender_isCorrupt),
          MAE = MAE(glm_predict_boolean2, test$tender_isCorrupt))

print(paste("Recall of Logistic Regression is:", glm_cm2$byClass["Recall"]))
print(paste("Precision of Logistic Regression is:", glm_cm2$byClass["Precision"]))
print(paste("F1 of Logistic Regression is:", glm_cm2$byClass["F1"]))

roc_glm2 <- roc(test$tender_isCorrupt, glm_predict_boolean2)
plot.roc(roc_glm2, col="orange")
text(0.36, 0.53, labels=sprintf("AUC(GLM with CV)): %0.3f", auc(roc_glm2)), col="orange")

```

```{r}
#Decision Tree
set.seed(101)
samples <- sample.int(n = nrow(scaled_df), size = floor(0.8*nrow(scaled_df)), replace = F)
train <- scaled_df[samples,]
test <- scaled_df[-samples,]

tree <- rpart(tender_isCorrupt~., train)
pred_tree_scores <- predict(tree, test)

threshold <-  0.5 # We set threshold for obtained scores
tree_predict_boolean <- ifelse(pred_tree_scores >= threshold,1,0)

data.frame( R2 = R2(tree_predict_boolean, test$tender_isCorrupt),
          RMSE = RMSE(tree_predict_boolean, test$tender_isCorrupt),
          MAE = MAE(tree_predict_boolean, test$tender_isCorrupt))
tree_cm <- confusionMatrix(as.factor(tree_predict_boolean), as.factor(test$tender_isCorrupt))
tree_cm

print(paste("Recall of Decision Tree is:", tree_cm$byClass["Recall"]))
print(paste("Precision of Decision Tree is:", tree_cm$byClass["Precision"]))
print(paste("F1 of Decision Tree is:", tree_cm$byClass["F1"]))

roc_dt <- roc(test$tender_isCorrupt, predict(tree, test))
plot.roc(roc_dt, col="green")
text(0.36, 0.53, labels=sprintf("AUC(DT)): %0.3f", auc(roc_dt)), col="green")
```

```{r}
#Decision Tree with Cross Validation
set.seed(101)
# Define train control for k fold cross validation
train_control <- trainControl(method="cv", number=10, savePredictions = TRUE, classProbs = TRUE)
# Fit Decision Tree Model
modelDT <- train(tender_isCorrupt~., data=train, trControl=train_control, method="rpart")
# Summarise Results
print(modelDT)

pred_tree_scores2 <- predict(modelDT, test)

threshold <-  0.5 # We set threshold for obtained scores
tree_predict_boolean2 <- ifelse(pred_tree_scores2 >= threshold,1,0)
```

```
data.frame( R2 = R2(tree_predict_boolean2, test$tender_isCorrupt),
          RMSE = RMSE(tree_predict_boolean2, test$tender_isCorrupt),
          MAE = MAE(tree_predict_boolean2, test$tender_isCorrupt))
tree_cm2 <- confusionMatrix(as.factor(tree_predict_boolean2), as.factor(test$tender_isCorrupt))
tree_cm2

print(paste("Recall of Decision Tree with Cross Validation is:", tree_cm2$byClass["Recall"]))
print(paste("Precision of Decision Tree with Cross Validation is:", tree_cm2$byClass["Precision"]))
print(paste("F1 of Decision Tree with Cross Validation is:", tree_cm2$byClass["F1"]))


roc_dt2 <- roc(test$tender_isCorrupt, predict(modelDT, test))
plot.roc(roc_dt2, col="green")
text(0.36, 0.53, labels=sprintf("AUC(DT with CV)): %0.3f", auc(roc_dt2)), col="green")
```

```{r}
library(randomForest)
#Random Forest
set.seed(101)
samples <- sample.int(n = nrow(scaled_df), size = floor(0.8*nrow(scaled_df)), replace = F)
train <- scaled_df[samples,]
test <- scaled_df[-samples,]

rf <- randomForest(tender_isCorrupt ~ ., train, ntree = 500, mtry = 5, )
pred_rf_scores <- predict(rf, test)
threshold <-  0.5 # We set threshold for obtained scores
rf_predict_boolean <- ifelse(pred_rf_scores >= threshold,1,0)

rf_cm <- confusionMatrix(as.factor(rf_predict_boolean), as.factor(test$tender_isCorrupt))
rf_cm

data.frame( R2 = R2(rf_predict_boolean, test$tender_isCorrupt),
          RMSE = RMSE(rf_predict_boolean, test$tender_isCorrupt),
          MAE = MAE(rf_predict_boolean, test$tender_isCorrupt))

print(paste("Recall of Random Forest is:", rf_cm$byClass["Recall"]))
print(paste("Precision of Random Forest is:", rf_cm$byClass["Precision"]))
print(paste("F1 of Random Forest is:", rf_cm$byClass["F1"]))

roc_rf <- roc(test$tender_isCorrupt, rf_predict_boolean)
plot.roc(roc_rf, col="orange")
text(0.36, 0.53, labels=sprintf("AUC(RF)): %0.3f", auc(roc_rf)), col="orange")


```

```{r}
#XGBOOST
set.seed(101)
samples <- sample.int(n = nrow(scaled_df), size = floor(0.8*nrow(scaled_df)), replace = F)
train <- scaled_df[samples,]
test <- scaled_df[-samples,]
```

```{r}
matrix_train <- xgb.DMatrix(as.matrix(train %>% select(-tender_isCorrupt)), label = train$tender_isCorrupt)
matrix_test <- xgb.DMatrix(as.matrix(test %>% select(-tender_isCorrupt)), label = test$tender_isCorrupt)
```

```{r}
watchlist <- list(train = matrix_train, cv = matrix_test)

params <- list(
  "objective" = "binary:logitraw", #Specify the learning task and the corresponding learning objective.
  "eval_metric" = "auc"
)

model_xgb <- xgb.train(params=params,
```

```
                data=matrix_train,
                maximize=TRUE, # larger the evaluation score the better
                nrounds=50, # max number of boosting iterations
                nthread=3, # number of threads
                early_stopping_round=10, # stop if the performance doesn't improve for 10 rounds
                watchlist = watchlist, # named list of xgb.DMatrix datasets to use for evaluating model performance
                print_every_n=5) # prints metrics for every 5th iteration
```

Variable importance

```{r}
importance <- xgb.importance(colnames(matrix_train), model = model_xgb)
xgb.ggplot.importance(importance)
```

Ploting AUC

Let's see how AUC changes with next iterations of model training

```{r}
melted <- melt(model_xgb$evaluation_log, id.vars="iter")
ggplot(data=melted, aes(x=iter, y=value, group=variable, color = variable)) + geom_line()
```

Prediction

Now let's predict fraud using test data:

```{r}
xgb_predict_scores <- predict(model_xgb, matrix_test)
threshold <-  0.5 # We set threshold for obtained scores
xgb_predict_boolean <- ifelse(xgb_predict_scores >= threshold,1,0)
```

And evaluate the result by creating confusion matrix:

```{r}
xgb_cm <- confusionMatrix(as.factor(xgb_predict_boolean), as.factor(test$tender_isCorrupt))
xgb_cm
```

We can access different metrics from our confusion matrix.

```{r}
print(paste("Recall of XGBoost is:", xgb_cm$byClass["Recall"]))
print(paste("Precision of XGBoost is:", xgb_cm$byClass["Precision"]))
print(paste("F1 of XGBoost is:", xgb_cm$byClass["F1"]))
```

We can also create ROC and find AUC

```{r}
roc_xgb <- roc(test$tender_isCorrupt, predict(model_xgb, matrix_test, type = "prob"))
plot.roc(roc_xgb)
text(0.36, 0.53, labels=sprintf("AUC(XGB)): %0.3f", auc(roc_xgb)), col="black")
```

Result

Let's compare all 4 AUC's and select the best one.

```{r}
plot.roc(roc_glm, col="green")
text(0.36, 0.53, labels=sprintf("AUC(GLM)): %0.3f", auc(roc_glm)), col="green")

lines(roc_dt, col="orange")
text(0.36, 0.33, labels=sprintf("AUC(DT)): %0.3f", auc(roc_dt)), col="orange")
```

```
lines(roc_xgb, col="black")
text(0.36, 0.43, labels=sprintf("AUC(XGB): %0.3f", auc(roc_xgb)), col="black")

lines(roc_rf, col="red")
text(0.36, 0.23, labels=sprintf("AUC(RF): %0.3f", auc(roc_rf)), col="red")

lines(roc_dt2, col="blue")
text(0.36, 0.13, labels=sprintf("AUC(DT_CV): %0.3f", auc(roc_dt2)), col="blue")

lines(roc_glm2, col="purple")
text(0.36, 0.03, labels=sprintf("AUC(GLM_CV): %0.3f", auc(roc_glm2)), col="purple")
```

Dataset with TOP10 most important features

```{r}
important_df <- scaled_df %>%
  select(tender_finalPrice, bid_price, tender_description_length, tender_economicRequirements_length, tender_technicalRequirements_length,
tender_personalRequirements_length, tender_lots_count, tender_indicator_INTEGRITY_SINGLE_BID, tender_ indicator_PROCEDURE_TYPE,
tender_publications_count, tender_isCorrupt)
```

#Models with dataset with just important features
```{r}
set.seed(101)
samples <- sample.int(n = nrow(important_df), size = floor(0.8*nrow(important_df)), replace = F)
train_imp <- important_df[samples,]
test_imp <- important_df[-samples,]
#Logistic Regression with Top10 features
glm_imp <- glm(tender_isCorrupt ~ ., train_imp, family=binomial)
pred_glm_scores_imp <- predict(glm_imp, test_imp)
threshold <-  0.5 # We set threshold for obtained scores
glm_predict_boolean_imp <- ifelse(pred_glm_scores >= threshold,1,0)

glm_cm_imp <- confusionMatrix(as.factor(glm_predict_boolean_imp), as.factor(test_imp$tender_isCorrupt))
glm_cm_imp

data.frame( R2 = R2(glm_predict_boolean_imp, test_imp$tender_isCorrupt),
          RMSE = RMSE(glm_predict_boolean_imp, test_imp$tender_isCorrupt),
          MAE = MAE(glm_predict_boolean_imp, test_imp$tender_isCorrupt))

print(paste("Recall of Logistic Regression is:", glm_cm_imp$byClass["Recall"]))
print(paste("Precision of Logistic Regression is:", glm_cm_imp$byClass["Precision"]))
print(paste("F1 of Logistic Regression is:", glm_cm_imp$byClass["F1"]))

roc_glm_imp <- roc(test_imp$tender_isCorrupt, glm_predict_boolean_imp)
plot.roc(roc_glm_imp, col="orange")
text(0.36, 0.53, labels=sprintf("AUC(GLM_IMP)): %0.3f", auc(roc_glm_imp)), col="orange")

```

```{r}
set.seed(101)
# Define train control for k fold cross validation
train_control <- trainControl(method="cv", number=10, savePredictions = TRUE, classProbs = TRUE)
# Fit Logistic Regression Model with Top10 features
modelGLM_imp <- train(tender_isCorrupt~., data=train_imp, trControl=train_control, method="glm")
# Summarise Results
print(modelGLM_imp)
#GLM with Cross Validation
pred_glm_scores2_imp <- predict(modelGLM_imp, test_imp)
threshold <-  0.5 # We set threshold for obtained scores
glm_predict_boolean2_imp <- ifelse(pred_glm_scores2_imp >= threshold,1,0)

glm_cm2_imp <- confusionMatrix(as.factor(glm_predict_boolean2_imp), as.factor(test_imp$tender_isCorrupt))
glm_cm2_imp

data.frame( R2 = R2(glm_predict_boolean2_imp, test_imp$tender_isCorrupt),
          RMSE = RMSE(glm_predict_boolean2_imp, test_imp$tender_isCorrupt),
```

```
        MAE = MAE(glm_predict_boolean2_imp, test_imp$tender_isCorrupt))

print(paste("Recall of Logistic Regression is:", glm_cm2_imp$byClass["Recall"]))
print(paste("Precision of Logistic Regression is:", glm_cm2_imp$byClass["Precision"]))
print(paste("F1 of Logistic Regression is:", glm_cm2_imp$byClass["F1"]))

roc_glm2_imp <- roc(test_imp$tender_isCorrupt, glm_predict_boolean2_imp)
plot.roc(roc_glm2_imp, col="orange")
text(0.36, 0.53, labels=sprintf("AUC(GLM_IMP with CV)): %0.3f", auc(roc_glm2_imp)), col="orange")
```

```{r}
#Decision Tree with Top10 features
set.seed(101)
samples <- sample.int(n = nrow(important_df), size = floor(0.8*nrow(important_df)), replace = F)
train_imp <- important_df[samples,]
test_imp <- important_df[-samples,]

tree_imp <- rpart(tender_isCorrupt~., train_imp)
pred_tree_scores_imp <- predict(tree_imp, test_imp)

threshold <-  0.5 # We set threshold for obtained scores
tree_predict_boolean_imp <- ifelse(pred_tree_scores_imp >= threshold,1,0)

data.frame( R2 = R2(tree_predict_boolean_imp, test_imp$tender_isCorrupt),
        RMSE = RMSE(tree_predict_boolean_imp, test_imp$tender_isCorrupt),
        MAE = MAE(tree_predict_boolean_imp, test_imp$tender_isCorrupt))
tree_cm_imp <- confusionMatrix(as.factor(tree_predict_boolean_imp), as.factor(test_imp$tender_isCorrupt))
tree_cm_imp

print(paste("Recall of Decision Tree is:", tree_cm_imp$byClass["Recall"]))
print(paste("Precision of Decision Tree is:", tree_cm_imp$byClass["Precision"]))
print(paste("F1 of Decision Tree is:", tree_cm_imp$byClass["F1"]))

roc_dt_imp <- roc(test_imp$tender_isCorrupt, predict(tree_imp, test_imp))
plot.roc(roc_dt_imp, col="green")
text(0.36, 0.53, labels=sprintf("AUC(DT_IMP)): %0.3f", auc(roc_dt_imp)), col="green")
```

```{r}
#Decision Tree with Cross Validation and Top10 features
set.seed(101)
# Define train control for k fold cross validation
train_control <- trainControl(method="cv", number=10, savePredictions = TRUE, classProbs = TRUE)
# Fit Decision Tree Model
modelDT_imp <- train(tender_isCorrupt~., data=train_imp, trControl=train_control, method="rpart")
# Summarise Results
print(modelDT_imp)

pred_tree_scores2_imp <- predict(modelDT_imp, test_imp)

threshold <-  0.5 # We set threshold for obtained scores
tree_predict_boolean2_imp <- ifelse(pred_tree_scores2_imp >= threshold,1,0)

data.frame( R2 = R2(tree_predict_boolean2_imp, test_imp$tender_isCorrupt),
        RMSE = RMSE(tree_predict_boolean2_imp, test_imp$tender_isCorrupt),
        MAE = MAE(tree_predict_boolean2_imp, test_imp$tender_isCorrupt))
tree_cm2_imp <- confusionMatrix(as.factor(tree_predict_boolean2_imp), as.factor(test_imp$tender_isCorrupt))
tree_cm2_imp

print(paste("Recall of Decision Tree with Cross Validation is:", tree_cm2_imp$byClass["Recall"]))
print(paste("Precision of Decision Tree with Cross Validation is:", tree_cm2_imp$byClass["Precision"]))
print(paste("F1 of Decision Tree with Cross Validation is:", tree_cm2_imp$byClass["F1"]))


roc_dt2_imp <- roc(test_imp$tender_isCorrupt, predict(modelDT_imp, test_imp))
plot.roc(roc_dt2_imp, col="green")
text(0.36, 0.53, labels=sprintf("AUC(DT with CV_IMP)): %0.3f", auc(roc_dt2_imp)), col="green")
```

```

```{r}
library(randomForest)
#Random Forest with Top10 features
set.seed(101)
samples <- sample.int(n = nrow(important_df), size = floor(0.8*nrow(important_df)), replace = F)
train_imp <- important_df[samples,]
test_imp <- important_df[-samples,]
#Random Forest
rf_imp <- randomForest(tender_isCorrupt ~ ., train_imp, mtry=5, ntree=500)
pred_rf_scores_imp <- predict(rf_imp, test_imp)
threshold <-  0.5 # We set threshold for obtained scores
rf_predict_boolean_imp <- ifelse(pred_rf_scores_imp >= threshold,1,0)

rf_cm_imp <- confusionMatrix(as.factor(rf_predict_boolean_imp), as.factor(test_imp$tender_isCorrupt))
rf_cm_imp

data.frame( R2 = R2(rf_predict_boolean_imp, test_imp$tender_isCorrupt),
            RMSE = RMSE(rf_predict_boolean_imp, test_imp$tender_isCorrupt),
            MAE = MAE(rf_predict_boolean_imp, test_imp$tender_isCorrupt))

print(paste("Recall of Random Forest is:", rf_cm_imp$byClass["Recall"]))
print(paste("Precision of Random Forest is:", rf_cm_imp$byClass["Precision"]))
print(paste("F1 of Random Forest is:", rf_cm_imp$byClass["F1"]))

roc_rf_imp <- roc(test_imp$tender_isCorrupt, rf_predict_boolean)
plot.roc(roc_rf_imp, col="orange")
text(0.36, 0.53, labels=sprintf("AUC(RF_IMP)): %0.3f", auc(roc_rf_imp)), col="orange")


```

```{r}
#XGBOOST with Top10 features
set.seed(101)
samples <- sample.int(n = nrow(important_df), size = floor(0.8*nrow(important_df)), replace = F)
train_imp <- important_df[samples,]
test_imp <- important_df[-samples,]
```

```{r}
matrix_train_imp<- xgb.DMatrix(as.matrix(train_imp %>% select(-tender_isCorrupt)), label = train_imp$tender_isCorrupt)
matrix_test_imp <- xgb.DMatrix(as.matrix(test_imp %>% select(-tender_isCorrupt)), label = test_imp$tender_isCorrupt)
```

```{r}
watchlist_imp <- list(train = matrix_train_imp, cv = matrix_test_imp)

params <- list(
  "objective" = "binary:logitraw", #Specify the learning task and the corresponding learning objective.
  "eval_metric" = "auc"
)

model_xgb_imp <- xgb.train(params=params,
                    data=matrix_train_imp,
                    maximize=TRUE, # larger the evaluation score the better
                    nrounds=50, # max number of boosting iterations
                    nthread=3, # number of threads
                    early_stopping_round=10, # stop if the performance doesn't improve for 10 rounds
                    watchlist = watchlist_imp, # named list of xgb.DMatrix datasets to use for evaluating model performance
                    print_every_n=5) # prints metrics for every 5th iteration
```

Ploting AUC

Let's see how AUC changes with next iterations of model training

```{r}

```
melted_imp <- melt(model_xgb_imp$evaluation_log, id.vars="iter")
ggplot(data=melted_imp, aes(x=iter, y=value, group=variable, color = variable)) + geom_line()
```

Prediction

Now let's predict fraud using test data:

```{r}
xgb_predict_scores_imp <- predict(model_xgb_imp, matrix_test_imp)
threshold <-  0.5 # We set threshold for obtained scores
xgb_predict_boolean_imp <- ifelse(xgb_predict_scores_imp >= threshold,1,0)
```

And evaluate the result by creating confusion matrix:

```{r}
xgb_cm_imp <- confusionMatrix(as.factor(xgb_predict_boolean_imp), as.factor(test_imp$tender_isCorrupt))
xgb_cm_imp
```

We can access different metrics from our confusion matrix.

```{r}
print(paste("Recall of XGBoost is:", xgb_cm_imp$byClass["Recall"]))
print(paste("Precision of XGBoost is:", xgb_cm_imp$byClass["Precision"]))
print(paste("F1 of XGBoost is:", xgb_cm_imp$byClass["F1"]))

```

We can also create ROC and find AUC

```{r}
roc_xgb_imp <- roc(test_imp$tender_isCorrupt, predict(model_xgb_imp, matrix_test_imp, type = "prob"))
plot.roc(roc_xgb_imp)
text(0.36, 0.53, labels=sprintf("AUC(XGB_IMP)): %0.3f", auc(roc_xgb_imp)), col="black")
```

Result

Let's compare all 4 AUC's and select the best one.

```{r}
plot.roc(roc_glm_imp, col="green")
text(0.36, 0.53, labels=sprintf("AUC(GLM_IMP): %0.3f", auc(roc_glm_imp)), col="green")

lines(roc_dt_imp, col="orange")
text(0.36, 0.33, labels=sprintf("AUC(DT_IMP): %0.3f", auc(roc_dt_imp)), col="orange")

lines(roc_xgb_imp, col="black")
text(0.36, 0.43, labels=sprintf("AUC(XGB_IMP): %0.3f", auc(roc_xgb)), col="black")

lines(roc_rf_imp, col="red")
text(0.36, 0.23, labels=sprintf("AUC(RF_IMP): %0.3f", auc(roc_rf_imp)), col="red")

lines(roc_dt2_imp, col="yellow")
text(0.36, 0.13, labels=sprintf("AUC(DT_IMP_CV): %0.3f", auc(roc_dt2_imp)), col="yellow")

lines(roc_glm2_imp, col="purple")
text(0.36, 0.03, labels=sprintf("AUC(GLM_IMP_CV): %0.3f", auc(roc_glm2_imp)), col="purple")

```

Appendix 3 – Descriptive statistics final data

```
tender_estimatedDurationInDays tender_finalPrice   tender_description_length tender_personalRequirements_length
Min.   :   1.0                 Min.   :      102   Min.   :   1.0            Min.   :  107
1st Qu.:   4.0                 1st Qu.:    17415   1st Qu.:  60.0            1st Qu.: 2184
Median :  15.0                 Median :    36036   Median : 110.0            Median : 2662
Mean   :  39.1                 Mean   :   363123   Mean   : 208.4            Mean   : 2866
3rd Qu.:  36.0                 3rd Qu.:    88777   3rd Qu.: 225.0            3rd Qu.: 3440
Max.   :2283.0                 Max.   :636824025   Max.   :3819.0            Max.   :68527
tender_economicRequirements_length tender_technicalRequirements_length tender_awardCriteria_count tender_lots_count
Min.   :  93.0                     Min.   :   75                       Min.   : 1.000            Min.   : 1.000
1st Qu.: 292.0                     1st Qu.:  457                       1st Qu.: 1.000            1st Qu.: 1.000
Median : 411.0                     Median :  812                       Median : 1.000            Median : 1.000
Mean   : 552.7                     Mean   : 1524                       Mean   : 1.346            Mean   : 4.592
3rd Qu.: 578.0                     3rd Qu.: 1647                       3rd Qu.: 1.000            3rd Qu.: 3.000
Max.   :9482.0                     Max.   :40341                       Max.   :82.000            Max.   :95.000
tender_publications_count tender_indicator_INTEGRITY_SINGLE_BID
Min.   : 1.000            Min.   :  0.00
1st Qu.: 2.000            1st Qu.:  0.00
Median : 2.000            Median :100.00
Mean   : 2.631            Mean   : 68.89
3rd Qu.: 3.000            3rd Qu.:100.00
Max.   :36.000            Max.   :100.00
tender_indicator_INTEGRITY_CALL_FOR_TENDER_PUBLICATION tender_indicator_INTEGRITY_ADVERTISEMENT_PERIOD
Min.   :  0.00                                         Min.   :  0.00
1st Qu.:  0.00                                         1st Qu.:  0.00
Median :100.00                                         Median :100.00
Mean   : 60.78                                         Mean   : 63.19
3rd Qu.:100.00                                         3rd Qu.:100.00
Max.   :100.00                                         Max.   :100.00
tender_indicator_INTEGRITY_PROCEDURE_TYPE tender_indicator_INTEGRITY_DECISION_PERIOD
Min.   :  0.00                            Min.   :  0.00
1st Qu.:100.00                            1st Qu.:  0.00
Median :100.00                            Median :100.00
Mean   : 93.65                            Mean   : 69.59
3rd Qu.:100.00                            3rd Qu.:100.00
Max.   :100.00                            Max.   :100.00
tender_indicator_ADMINISTRATIVE_COVERED_BY_GPA tender_indicator_ADMINISTRATIVE_FRAMEWORK_AGREEMENT
Min.   :  0.00                                 Min.   :  0.000
1st Qu.:  0.00                                 1st Qu.:  0.000
Median :  0.00                                 Median :  0.000
Mean   : 12.41                                 Mean   :  9.912
3rd Qu.:  0.00                                 3rd Qu.:  0.000
Max.   :100.00                                 Max.   :100.000
tender_indicator_ADMINISTRATIVE_NOTICE_AND_AWARD_DISCREPANCIES   lot_row_nr     lot_bidsCount       bid_row_nr
Min.   : 8.602                                                  Min.   : 1.000   Min.   :  1.000   Min.   : 1.000
1st Qu.:69.231                                                  1st Qu.: 1.000   1st Qu.:  1.000   1st Qu.: 1.000
Median :75.000                                                  Median : 1.000   Median :  3.000   Median : 1.000
Mean   :71.661                                                  Mean   : 2.767   Mean   :  3.871   Mean   : 1.496
3rd Qu.:81.818                                                  3rd Qu.: 1.000   3rd Qu.:  5.000   3rd Qu.: 2.000
Max.   :98.113                                                  Max.   :95.000   Max.   :135.000   Max.   :52.000
  bid_price         tender_isCorrupt
Min.   :      100   Min.   :0.0000000
1st Qu.:    12951   1st Qu.:0.0000000
Median :    27040   Median :0.0000000
Mean   :   231918   Mean   :0.0005842
3rd Qu.:    61341   3rd Qu.:0.0000000
```

Figure 7. Descriptive statistics for the final data, compiled by author, source: Appendix 2

Appendix 4 – Descriptive statistics initial data

```
##  tender_estimatedDurationInMonths tender_estimatedDurationInDays
##  Min.   :  0.00                   Min.   :   1.00
##  1st Qu.: 18.00                   1st Qu.:   3.00
##  Median : 24.00                   Median :  12.00
##  Mean   : 30.54                   Mean   :  24.87
##  3rd Qu.: 37.00                   3rd Qu.:  24.00
##  Max.   :300.00                   Max.   :2283.00
##  NA's   :96023                    NA's   :51380
##  tender_estimatedPrice tender_finalPrice   tender_description_length
##  Min.   :    125       Min.   :     102   Min.   :   1.0
##  1st Qu.:   65100       1st Qu.:    17188   1st Qu.:  59.0
##  Median :  210000       Median :    35280   Median : 110.0
##  Mean   : 1226396       Mean   :   304779   Mean   : 208.2
##  3rd Qu.:  634000       3rd Qu.:    84036   3rd Qu.: 224.0
##  Max.   :95867473       Max.   :636824025   Max.   :3819.0
##  NA's   :97787          NA's   :9817        NA's   :5597
##  tender_personalRequirements_length tender_economicRequirements_length
##  Min.   :  107                      Min.   :  93.0
##  1st Qu.: 2226                      1st Qu.: 294.0
##  Median : 2697                      Median : 423.0
##  Mean   : 2947                      Mean   : 588.3
##  3rd Qu.: 3587                      3rd Qu.: 618.0
##  Max.   :68527                      Max.   :9482.0
##  NA's   :47057                      NA's   :55393
##  tender_technicalRequirements_length tender_awardCriteria_count
##  Min.   :   75                       Min.   : 1.000
##  1st Qu.:  461                       1st Qu.: 1.000
##  Median :  844                       Median : 1.000
##  Mean   : 1692                       Mean   : 1.339
##  3rd Qu.: 1770                       3rd Qu.: 1.000
##  Max.   :40341                       Max.   :82.000
##  NA's   :48547                       NA's   :8064
##  tender_lots_count tender_publications_count
##  Min.   : 1.000    Min.   : 1.000
##  1st Qu.: 1.000    1st Qu.: 2.000
##  Median : 1.000    Median : 2.000
##  Mean   : 4.596    Mean   : 2.631
##  3rd Qu.: 3.000    3rd Qu.: 3.000
##  Max.   :95.000    Max.   :36.000
##  NA's   :2249
##  tender_indicator_INTEGRITY_SINGLE_BID
##  Min.   :  0.00
##  1st Qu.:  0.00
##  Median :100.00
##  Mean   : 68.44
##  3rd Qu.:100.00
##  Max.   :100.00
##  NA's   :15077
##  tender_indicator_INTEGRITY_CALL_FOR_TENDER_PUBLICATION
##  Min.   :  0.00
##  1st Qu.:  0.00
##  Median :100.00
##  Mean   : 60.78
##  3rd Qu.:100.00
##  Max.   :100.00
##  NA's   :1
##  tender_indicator_INTEGRITY_ADVERTISEMENT_PERIOD
##  Min.   :  0.00
##  1st Qu.:  0.00
##  Median :100.00
##  Mean   : 63.19
##  3rd Qu.:100.00
##  Max.   :100.00
##  NA's   :7
```

```
##  tender_indicator_INTEGRITY_PROCEDURE_TYPE
##  Min.   :  0.00
##  1st Qu.:100.00
##  Median :100.00
##  Mean   : 93.65
##  3rd Qu.:100.00
##  Max.   :100.00
##  NA's   :1
##  tender_indicator_INTEGRITY_DECISION_PERIOD
##  Min.   :  0.00
##  1st Qu.:  0.00
##  Median :100.00
##  Mean   : 70.35
##  3rd Qu.:100.00
##  Max.   :100.00
##  NA's   :6972
##  tender_indicator_INTEGRITY_TAX_HAVEN
##  Min.   :  0.00
##  1st Qu.:100.00
##  Median :100.00
##  Mean   : 99.67
##  3rd Qu.:100.00
##  Max.   :100.00
##  NA's   :71638
##  tender_indicator_ADMINISTRATIVE_COVERED_BY_GPA
##  Min.   :  0.00
##  1st Qu.:  0.00
##  Median :  0.00
##  Mean   : 12.17
##  3rd Qu.:  0.00
##  Max.   :100.00
##  NA's   :3260
##  tender_indicator_ADMINISTRATIVE_FRAMEWORK_AGREEMENT
##  Min.   :  0.000
##  1st Qu.:  0.000
##  Median :  0.000
##  Mean   :  9.723
##  3rd Qu.:  0.000
##  Max.   :100.000
##  NA's   :3260
##  tender_indicator_ADMINISTRATIVE_ENGLISH_AS_FOREIGN_LANGUAGE
##  Min.   :  0.00
##  1st Qu.:  0.00
##  Median :  0.00
##  Mean   : 10.27
##  3rd Qu.:  0.00
##  Max.   :100.00
##  NA's   :94564
##  tender_indicator_ADMINISTRATIVE_NOTICE_AND_AWARD_DISCREPANCIES
##  Min.   : 8.60
##  1st Qu.:66.67
##  Median :75.00
##  Mean   :69.37
##  3rd Qu.:76.92
##  Max.   :98.11
##  NA's   :42897
##    lot_row_nr     lot_estimatedPrice  lot_bidsCount
##  Min.   : 1.000   Min.   :      111   Min.   :  1.000
##  1st Qu.: 1.000   1st Qu.:    23250   1st Qu.:  1.000
##  Median : 1.000   Median :    64970   Median :  3.000
##  Mean   : 2.769   Mean   :   839518   Mean   :  3.632
##  3rd Qu.: 1.000   3rd Qu.:   300000   3rd Qu.:  4.000
##  Max.   :95.000   Max.   :636824025   Max.   :135.000
##  NA's   :2250     NA's   :96671       NA's   :21399
##  lot_description_length   bid_row_nr       bid_price
##  Min.   :  3.0          Min.   : 1.000   Min.   :      100
##  1st Qu.:  35.0         1st Qu.: 1.000   1st Qu.:    12493
##  Median :  64.0         Median : 1.000   Median :    24954
##  Mean   : 114.2         Mean   : 1.513   Mean   :   128583
```

```
##  3rd Qu.: 131.0        3rd Qu.: 2.000   3rd Qu.:    51595
##  Max.   :2743.0        Max.   :52.000   Max.   :486966498
##  NA's   :88336         NA's   :14646    NA's   :23810
##  tender_isCorrupt
##  Min.   :0.0000000
##  1st Qu.:0.0000000
##  Median :0.0000000
##  Mean   :0.0005842
##  3rd Qu.:0.0000000
##  Max.   :1.0000000
```

Non-exclusive licence to reproduce thesis and make thesis public

I, Mart Kevin Põlluste,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

DETECTING CORRUPTION IN PUBLIC PROCUREMENT THROUGH OPEN DATA ANALYSIS supervised by Rajesh Sharma.

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.
3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.
4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Mart Kevin Põlluste

15/08/2019