VILNIUS UNIVERSITY

FACULTY OF MATHEMATICS AND INFORMATICS

MODELLING AND DATA ANALYSIS MASTER'S STUDY PROGRAMME

Master's thesis

Unsupervised Anomaly Detection in Value-Added Tax Return Forms

Anomalijų nustatymas pridėtinės vertės mokesčio deklaracijose neprižiūrimo mokymosi metodais

Kamila Prokopovič

Supervisor Dr. Viktor Skorniakov

Vilnius, 2021

Abstract

Value Added Tax (VAT) is an important source of income in most European countries. In Lithuania, VAT accounts for 44 % of all income, and that makes the collection of VAT an important task as it is essential to support state allocations and sustainability. However, there are many ways the VAT taxation system can be exploited. In Lithuania, registered VAT taxpayers are obliged to submit monthly VAT declaration form FR0600. Obligation to submit VAT declarations makes such acquired data a perfect choice for the identification of fraudulent activities. The aim of this thesis is to identify possible VAT fraud cases by applying selected unsupervised anomaly detection methods.In order to reliably identify VAT fraud cases, several tasks need to be addressed, including analyzing, implementing, comparing, and summarizing the fraud detection alternatives that exist in the field of unsupervised learning. Fraud is considered to be rare and markedly different from legitimate observations event; therefore, anomaly detection techniques were considered.

Keywords: Value Added Tax, Anomaly Detection, Unsupervised Learning, Fraud Detection.

Santrauka

Pridėtinės vertės mokestis (PVM) yra svarbus pajamų šaltinis daugumoje Europos šalių. Lietuvoje PVM sudaro 44 % visų pajamų, todėl PVM surinkimas yra svarbi užduotis, nes tai būtina palaikant valstybės asignavimus ir tvarumą. Tačiau yra daugybė būdų, kaip galima išnaudoti PVM apmokestinimo sistemą. Lietuvoje registruoti PVM mokesčių mokėtojai privalo kas mėnesį pateikti PVM deklaracijos FR0600 formą. Dėl pareigos teikti PVM deklaracijas, tokie duomenys yra puikus pasirinkimas nesąžiningai veiklai nustatyti. Šio baigiamojo darbo tikslas yra identifikuoti galimus sukčiavimo PVM atvejus taikant pasirinktus anomalijų nustatymo metodus. Norint patikimai nustatyti PVM sukčiavimo atvejus, reikia išspręsti kelias užduotis, įskaitant išanalizuoti, įgyvendinti, palyginti ir apibendrinti sukčiavimo nustatymo alternatyvas, kurios egzistuoja neprižiūrimo mokymosi srityje. Sukčiavimas laikomas retu ir akivaizdžiai besiskiriančiu nuo įstatyminių atvejų įvykiu, todėl buvo apžvelgti anomalijų nustatymo metodai.

Raktažodžiai: pridėtinės vertės mokestis, anomalijų nustatymas, neprižiūrimas mokymasis, sukčiavimo identifikavimas.

Contents

List of Figures

List of Tables

List of abbreviations

k-NN k-Nearest Neighbor

AD Anomaly Detection

AUC Area Under Curve

CASE Center for Social and Economic Research

EU European Union

iforest Isolation Forest

MCC Matthews Correlation Coefficient

MLC Multilateral Control

MTIC Missing Trader Intra-Community

PCA Principal Component Analysis

ROC Receiver Operating Characteristic Curve

SCAC The Standing Committee on Administrative Cooperation

VAT Value Added Tax

VIES VAT Information Exchange System

VTTL Total VAT Liability

WF Working Field

Introduction

Value Added Tax (VAT) is an important source of income in most European countries[36]. In Lithuania, VAT accounts for 44 % of all income[19], and that makes the collection of VAT an important task as it is essential to support state allocations and sustainability. However, there are many ways the VAT taxation system can be exploited, such as exaggerated purchases, under-reported sales, feigning of foreign sales, failure to register as VAT taxpayer, and many more.

All in Lithuania registered VAT taxpayers are obliged to submit monthly VAT declaration form FR0600 (Appendix C). Obligation to submit VAT declarations makes such acquired data a perfect choice for identification of fraudulent activity, whereas data mining is a promising field that can help gain insights from available data and discover new patterns that may have been overlooked.

Supervised learning requires identified examples of fraud, which means that the results of performed audits are necessary. However, audits are costly and time-consuming, so there are very few investigated and even fewer identified fraud cases in comparison to the general population. Moreover, entities for audits are selected on the basis of expert opinion and experience, which leads to bias in the selection of the sample. In such a sample, although fraud is considered a rare occurrence, the proportion of fraudsters is quite high[9].

To overcome the above-mentioned reasons, unsupervised methods will be explored in this thesis. Fraud is considered to be rare and markedly different from legitimate observations event; therefore, anomaly detection techniques will be considered. However, models can only indicate possible fraud or error but not to guarantee that the highest scores were assigned correctly to fraudsters. The expert in this field should always assess the results. On the other hand, unsupervised methods are capable of discovering unseen patterns which can indicate new, yet unknown to tax authorities forms of fraud schemes.

Aim

Identify possible VAT fraud cases by applying selected unsupervised anomaly detection methods.

Goals

In order to reliably identify VAT fraud cases following tasks need to be addressed:

- Review literature on VAT fraud identification.
- Define input variables based on VAT domain knowledge.
- Analyze and preprocess raw data to achieve more accurate results.
- Generate outliers to make the evaluation of unsupervised models possible.
- Compare and summarize results obtained by different anomaly detection algorithms.
- Identify the algorithm the most suitable to detect fraudulent activities.

1 VAT and VAT fraud

The concept of value-added tax (VAT) was presented in the 1960s to replace a complex set of historical indirect taxes. The aim was to facilitate cross-border trade and provide guidelines for creating a single market in Europe. VAT was introduced in Lithuania on the 1st of May 1994. Since the 1st of July 2002, the new Value Added Tax of the Republic of Lithuania Law amendments have entered into force, which has implemented all the fundamental provisions of the European Union (EU) legal acts regulating VAT taxation; however, the law did not transpose the requirements which can only be applied to a member of the European Union and single market. The amendment to the Law on Value Added Tax of the Republic of Lithuania was adopted on the 15th of January 2004, which entered into force on the 1st of May 2004. This law definitively transposed the legal acts of the European Union regulating the VAT taxation procedure and provisions. The VAT object is the supply of goods and services effected for consideration within the territory of the country when these goods and services are provided by a taxable person engaged in economic activity. VAT object is also the acquisition of goods for consideration within the territory of the country from another Member State. The object of import VAT is the import of goods when the goods are considered to have been imported into the territory of the country[20].

1.1 VAT

VAT in the European Union is a general, widely based consumption tax calculated on the value-added of goods and services. EU law only requires that a standard VAT rate would be of at least 15 % and a reduced rate of at least five %. Exemptions are also possible when zero VAT taxation is allowed[32]. VAT is levied on the majority of goods and services that are purchased and sold for use or consumption in the European Union. Goods or services exported to other countries of the Community are not subject to VAT[18]. Imports are taxed so that the system is fair for EU producers so that they could compete on the European market with suppliers outside the EU on an equal footing[53].

VAT - and indirect tax, which applies to the created added value. VAT payers are final consumers of goods and services who purchase them for personal purposes and do not use them in further commercial activities. VAT is paid at each stage of the production or supply chain[47].

Since the 1st of September 2009, the standard VAT rate in Lithuania is 21 %. Figure 1 shows an example of VAT calculation: the producer manufactures goods that are sold to a distributor for €100 + 21 % VAT. A distributor pays a producer €121, of which €21 the producer must pay to the state budget. In the next step, the distributor sells those goods to a seller for €200 + €42 VAT. The distributor receives €42 VAT, of which €21 pays to the state budget and the remaining €21 previously paid to the producer when purchased goods. A buyer pays €484 for the same goods, of which €84 is VAT. The seller pays €42 VAT to the state budget; the rest was paid during a transaction between the seller and the distributor. This way, the entire VAT amount is paid by the final consumer, even though VAT is paid at each stage.

VAT payers in Lithuania are natural and legal persons who carry out economic activities of any kind in

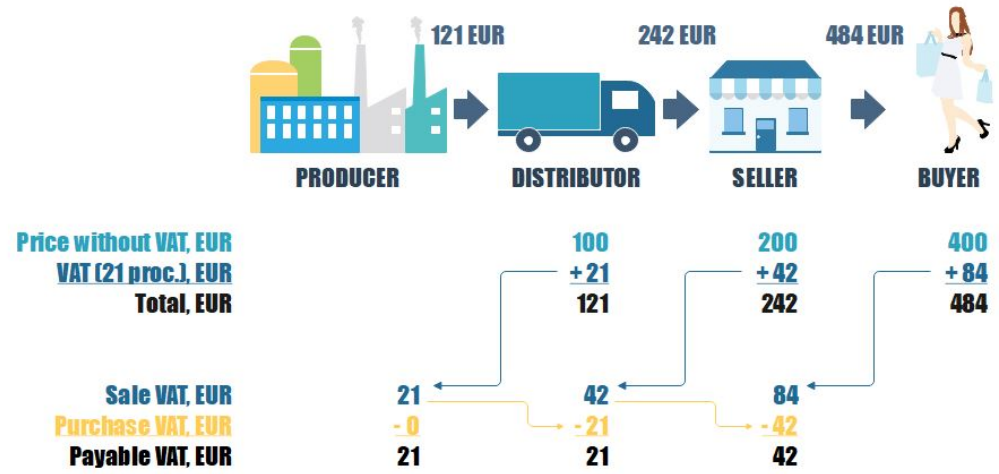


Figure 1: VAT scheme

Lithuania, as well as a collective investment undertaking established in the Republic of Lithuania without the status of a legal person, whose the form of activity is an investment fund [2].

VAT is one of the easiest ways to supplement the country's budget [40]. The tax base is very broad, and the tax collection is not very complex; therefore, the value-added tax is widely used in Europe. As shown in Figure 2, in Lithuania, VAT accounts for the largest share of the country's total budget revenue. The collection of VAT is a key component of the budget for maintaining public allocations and sustainability.

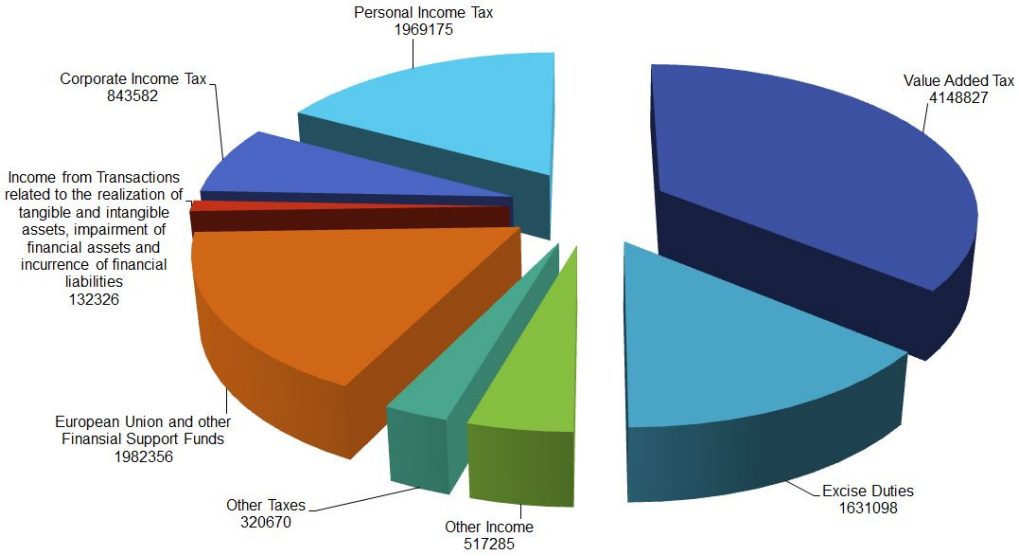


Figure 2: Lithuania budget for 2020 [19]

In the European Union, the VAT system is strictly regulated, given that this tax, as a consumption tax, has an essential influence on the creation of an internal market without borders. The main document governing the taxation of VAT at the European Union level is the Council Directive 2006/112/EC [32] of the 28th of November 2006 on the common system of value-added tax. This Directive regulates practically all aspects of

the application of VAT [20]. The EU has a few directives regulating specific subjects:

- Cases and procedure for applying for VAT exemptions on imports - Council Directive 2006/79/EC of the 5th of October 2006 on the exemption from taxes of imports of small consignments of goods of a non-commercial character from third countries and Council Directive 2009/132/EC of the 19th of October 2009 determining the scope of Article 143(b) and (c) of Directive 2006/112/EC as regards exemption from value-added tax on the final importation of certain goods.
- Return of VAT to foreign taxable persons - Council Directive 2008/9/EC of the 12th of February 2008 laying down detailed rules for the refund of value-added tax, provided for in Directive 2006/112/EC, to taxable persons not established in the Member State of refund but established in another Member State, also partially amending it Regulation 66/2010 of the European Parliament and of the Council of the 25th of November 2009 and Thirteenth Council Directive 86/560/EEC of the 17th of November 1986 on the harmonization of the laws of the Member States relating to turnover taxes - arrangements for the refund of value-added tax to taxable persons not established in Community territory [20].

The common VAT rules aim to prevent larger international companies from gaining a competitive advantage by selling goods from countries with lower VAT rates. For example, if a country applies significantly higher VAT rates than in the other Member States of the Community, this could negatively affect the companies of that country when competitors establish in another country with lower VAT rates and because of that are able to offer goods or services at lower prices. The rules also simplify the operation of smaller companies in a single market outside their own country.

On the 1st of January 1993, the single market was established in the European Union, border controls within the Community were abolished, and the Community members adopted the common EU VAT rules [47]. Under the adopted legislation, the principle of taxation in the country of destination has entered info force for cross-border transactions, that is, a purchaser acquiring goods or services that are subject to zero rate VAT from another Community country, later sells them within a country VAT taxed; if goods or services are exported, a seller acquires the basis for recovering the VAT paid [12]. Such a VAT taxation model has loopholes that are used to create complex VAT fraud schemes.

1.2 VAT Fraud

1.2.1 Fraud Types

From a budget revenue collection point, VAT can be regarded as a successful tax; however, like the other taxes, VAT is not protected from evasion and fraud. Various economic agents seek to avoid taxation [22]. There are various types of fraud [36]:

- Undeclared sales. A trader does not declare part or all of the sales. This group includes companies and individuals providing services to the end-user, such as beauty and aesthetic procedures or finishing and repair works of buildings or other structures. Usually, the added value at the last stage of the VAT chain

is significantly higher than the VAT costs; therefore, the choice is not to issue invoices.

- Avoidance of registration. Persons whose remuneration reaches the VAT registration threshold avoid registering as VAT payers and do not submit accountancy of the value-added tax payable by a person or entity not registered as a value-added taxpayer in order to avoid VAT liability.
- Incorrect classification of goods. VAT payable is reduced by declaring sales of goods or services under the guise of other goods or services which are subject to reduced or zero rate VAT.
- Excluded manufactured goods or services. Goods or services produced in the company and consumed by the owner or employees of that company are not declared, and taxes are not paid.
- Collected but not paid taxes. The collected VAT may not be paid to the state budget in a number of ways: through incorrect accounting, declaring bankruptcy, or by getting involved in missing trader in the Community schemes;
- Non-payment of VAT on importation. Imported goods are sold at a zero VAT rate domestically on the black market.

Some fraud mechanisms are specific to VAT[36]:

- Illegal VAT refund claims. The whole VAT taxation principle makes it possible to apply for VAT refund by presenting forged invoices for non-existent or exaggerated purchases, which allow the recovery of allegedly paid VAT.
- VAT reduction. In one company producing goods that are subject to VAT and which are exempted, it is possible to take advantage of the situation and to allocate costs to taxable goods in order to obtain a VAT refund (VAT can not be refunded if the goods are not taxable).
- Fictitious traders. Businesses can be set up solely to generate invoices that allow deduction of VAT.

Four VAT fraud driving forces can be distinguished[23]:

- Increased supply of high-value low weight goods.
- Legislation allowing importers not to pay VAT on purchases from another EU country.
- Legislation allowing exporters to reclaim VAT paid on purchases of goods although the tax administrator has not received contributions in the previous steps of the chain.
- Restrictions related to the prohibition on impeding the free movement of goods and the slowing of inspection procedures, which leads to the review of transactions only after some period of time.

1.2.2 Missing Trader Intra-Community Fraud Schemes

The Member States of the European Union lose billions of euros every year due to VAT evasion or unfair claims to refund VAT from national authorities[39]. The biggest damage is caused by missing trader schemes[47] that fall into collected but not paid to the state budget fraud category. Missing trader fraud scheme, or more precisely Missing Trader Intra-Community (MTIC) fraud scheme by Europol is defined as VAT theft from authorities carried out by organized criminal groups[1]. These are very complex schemes that use VAT rules stating that cross-border transactions within the EU are not subject to VAT.

Several conditions are necessary for the missing trader scheme to function[31]:

- the goods supplied must be subject to VAT;
- the transported goods must reach the country of destination;
- the supply and sale of goods must be carried out as part of economic activity;
- the buyer and seller must be registered for VAT.

Four entities are usually involved in cross-border fraud schemes[21]:

- Missing Trader - a company registered for VAT fraud purposes that carry out or imitates cross-border VAT-exempt transactions in order to profit from subsequent VAT-taxable transactions. When the time comes to pay VAT to the tax administrator missing trader disappears.
- Buffer Company - a company that operates as an ordinary trader in the domestic market, that buys and/or sells goods or services and pays taxes. In the fraudulent chain buffer company misleads supervising authorities by creating an image of a fair business.
- Broker - the last piece in the fraudulent chain located in the same country as the missing trader. Broker purchases goods or services from buffer company and sells them to another member state of the Community with a right to a VAT refund.
- Conduit Company - trader that fictively or actually supplies goods to a company in another member state of the Community which does not pay taxes.

The main missing trader fraud models are fraud by purchasing goods, carousel fraud, and contra trading[52].

1.3 VAT Gap

It is important for the member states of the European Union to assess the extent of VAT fraud and the resulting financial losses. The difference between the VAT collected and the total VAT liability (VTTL) defined by law, the so-called VAT Gap, shows how much VAT is not paid because not all taxpayers honestly declare and pay the due taxes. It is important to note that VAT fraud is not an equivalent of a VAT Gap; it is just one of the components. The Gap calculations also cover VAT losses due to insolvency, bankruptcies, administrative errors, and tax optimization[48]. The VAT Gap can be assessed in three main ways[34]:

- Top-down - using generalized statistical methods. Accuracy of the estimates Gap is determined by the accuracy and completeness of the national accounts data used in the calculations.
- Bottom-up - using data from VAT declaration forms.
- Econometric methods such as frontier analysis or time series analysis. Econometric methods are sensitive to the choice of assumptions and factors, therefore, are not recommended in VAT Gap estimation.

Every year since 2013, the Center for Social and Economic Research (CASE) has been preparing a report for the European Commission assessing the top-down VAT Gap for each EU country. Due to data availability, the calculations are performed with a two-year lag, so currently, only calculations for 2018 are available. VAT

2017 and 2018 Gaps as a percentage of the total VAT liability for EU countries are presented in Figure 3.

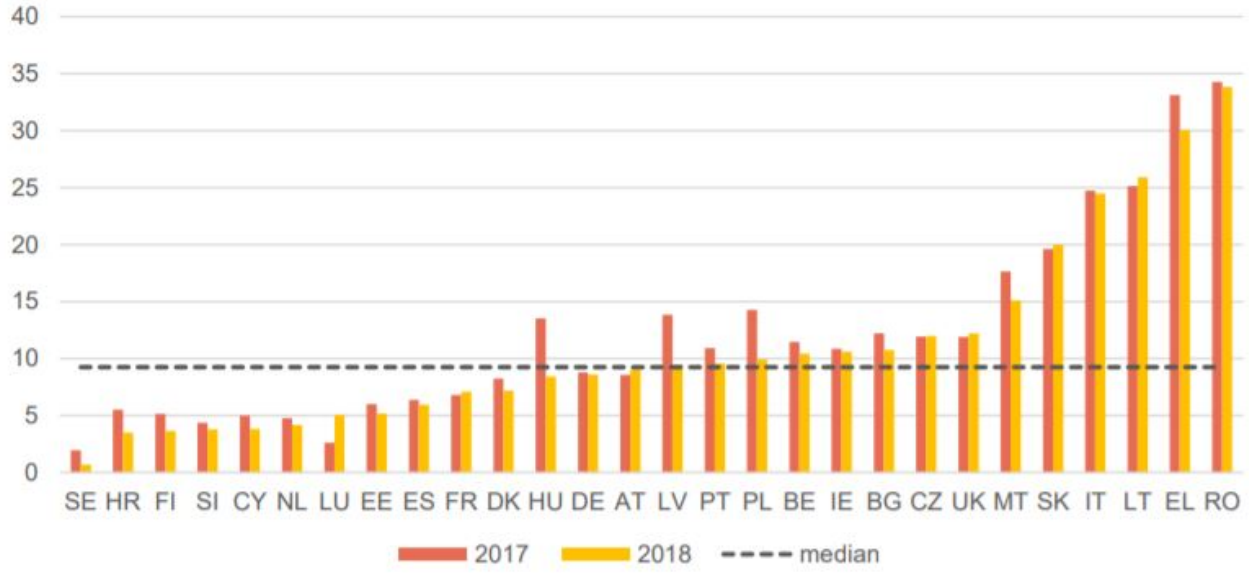


Figure 3: VAT gap top-down[48]

In 2018 in comparison to 2017, the VAT Gap increased in seven countries, mostly in Luxembourg (LU, 2.5 %), Lithuania (LT, 0.8 %), and Austria (AT, 0.5 %). The biggest decrease in VAT Gap was observed in Hungary (HU, 5.1 %), Latvia (LV, 4.4 %), and Poland (PL, 4.3 %). The smallest Gap is in Sweden (SE, 0.7 %), while the largest in Romania (RO, 33.8 %). The median VAT Gap was 9.2 %. In 2018 European Union lost €140 billion in VAT[48]. For more details see Appendix D.

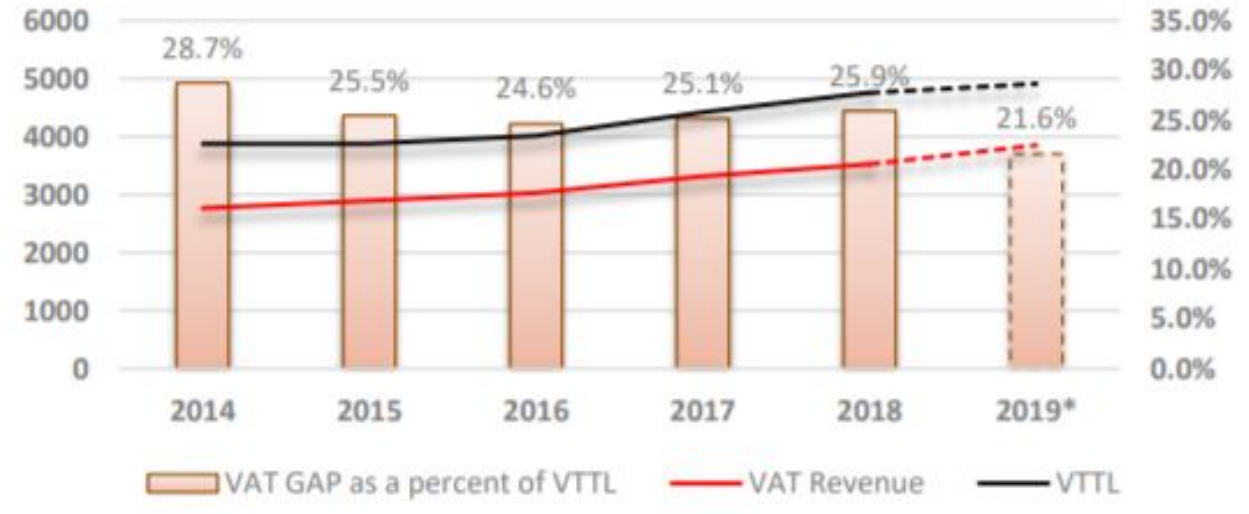


Figure 4: Lithuanian VAT Gap[48]

Lithuania has one of the largest VAT Gap - 25.9 %, which accounts for €1.2 billion lost VAT revenue. As shown in Figure 4 the Gap has not changed much since 2015 and fluctuates around 25 %. However, CASE predicts that the Gap estimate will significantly fall in 2019.

1.4 Measures to Fight VAT Fraud

Well-organized groups by taking advantage of gaps in legislation misappropriate unpaid taxes. To prevent revenue losses, member states are seeking to develop new modern surveillance tools. The supervisory authorities and other competent authorities or working groups assess crimes and try to anticipate and prevent the spread of VAT fraud.

1.4.1 Legal Regulation

On the 16th of March 2010, European Union member states adopted Council Directive 2010/24/EU concerning mutual assistance for the recovery of claims relating to taxes, duties, and other measures. It defines rules for recovery of all taxes, including VAT from other Community countries. The possibility of receiving support from several EU institutions, bodies, and committees in the fight against fraud was also introduced [3]. The main assistance means to be taken by member states that have transposed this Directive are:

- To transmit documents or necessary information related to tax crimes, as well as to assist in the investigation of crimes.
- To participate in the courts or tax authorities of another state, to investigate and interview suspects.
- to share information with third parties if required.

In accordance with the 904/2010 regulation adopted on the 7th of October 2010, European Union member states undertook to exchange relevant information on the calculation and collection of VAT. Key aspects:

- At the request of member states, it is mandatory to exchange all information that may help in the assessment of VAT transactions.
- Automatic exchange of information should take place where are obvious grounds to believe that VAT laws are violated, or the risk of tax loss exists.
- Information requested should be exchanged as soon as possible but no later than three months from the date of the received request.
- Exchange of information should be carried out through VIES electronic database.
- Rapid exchange of targeted information on suspicious traders and similar problems should take place through the Eurofisc system.

E-commerce package adopted on the 5th of December 2017 [54] aims at improving the collection of VAT in the distance sales sector. The changes in this package relate to business-to-consumer (B2C) intra-Community distance sales and B2C imports from third countries of less than €150. The new e-commerce package is being implemented in two stages:

- In the first stage, since 2019, VAT rules have been facilitated for small and medium-sized businesses providing telecommunications, broadcasting, and electronic services. Service providers with a cross-border B2C turnover of less than €10 000 may apply the VAT rules of the country in which they are established. The use of the €10 000 thresholds is not mandatory; that is, a service provider can choose to charge VAT

on the services in the country of the purchaser.

- In the second stage, in 2021, the VAT exempt threshold of €22 for transactions with third countries will be abolished in order to harmonize conditions for EU and non-EU VAT payers. Imported into EU goods up to €150 will be subject to VAT by filling in an electronic simplified customs declaration. Goods exceeding €150 will continue to be subject to normal customs procedure.

VAT fraudsters have a direct impact on the EU's financial interests; therefore, members of the Community undertook to criminalize VAT fraud under the Directive 2017/1371 [?] of the European Parliament and of the Council on the fight against fraud to the Union's financial interests by means of criminal law adopted on the 5th of July 2017. Community members[39]:

- Undertake to criminalize intentional fraud affecting the European Union's financial interests.
- Infringements of the common VAT system consider being aggravated if fraud schemes seeking to benefit from the common VAT system involve two or more countries and caused damaged accounts for at least €10 000 000.
- May continue to apply administrative measures and penalties in the field covered by this Directive but must ensure that the criminal sanctions, administrative measures, and penalties provided in this Directive do not infringe EU Charter of Fundamental Rights.
- Cooperate with each other and with the EU institutions in providing technical and operational assistance, sharing available information in order to combat VAT fraud.

1.4.2 Administrative Cooperation

In fraud schemes, a number of countries are being involved, so cooperation between tax administrations is necessary to curb tax fraud. Due to the existing VAT system, it is necessary to have information on transactions in other member states in order to collect VAT in one's own country. Currently, there are a few tools for cross-border administrative cooperation: VIES, Eurofisc, SCAC, and multilateral controls. Figure 5 shows the classification of administrative cooperation systems according to the speed and detail of information provision.

VIES

VIES (VAT Information Exchange System) - is a computerized system that makes it easy and quick to check whether a purchaser in another Community country is a taxable person and has a valid VAT identification number[55]. The European Commission maintains a page where the validity of the VAT code can be checked: https://ec.europa.eu/taxation_customs/vies/vieshome.do?selectedLanguage=en

Eurofisc

Eurofisc - is an early warning system launched in 2010 to strengthen administrative cooperation within the Community in the fight against VAT fraud. Eurofisc consists of the Eurofisc Group, which identifies working fields, evaluates the effectiveness of the system and six working fields (WF), each specializing in one of the

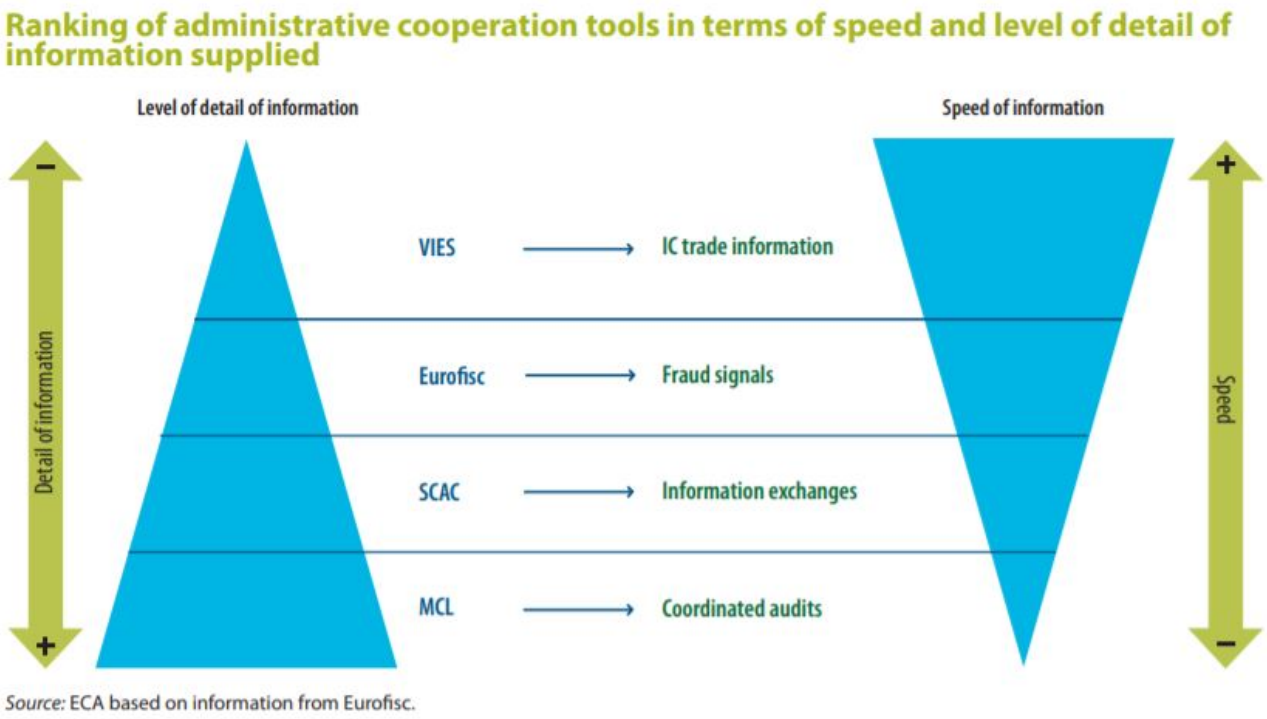


Figure 5: Classification of administrative cooperation tools[8]

following areas[46]:

- WF1: missing trader intra-Community fraud schemes.
- WF2: fraud related to vehicles (cars, ships, and planes).
- WF3: fraud related to abuse of customs procedure 42.
- WF4: monitoring trends and changes in VAT fraud.
- WF5: fraud related to cross-border e-commerce.
- WF6: implementation of Transaction Network Analysis (TNA).

All Community countries are connected to the Eurofisc network but can choose the working fields they are interested in. By joining the working field, the party undertakes to take an active part, that is to exchange information and provide feedback[46].

SCAC

SCAC (The Standing Committee on Administrative Cooperation) - The Standing Committee on Administrative Cooperation that works under Article 58 of Regulation 904/2010.

Multilateral Control

Multilateral Control (MLC) - a coordinated inspection of one or more related traders of the Community, if a member state considers that such an inspection is more effective than an inspection carried out by a single state. During the multilateral control, members of the Community agree that auditors from other countries can observe but not actively participate in administrative investigations. Such coordinated inspections allow for a

faster exchange of targeted information.

1.4.3 Analytical Tools

Tax administrations typically use rule-based systems to identify cases of fraud. Such systems report possible fraud when a case is detected that complies with defined rules. The rules are drawn up by experts on the basis of historical cases of fraud, and their own experience[16]. Those rule-based systems can only identify cases similar to those that have occurred in the past, for which the mechanism has been found out, and the relevant risk rules described.

With the increasing amount of information stored in databases and the complexity of fraud schemes, a need for automation of fraud identification and more quick responses to changes emerged. New methods are being used to detect tax fraud. Data mining techniques that help to better understand hidden patterns and identify useful relationships are becoming more dominant. Data mining is the process of analyzing big data to detect previously unknown patterns in the data. This process consists of five main steps[43]:

1. Problem definition, goal setting.
2. Assessment of the adequacy of the data.
3. Definition of rules, data preparation.
4. Algorithm selection, modeling.
5. Evaluation of the results.

Two main types of data mining are used when searching for financial fraud structures. First, Supervised Learning, when data used is labeled, that is, it has a pre-identified class. For example, in fraud investigation, one class is identified as fraud, another class of not fraudulent observations. Second, Unsupervised Learning, when data is not labeled. A hybrid approach also exists, so-called Semi-Supervised Learning. In this case, a small proportion of instances is labeled, and the rest is not.

1.4.3.1 Supervised Learning

In a literature review conducted by Albashrawi M. on financial fraud of different types such as financial statement, credit card, insurance, etc., was summarized that the supervised learning techniques are the most widely used. Leading methods in this category were logistic regression, neural networks, and decision trees[6].

Logistic Regression

Logistic regression is a statistical method used to estimate the probability of a dependent variable falling into one of the categories. Probability values range from 0 to 1, so in the case of binary regression, the dependent variable acquires a value of 0 when the calculated probability is below the selected threshold and a value of 1 otherwise.

Neural Networks

A neural network is a multilayer network of interconnected neurons. The neural network visualized in Figure 6 consists of:

- Input layer that receives information.
- Connections weights which show how much impact a neuron has on neurons in the layer.
- One or more hidden layers where calculations take place.
- Output layer, which presents results as a probabilistic estimate.

The input layer receives information that travels further in the direction of arrows through the hidden layers where the data is being processed to the output layer. The neural network calculates the losses between the obtained prediction estimates and the known outputs. The network learns a selected number of iterations minimizing network loss each time and changing connection weights. Such a neural network does not need rules defined in advance; it learns from the data provided.

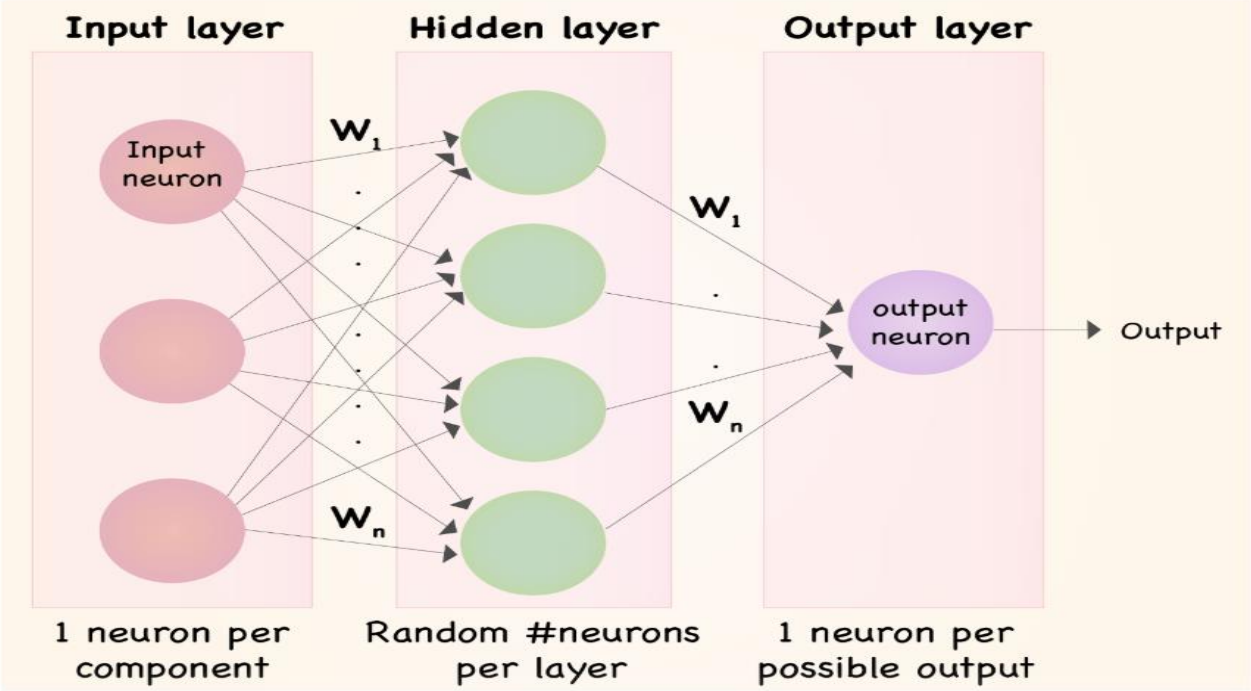


Figure 6: Neural Network[24]

Decision Trees

Decision tree is a tree-shaped structure as shown on Figure 7 containing:

- Root which stands for the entire population.
- Nodes, each representing a particular attribute or property.
- Branches connecting nodes in accordance with the rules laid down.
- Leaves picturing the results.

Decision trees are used to solve classification tasks when the dependent variable is categorical and predic-

tion tasks when the dependent variable is quantitative. The purpose of those trees is to group data in such a way that the observations on one node (leaf) are as similar as possible and the observations on different nodes (leaves) are as different as possible[38].

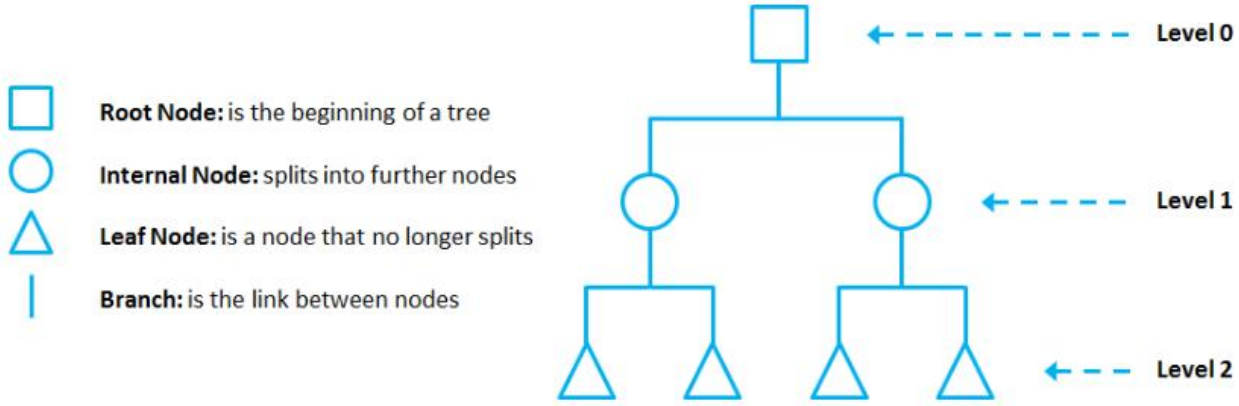


Figure 7: Decision Tree[38]

1.4.3.2 Unsupervised Learning

According to the aforementioned review by Albashrawi M., a few unsupervised learning techniques were also used to identify various fraud cases, for example, k-means clustering and self-organizing maps[6].

K-Means Clustering

K-Means clustering is an algorithm that allows finding clusters in data, that is, groups of similar observations. Using this method, it is necessary to redefine the value of k - the number of centroids denoting the centers of clusters[35]. Visually clustering is presented in Figure 8.

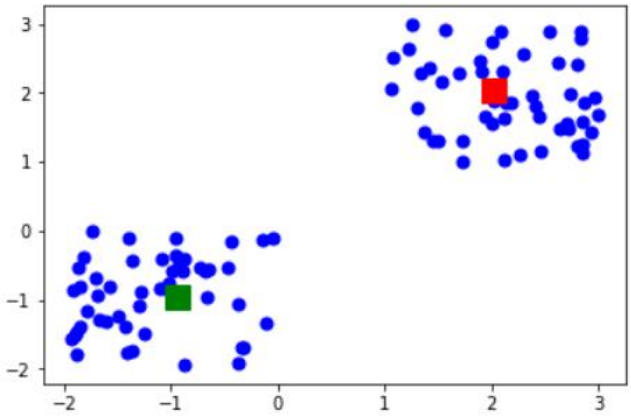


Figure 8: K-Means clustering[35]

Self-Organizing Maps

Self-organizing neural networks are one of the types of neural network that reduces the dimension of the data.

Multidimensional data is projected so that similar observations on the map would be next to each other as pictured in Figure 9.

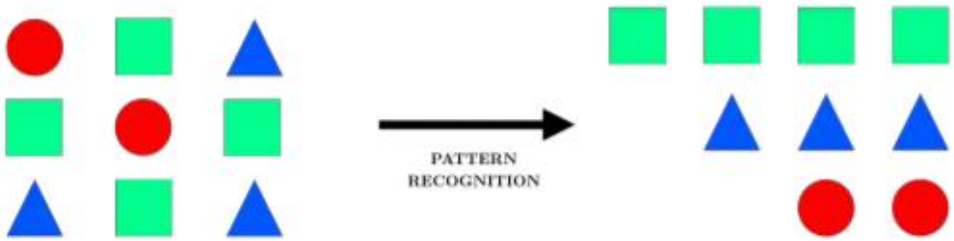


Figure 9: Self-Organizing Map [37]

2 Anomaly Detection

Anomaly detection is the process of finding unusual, significantly different data points in a data set. Anomaly is also referred to as outlier or abnormality, and in this work will be used interchangeably. By the definition given by Grubbs[30]: "An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs." At first, anomaly detection was important as a data cleansing step because many algorithms are sensitive to anomalies. Nowadays, outliers themselves are interesting because they contain information about characteristics that depart from the norm. A few domains where identification of unusual data points is useful are[4]:

- Fraud detection: Fraud can occur in different fields such as credit card fraud, insurance claim fraud, tax fraud, etc. No matter the field, the goal is to identify criminal activity to prevent financial losses.
- Medical diagnosis: Various disease conditions can be interpreted as anomalies in medical applications covering data obtained from an electrocardiogram, magnetic resonance imaging, computed tomography, and other devices. As this is a very sensitive area, high accuracy is required.
- Intrusion detection: Intrusions, in other words, malicious activities as anomalies are interesting in computer systems and security. Many different system activities are being constantly monitored to check if systems are working properly. It is important to notice as early as possible all kinds of abnormal behavior that can happen in the form of break-ins, attacks, or other abuse.

2.1 General Information on Anomaly Detection

2.1.1 Types of Anomalies

Anomalies is a data point that significantly differs from other observations. However, in practice, it is not easy to state firmly, which points are anomalous. Figure 10 illustrates an example of different types of anomalies that can be categorized into[13]:

- Point Anomaly - a single instance of data is anomalous if its value differs greatly from the rest. Most of the research done on anomaly detection focuses on this type of anomalies as it is the simplest and most common one. Point anomaly is illustrated in Figure 10a. As a use case scenario can be mentioned the detection of credit card fraud based on the amount spent in a transaction. The point anomaly would be an unusually high amount that would fall outside of the expenditure range in comparison to what this specific individual would typically spend in a transaction.
- Contextual Anomaly - is based on a context; it is common in time-series data. A point which is considered as an anomaly by taking time into account is illustrated in Figure 10b. A case scenario related to credit card fraud would appear when an individual spends high amounts of money during unusual times of the year. An individual who typically spends any given amount spent five times as much, so it does not conform with the typical behavior portrayed by this person. Not every sudden increase in expenditures would be an anomaly as there are times of the year like Christmas where these increments are normal; therefore, the

choice to applying a contextual anomaly detection strategy would be determined by the existing context of the specific domain.

- Collective Anomaly - a set of data instances are anomalous with respect to the entire data set. Usually explored in a sequence data and is illustrated in Figure 10c. An example of a collective anomaly can be a potential cyber-attack where someone tries to retrieve data from a remote to a local machine without any authorization.

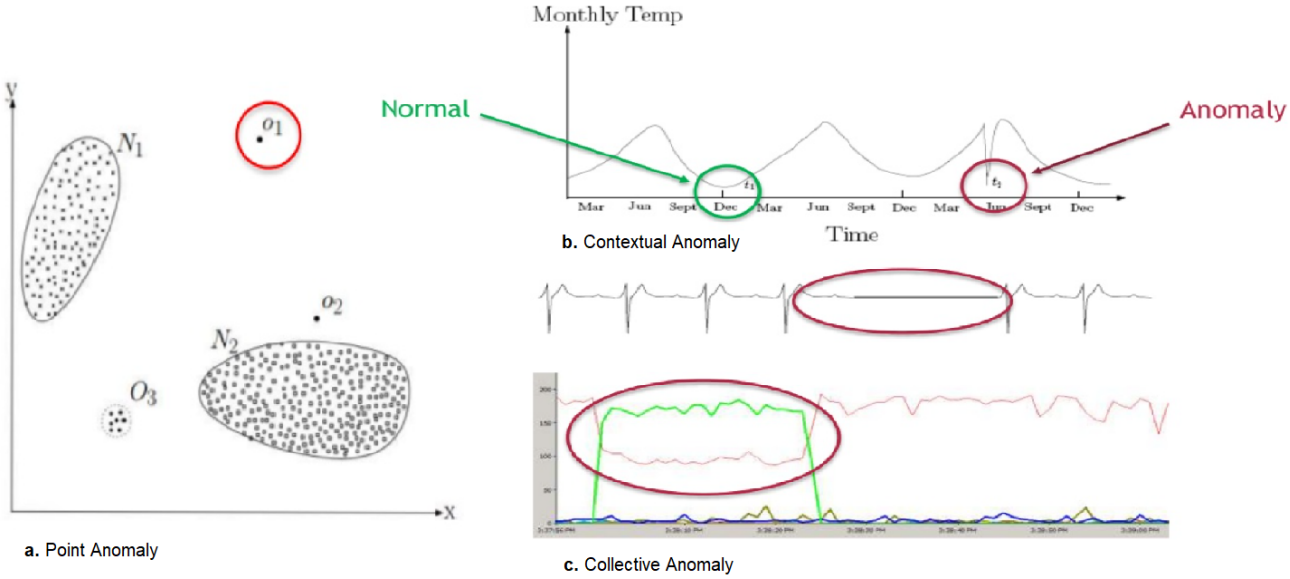


Figure 10: Anomaly Detection Types[11]

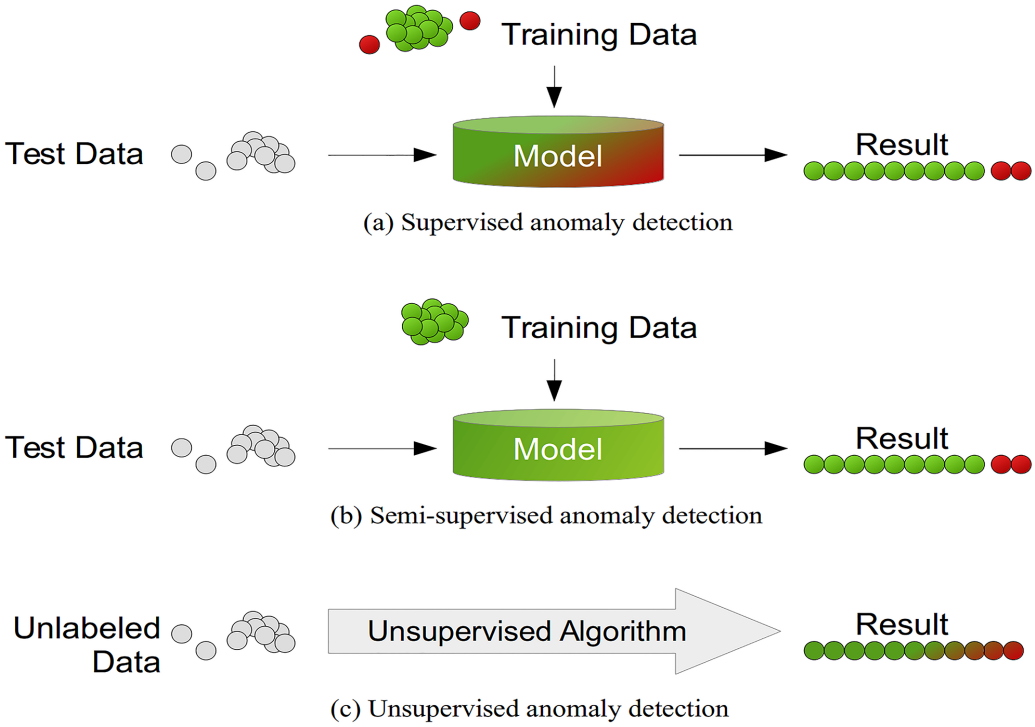
2.1.2 Anomaly Detection Modes

Anomaly detection modes depend on the labels available and are divided as follows[13]:

- Supervised Anomaly Detection (Figure 11a) - mode where both training and testing data sets are fully labeled. Although this scenario is researched the most, it is not very relevant in practice due to two reasons: unbalanced data and difficulties obtaining labels for anomalous instances. Typically, a predictive model is constructed that distinguishes between two classes, normal and anomalous, and predictions are made on test data.
- Semi-Supervised Anomaly Detection (Figure 11b) - mode where training data comprises of only one labeled class. It is difficult to obtain data set that consists of all possible anomaly cases as usual anomalies are not known; therefore, the more common approach is to train a model using only observations that are considered normal. Afterward, such a model is able to identify unseen cases that deviate from the norm it was trained on.
- Unsupervised Anomaly Detection (Figure 11c) - mode where no difference exists between training and testing data as labels are unavailable. The model works under the assumption that anomalies rarely occur in data and can be detected based on intrinsic characteristics of a data set only. This is the most flexible

approach; however, it is challenging to evaluate the model.

Altogether, supervised anomaly detection methods are used in an application-specific manner while unsupervised techniques to identify anomalous observations that raise suspicion. Generally, results obtained from unsupervised methods have to be examined further by a field specialist[4].



(a) Supervised anomaly detection

(b) Semi-supervised anomaly detection

(c) Unsupervised anomaly detection

Figure 11: Anomaly Detection Modes[27]

2.1.3 Anomaly Detection Outputs

Two types of outputs can be returned by anomaly detection methods[13]:

- Scores - assigned values that quantify how anomalous a data point is. This type of output allows to rank instances based on their score in order to select those with the highest degree of abnormality.
- Labels - assigned class whether a data point is an anomaly or not and is derived by imposing a specific threshold. Categorical output carries less information than scoring; thus, the possibility to select only a chosen number of most anomalous instances is not available.

2.1.4 Challenges in Anomaly Detection

Challenges to look out for in anomaly detection[5]:

- A technique suitable to detect anomalies well in different domains does not exist because abnormal behavior is domain-specific.
- It is difficult to draw a line between an actual anomaly and noise which data contains.
- Most often labeled data is not available because of high labeling cost.

- Anomalous behavior can change over time, or new cases of abnormal behavior can arise.
- Fraudsters may try to imitate behavior which is considered normal; thus, they may appear as legitimate.

2.2 Anomaly Detection in Tax Fraud Domain

Anomaly detection methods aim to detect abnormal instances, whereas tax fraud is considered to be a rare and anomalous phenomenon. Thus, tax fraud fulfills two main characteristics anomalies possess[56]:

- Anomalies in a data set occur far less than normal instances.
- Anomalies with respect to their features representation markedly differ from the rest of the data.

Castellón González & Velásquez's[29] analysis of known fraud cases showed that values of the features describing fraudulent instances usually occur among the extreme values, and this provides another basis for investigating anomaly detection methods in order to identify abnormal behavior.

In the field of tax, the acquisition of labeled data for identified VAT fraud cases is difficult because of costly and time-consuming investigations conducted by tax authorities. In order to identify fraud, documents must be reviewed, transactions verified, accounts checked, interviews conducted, etc. Because of the large quantities of gathered data and limited resources of tax administration, only a small fraction of all entities can be investigated. Aforementioned reasons lead to the following issues[56]:

- Audits results can be outdated as investigations take a long time.
- Audited sample is really small in comparison to the entire population of operating entities. Only less than 1 % are usually investigated.
- Audited sample is biased as tax authorities conduct investigations only in case of potentially high risk and avoid sacrificing resources on compliant entities.
- Due to the selection bias audited sample includes a bigger proportion of fraudsters than it is expected in the population.

Even if a label sample is available, it is not representative of the population; therefore, unsupervised anomaly detection methods should be considered. Another important reason to choose an unsupervised algorithm is the dynamic nature of the field. Entities are looking for ways to avoid paying taxes; therefore, fraudsters exploiting the system are always ahead of the tax authorities. While public authorities are trying to find fraudulent scheme patterns, fraudsters are already using new legal or illegal ways to evade payments[50]. Supervised methods can only be applied to detect new instances of known patterns, and as a result, authorities are always staying behind fraudsters. To confront this problem, unsupervised methods that look for newly emerged patterns without preliminary assumptions is a good choice.

Worth noting that a high score assigned by anomaly detection techniques is only an indication of potentially fraudulent activity that needs to be analyzed further and can not be considered ground truth.

2.3 Anomaly Detection Methods

2.3.1 Models Selection

A few different ways of classifying unsupervised anomaly detection methods are found in literature, of which the most common are: nearest-neighbor based, clustering-based, statistical, or subspace. A few methods, like neural networks or support vector machines, can not be assigned to any of roughly established groups [27].

- Nearest Neighbor methods are based on the assumption that anomalous points lie far from the closest neighbors while nonanomalous points are gathered together. The distance of every data point is measured to its k-th neighbors, where k is the number of chosen neighbors, and this is regarded as the anomaly score [4].
- Clustering-based methods group all data points into clusters of similar points. Those points that do not belong to any identified clusters are considered anomalous. Also, the distance of every point to the nearest cluster centroid can be measured. Normal data points lie close to the centroids thus have low anomaly score while anomalous points acquire high score [13].
- Statistical methods fit the statistical model to all data points. Statistical inference tests are used to determine whether a point is anomalous or not based on calculated probability. Normal data points have a high probability of belonging to the learned model, while anomalous points have a low probability [13].

Unsupervised anomaly detection methods were chosen based on their differences and sources, claiming that they are well suited to identify anomalies. The comprehensive study by Goldstein & Uchida [27] was shown that among 19 unsupervised algorithms aimed at detecting anomalies in 10 different data sets, nearest neighbor methods perform better in the majority of cases. The best performing algorithm in general with respect to the accuracy, the stability of scoring, sensitivity, and computational time was k-NN. This provided a reasonable basis for including k-NN in this study. Another selected method is Isolation forest based on results obtained by Domingues et al. [17] during the comparative evaluation of 14 different unsupervised anomaly algorithms on 15 data sets. In this study, robustness, precision, computation time, and memory consumption were evaluated. Although Isolation Forest did not perform the best on both of the latter, it achieved excellent performance in precision on many data sets, making it a reliable choice in many cases. With the growing popularity of deep learning, autoencoders were chosen as a representative of this field. This is a relatively new approach to detect anomalies that were evaluated in Nolle et al. [45] research where autoencoders were compared with seven state-of-the-art outlier detection methods and have proven to be worth attention for detecting anomalous events. Unfortunately, due to a lack of extensive research in unsupervised tax fraud detection, models were selected based on their overall performance rather than in a unique field of tax fraud.

2.3.2 k-Nearest Neighbor Anomaly Detection

k-nearest neighbor anomaly detection, hereafter k-NN AD, is an unsupervised anomaly detection method on the contrary to well known k-nearest neighbor, which is used to solve classification tasks in a supervised manner. In

k-NN AD, basically, for every observation in a data set, distances to k-neighbors are calculated, and anomaly scores are assigned. Two options, which differ in a number of neighbors, are available that is to take into account only one nearest neighbor or to estimate the average distance to the k-nearest neighbors[27]. An experiment conducted by Goldstein et al. was demonstrated that k-NN AD averaging over k-nearest neighbors performs better[28].

In Figure 12 k-NN is visualized in two-dimensional space. Bubble size corresponds to anomaly score, and color illustrates whether the data point is anomalous red or normal green.

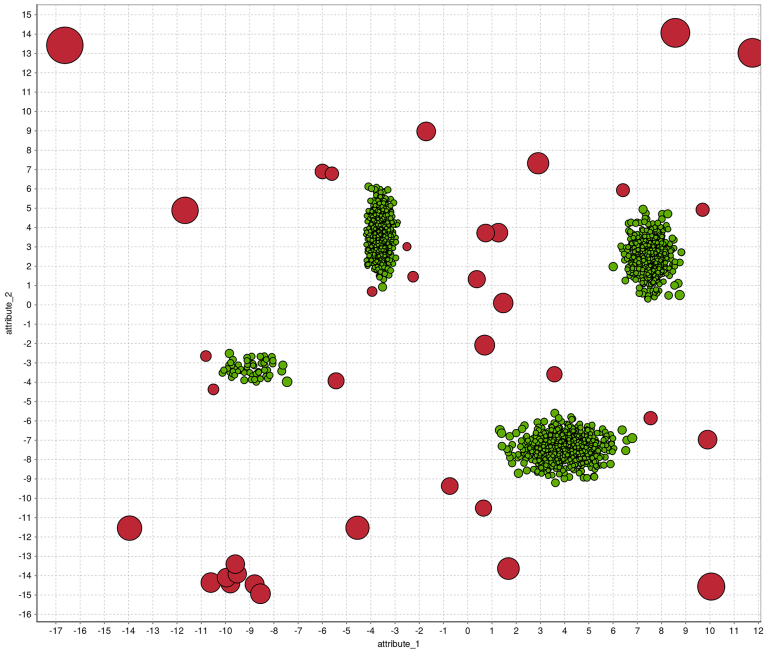


Figure 12: k-Nearest Neighbor Anomaly Detection Visualization[27]

2.3.3 Isolation Forest

Isolation Forest is an ensemble of isolation trees. In an isolation tree, the data is partitioned recursively at randomly selected partition points of randomly chosen features with cuts parallel to axes. The idea behind this method is to isolate anomalous points instead of profile normal points lies in the properties of anomalies. They are different and sparse, so they are more susceptible to isolation; thus, paths from the root to the leaves of such points are shorter than the normal data points. The path length is used to assign the anomaly score to each observation[41].

In Figure 13 is presented how partitioning is executed. The graph on the left has a longer path in a tree structure as isolation of normal point requires more partitions. The graph on the right isolating anomalous point; therefore, a number of partitions are small as an instance with features deviating from the norm is easier to separate.

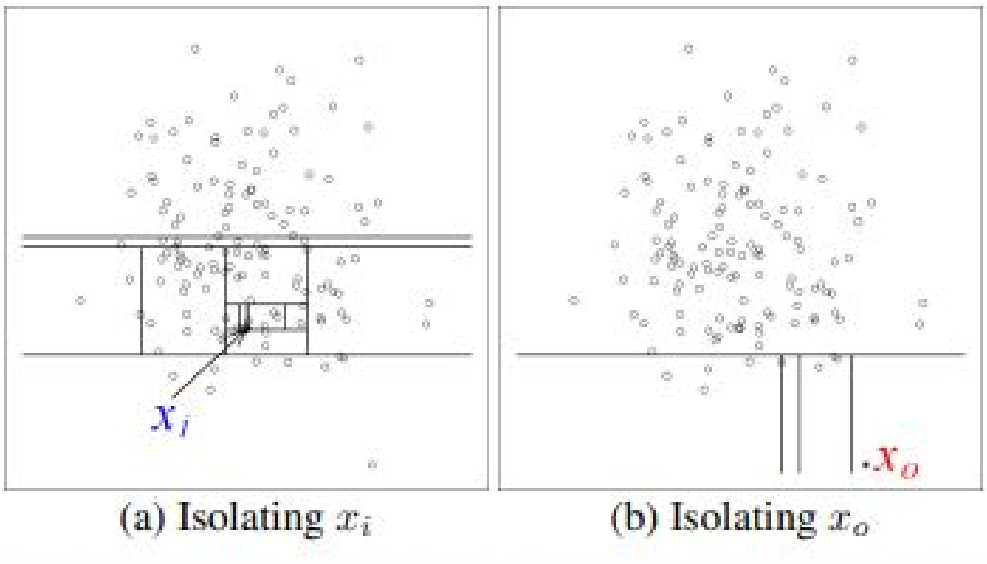


Figure 13: Isolation Forest [41]

2.3.4 Autoencoder Neural Network

Autoencoder is a type of artificial neural network that compresses data into a representation of a lower dimension and then learns to reconstruct the data back using only the compressed representation of the input data. Reconstructed data is not exactly the same as the original input because the noise in the data is eliminated. The goal of such a neural network is to learn to generalize but not to memorize precisely. As depicted on Figure 14 autoencoder has three main components:

- Encoder - compresses the input into lower dimension, encoded representation.
- Bottleneck - contains the compressed representation of the lowest dimension.
- Decoder - reconstructs the input only from compressed representation.

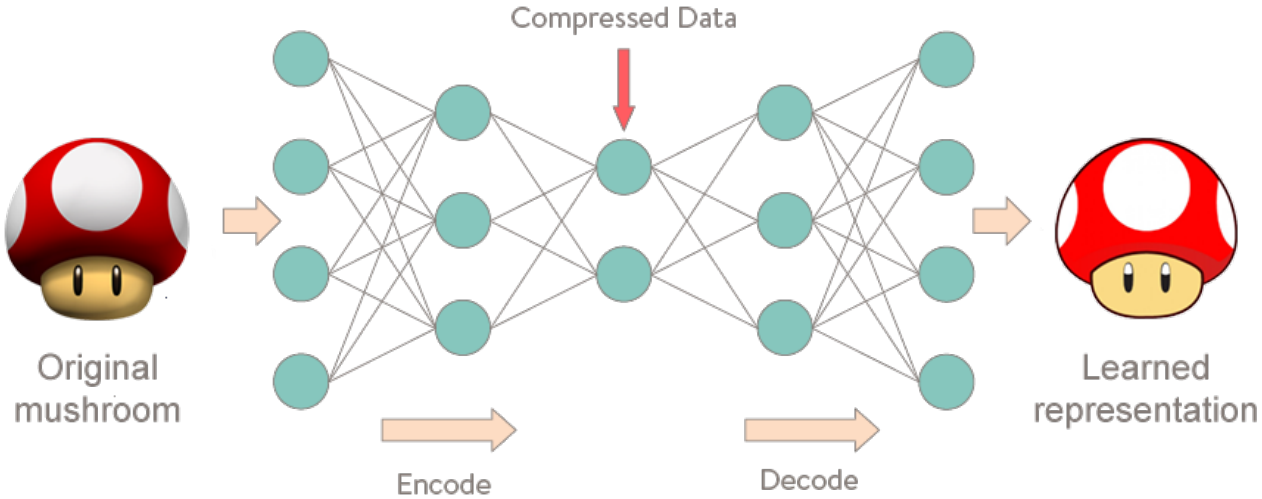


Figure 14: Autoencoder Neural Network [15]

A neural network learns by minimizing reconstruction loss, that is, by measuring how close is the recon-

structed output to the original input. As the majority of observations in the data set are normal, autoencoder learns to reconstruct them well, and reconstruction loss is small. In the case of anomalous entries, patterns due to their rarity are not learned, and reconstruction loss of those observations is higher in comparison to the majority of entries.

3 Data Analysis

3.1 Raw Data

The Lithuanian State Tax Inspectorate provided with fully anonymized data set of grids from VAT declaration FR0600 form (Appendix C). Submission of VAT declaration is obligatory to all in Lithuania VAT registered taxpayers and has to be filled up monthly. Some exceptions are possible when it is allowed to change tax period and to submit forms on a quarterly or half-yearly basis; however, those instances are quite rare in comparison to the entire population of taxpayers and will not be analyzed. The provided data set consists of almost 1.5 million instances during the period of 2018 - 2019.

Vanhoeyveld et al.[56] claim that fraudsters, in order to appear as honest taxpayers, may alternate values in VAT declaration forms. It is more difficult to correctly modify all grids in the form; therefore, the combinations of those values are used rather than raw submitted values. Thirty-one variables are manually derived from VAT form based on VAT domain knowledge, and it is presumed that the combination of them may indicate a committed tax fraud. Due to the sensitive nature of data, explanations of the variables will not be disclosed, and they are encoded as **VAR01 - VAR31**. Also, additional columns such a taxpayer ID, sector the company is active in, and time in months how long a taxpayer is registered as VAT payer are present.

In the data set presented in Table 1 only two variables are categorical **ID** and **Sector** while all other variables are continuous.

	ID	Sector	TP_VAT_age	VAR01	VAR02	VAR03	VAR04	VAR05	VAR06	VAR07	...	VAR27	VAR28	VAR29	VAR30	VAR31
0	00001	X	4	12298.00	nan	nan	nan	12298.00	12298.00	2582.00	...	3369.67	153.09	7.28	21.00	13.72
1	00001	X	5	8176.00	nan	nan	nan	8176.00	8176.00	1717.00	...	-3437.00	78.62	4.56	31.27	26.71
2	00001	X	6	9543.00	nan	nan	nan	9543.00	9543.00	2004.00	...	6041.76	503.52	16.83	21.00	4.17
3	00001	X	7	9074.00	nan	nan	nan	9074.00	9074.00	1906.00	...	-1668.71	81.12	-4.88	21.01	25.89
4	00001	X	8	9301.00	nan	nan	nan	9301.00	9301.00	913.00	...	183.29	89.56	-13.63	9.82	23.45
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1481978	72409	F	52	nan	nan	nan	nan	nan	nan	nan	...	-51677.05	nan	nan	nan	nan
1481979	72409	F	53	nan	nan	nan	nan	nan	nan	nan	...	-56219.95	nan	nan	nan	nan
1481980	72409	F	54	161233.00	nan	nan	nan	161233.00	161233.00	28361.00	...	100912.52	528.80	16.06	20.03	3.97
1481981	72409	F	55	95041.00	nan	nan	nan	95041.00	95041.00	19959.00	...	37510.90	251.84	16.74	25.08	8.34
1481982	72409	F	56	92277.00	nan	nan	nan	92277.00	92277.00	15877.00	...	72233.67	1833.32	16.16	17.31	1.15

Table 1: Example of data set with variables derived from VAT declaration form

Due to varying legal bases and dissimilar market conditions, there are differences between entities operating in different sectors. Consequently, fraudulent behavior also differs. State Tax Inspectorate, having limited resources every year, selects priority sectors to focus on. This year the most attention was paid to construction, used cars trade, repair and trade of cars parts, dental services, e-commerce and catering, and rural tourism[7]. Keeping in mind those differences, set priorities by tax authorities Health Care (Q) and Accommodation and Catering Services (I) sector were selected.

A glimpse at the Table 2 of descriptive statistics. Not all variables are presented as they all are similar.

From the percentile values is obvious that some highly anomalous data points exist in the data set while most of the data points are concentrated in a rather narrow range.

	TP_VAT_age	VAR01	VAR02	VAR03	VAR04	VAR05	VAR06	...	VAR30	VAR31
count	5298.0000	5298.0000	5298.0000	5298.0000	5298.0000	5298.0000	5298.0000	...	5298.0000	5298.0000
mean	133.2752	22045.9473	356994.0659	4751.8615	502.9153	379040.0132	383791.8747	...	19.3948	21.7627
std	93.1620	101677.1896	1386566.5738	43487.1858	7374.6382	1398345.7524	1402199.6922	...	805.3239	874.3772
min	0.0000	-26045.0000	-60989.0000	-2032.0000	-80.0000	-60989.0000	-60989.0000	...	-47.1800	-2306.2500
1%	2.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	...	0.0000	0.0000
25%	43.0000	183.0000	1212.2500	0.0000	0.0000	8363.0000	9664.0000	...	0.2000	0.0000
50%	124.0000	2772.0000	24767.5000	0.0000	0.0000	36838.0000	41544.5000	...	1.5500	0.4400
75%	232.0000	12215.0000	127784.0000	0.0000	0.0000	161142.5000	165641.0000	...	12.4650	5.2150
99%	298.0300	367448.3200	4502261.7900	124241.3900	10555.1300	4518137.7500	4518137.7500	...	39.6990	67.3006
max	307.0000	2259884.0000	20309962.0000	2708035.0000	407590.0000	20392272.0000	20433063.0000	...	58483.3300	58922.2200

Table 2: Descriptive Statistics of Features (Sector Q)

3.2 Data Preprocessing and Synthetic Outlier Generation

Raw, real-world data is often incomplete, noisy, and contains inconsistencies due to various reasons like incorrect data entries, system errors, etc. Data preprocessing is an essential step required to improve data quality and therefore achieve more accurate and reliable results[26]. Data preprocessing can be divided into three sections[25]:

1. Data Integration.
2. Data Cleaning.
3. Data Normalization.
4. Data Reduction.

Both chosen sectors (Q, I) undergo the same data preprocessing, synthetic outlier generation, and model implementation. All mentioned steps, with the exception of data integration, which is not applicable in this case, will be discussed in more detail.

3.2.1 Data Cleaning

The majority of data mining and other methods assume that data is complete and free of noise. However, real-world data is not clean because of data entry or transmission problems, faulty collection of data, or inconsistencies in data processing systems. To deal with those issues, a data cleaning step is required.

3.2.1.1 Missing Values Imputation

Different solutions to impute missing values are available, but there is no good way. In the case of large data set, instances with missing values can be removed from analysis though this approach may produce a bias by discarding observations with relevant information[26]. Missing values can be replaced by measures of central tendency: mean, mode, or median, but it is not an accurate solution and can not be used on categorical

variables. Missing values can also be replaced by zero or any other constant. Machine learning techniques are used to predict values based on feature similarity. Improper treatment of missing values can lead to wrong conclusions, so they have to be handled with caution.

During exploratory data analysis of VAT, declarations were noticed that many variables have a big percentage of missing values, as can be seen in Table 3. Data comes from tax declarations, and entities are obliged to fill up only relevant fields and are allowed to leave the rest empty. Thereby omitted values can be replaced by zero.

	Q	I	F	G		Q	I	F	G		Q	I	F	G
ID	0.00	0.00	0.00	0.00	**VAR10**	3.45	0.65	1.95	1.39	**VAR22**	54.49	74.21	82.01	55.24
Sector	0.00	0.00	0.00	0.00	**VAR11**	3.67	0.65	1.95	1.37	**VAR23**	0.15	0.40	0.96	0.71
TP_VAT_age	0.00	0.00	0.00	0.00	**VAR12**	71.43	79.98	91.17	81.30	**VAR24**	10.55	0.94	2.51	1.74
VAR01	17.73	3.27	16.42	8.80	**VAR13**	3.67	0.65	1.96	1.38	**VAR25**	0.55	0.44	1.16	0.83
VAR02	20.77	81.44	90.74	86.25	**VAR14**	3.45	0.65	1.95	1.39	**VAR26**	1.58	0.44	1.48	0.90
VAR03	87.58	95.29	84.50	67.21	**VAR15**	3.32	0.65	1.93	1.39	**VAR27**	0.13	0.39	0.94	0.70
VAR04	96.70	97.98	95.70	80.40	**VAR16**	96.12	98.75	97.77	89.89	**VAR28**	5.45	3.57	12.80	5.63
VAR05	3.67	3.08	15.87	8.45	**VAR17**	96.12	98.75	97.77	89.89	**VAR29**	11.91	4.05	13.99	5.88
VAR06	1.83	2.85	11.06	4.37	**VAR18**	53.48	74.07	81.45	53.04	**VAR30**	11.02	4.26	23.91	7.86
VAR07	17.81	3.42	26.54	9.33	**VAR19**	99.49	99.67	99.63	99.32	**VAR31**	22.01	4.17	14.09	6.02
VAR08	1.83	2.85	11.03	4.35	**VAR20**	20.56	0.68	2.03	1.43					
VAR09	10.50	2.87	19.83	6.28	**VAR21**	54.51	74.26	82.03	55.33					

Table 3: Percentages of missing data in each variable by sector

3.2.1.2 Noise Treatment

Two noise types can be distinguished: class noise and attribute noise. Class noise refers to incorrectly labeled observations while attribute noise to erroneous feature values. To deal with noisy data, two main approaches are described in the literature. First, to use a data polishing method that aims to correct noise in the data set and second, to use noise filters that able to identify noisy instances [25].

3.2.2 Synthetic Outlier Generation

In unsupervised learning, data is not labeled; thus, it is problematic to evaluate how well the model performs. To test the performance and validity of the model's synthetic outliers may be injected into real data set [33]. Synthetic outlier generation topic appears in the middle of data preprocessing steps because anomalies are generated right after missing values imputation.

Outliers are generated from distribution tails of randomly selected features by randomly selecting values from the interval [min, 1 % percentile] or [99 % percentile, max] of the corresponding feature. Interval values can be found in Table 2. Anomalies make up only a small fraction of the population, and it is also assumed that they are present in the original data set; therefore, only 1 % of all observations in a particular sector are added

as artificially generated outliers.

Data normalization and reduction are fitted using only training data, but the transformation is performed on the data containing generated outliers. Transformation statistics are learned only from the training set to prevent data leakage, which would occur by fitting on the data with generated outliers. Artificially created outliers are also not included in the training process of the neural network. They are used only to evaluate the goodness of the model.

3.2.3 Data Normalization

Many learning methods work best when features vary on comparable scales that are on a similar scale. Normalization is a technique that transforms data attributes to a common scale. The most well-known transformations are min-max normalization and z-score normalization.

In Table 2 data ranges of the few variables are presented. As attributes have quite different scales, normalization is necessary. Though many available normalization methods were tried, for the final experiments, standardization and min-max normalization were selected as the most appropriate.

3.2.3.1 Min-Max Normalization

Min-max normalization rescales numerical features to a specified range. Although any range can be selected, typical ones are [0, 1] and [-1, 1]. This kind of normalization preserves the original distribution but removes the effect of the dominance of those variables with very wide [min, max] interval in distance-based methods.

$$\tilde{x} = \frac{x - min(x)}{max(x) - min(x)}$$

If x is a feature value, then $min(x)$ and $max(x)$ are minimum and maximum values of this particular feature in the given data set.

3.2.3.2 z-Score Normalization

z-score normalization results in the feature distribution with a mean equal 0 and a standard deviation 1. This method of normalization is also referred to as standardization.

$$\tilde{x} = \frac{x - mean(x)}{sqrt(var(x))}$$

Here x is a feature value and $mean(x)$ and $var(x)$ are, respectively, mean and variance of this feature in the entire data set.

3.2.4 Data Reduction

To avoid the curse of dimensionality[10] or simply to save processing time or resources, data reduction techniques that reduce the original data set while maintaining the essential structure can be applied. The curse of dimensionality is a phenomenon of data sparsity, which occurs when a number of features (dimensions) grows.

To downsize the amount of data, feature selection, instance selection, discretization, or feature extraction can be performed.

3.2.4.1 Principal Component Analysis

Principal Component Analysis (PCA) is one of the multivariate data reduction techniques, and it was applied to the data in order to find a better representation of highly correlated variables (Figure 15). This method de-correlates variables by defining new data set representation by finding the directions where the variance is maximal. Transformed variables resulted from PCA, are called principal components, and often they represent data where patterns, as well as anomalies, are more clear than in the original data set. Another useful PCA feature is data de-noising[49].

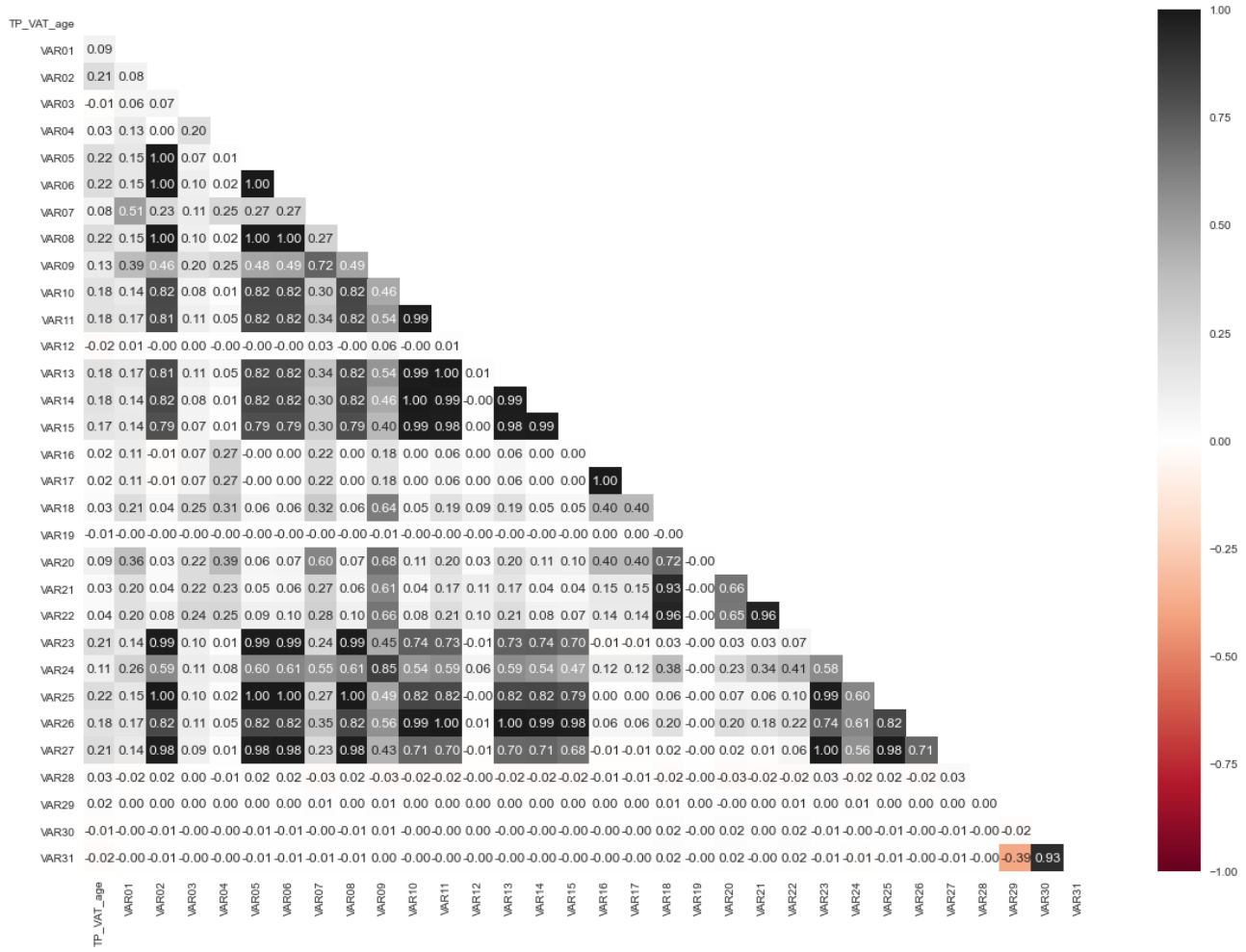


Figure 15: Correlation Matrix of Standardized Features

After data normalization, PCA was applied. A number of principal components were chosen such, so that explained variance would reach 95 %. However, principal components exhibit the drawback is their interpretation. Components are not easy to perceive because they are the combination of all features in the original data[44].

Worth reminding that principal components were found using data without artificially generated outliers, and then data with injected outliers was projected on those previously extracted components.

3.3 Results

Four different combinations, presented in Table 4 were examined to evaluate the performance of selected algorithms in two cases, when:

- 5 features were generated from tails of respective attributes, that is, in one instance, five features belong to the anomalous region and the rest to normal.
- 9 features in each generated observation are anomalous.

1 % of all instances in a sector were added as anomalies.

Number of Synthetic Values in an instance	5				9			
Scaling Method	Standardization		Normalization		Standardization		Normalization	
PCA transformation	No	Yes	No	Yes	No	Yes	No	Yes

Table 4: Different combinations to analyze performance of selected methods

Hyperparameters that were analyzed are shown in the Table 5. In unsupervised learning, parameters can not be tuned precisely; therefore, a range of them was evaluated. As the original data set is not labeled, models and performance of chosen hyperparameters were tested on generated data points that are known anomalies.

Model	Software	Hyperparameters
Autoencoder	Python, Keras library	activation = [relu, tanh] optimizer = [Adam] learning_rate = [0.0001] loss = [MSE]
k-NN Anomaly Detection	Python, pycaret library	n_neighbors = [5, 10, 15, 20]
Isolation Forest	Python, pycaret library	max_features = [1.0, 0.9, 0.8, 0.7]

Table 5: Hyperparameters used in models

The performance of autoencoder, k-nearest neighbor anomaly detection, and isolation forest was evaluated by such metrics:

- Recall of 100 % of all generated anomalies[51]. The threshold was determined by the lowest-scoring injected known anomalous observation.
- Top-k precision[51], where k is the number of generated anomalies. Precision is calculated only for top-k observations.
- Matthews correlation coefficient (MCC) that has been proven to be more reliable than the most well-known statistics like Accuracy and F_1 score[14]. Both of the latter are not suitable to measure the reliability of anomaly detection methods.
- The area under the receiver operating characteristic (ROC) curve (AUC) that is used in many studies[27],[17]. However, there are claims that ROC AUC should not be used in unbalanced data sets [42].

3.3.1 Sector Q - Health Care

The total number of instances in Q sector was 5307. After removing observations in which all feature values were missing 5298 left. In this group 53 anomalies were added. First metrics were calculated with 5 randomly selected anomalous features in each of 53 generated observations, then with 9.

Regardless of the applied scaling method features were highly correlated. An example of graphical representation of correlations is shown in Figure 15. Correlation matrices for all combinations can be found in Appendix A.

With principal components were sought to explain 95 % of variance. Different number of components was required depending on the data transformation performed. On Figure 16 is illustrated how many principal components was necessary to reach 95 % after standardization while on Figure 17 after normalization. Less components are needed after normalization, that is 4, out of which first component explains 63 % of variance followed by 23 % which explains second component. After standardization, first component captures only 39.6 % of variance. On the left side of both figures, 17 and 16, percentage of variance explained by individual components is presented while on the right side cumulative percentage of variance.

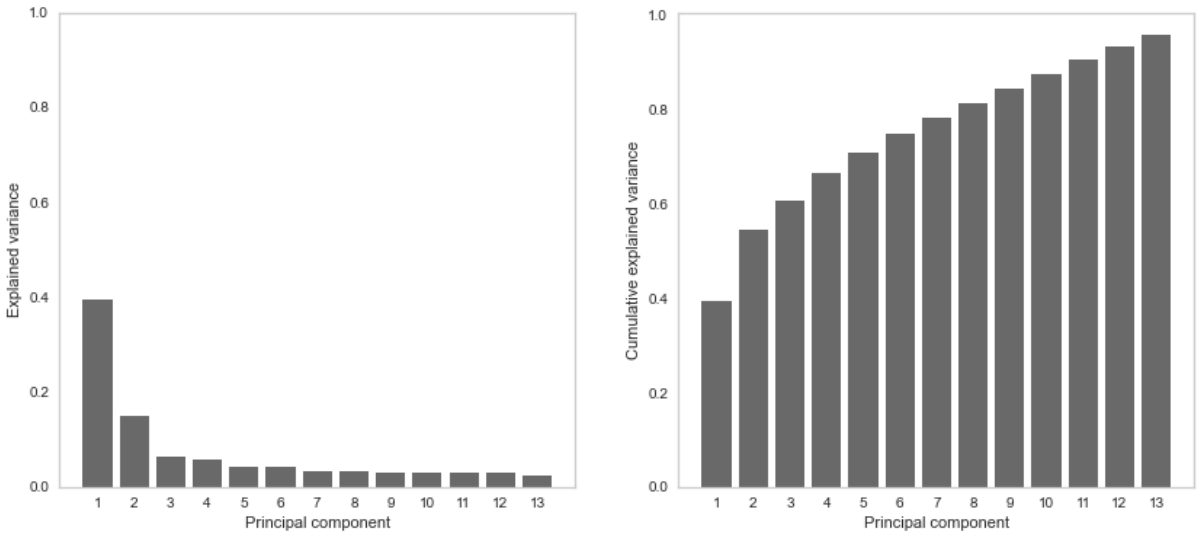


Figure 16: Variance Explained by Components after Standardization

As an example, Figure 18 and Figure 19 show how difficult can be to interpret components when many of original features are interconnected and how that affect principal components.

First two most important components are plotted in Figure 20a and Figure 20b. From visual representation, in Figure 20b anomalies are more scattered than in Figure 20a. This was expected, as normalized before transformation variables were explained by a smaller number of principal components.

More plots of various analyzed combinations in sector Q can be found in Appendix A.

Anomaly detection was performed with three models: k-NN anomaly detection, isolation forest and autoencoder. The results, when 5 out of 31 variables were randomly injected as anomalous in 53 generated

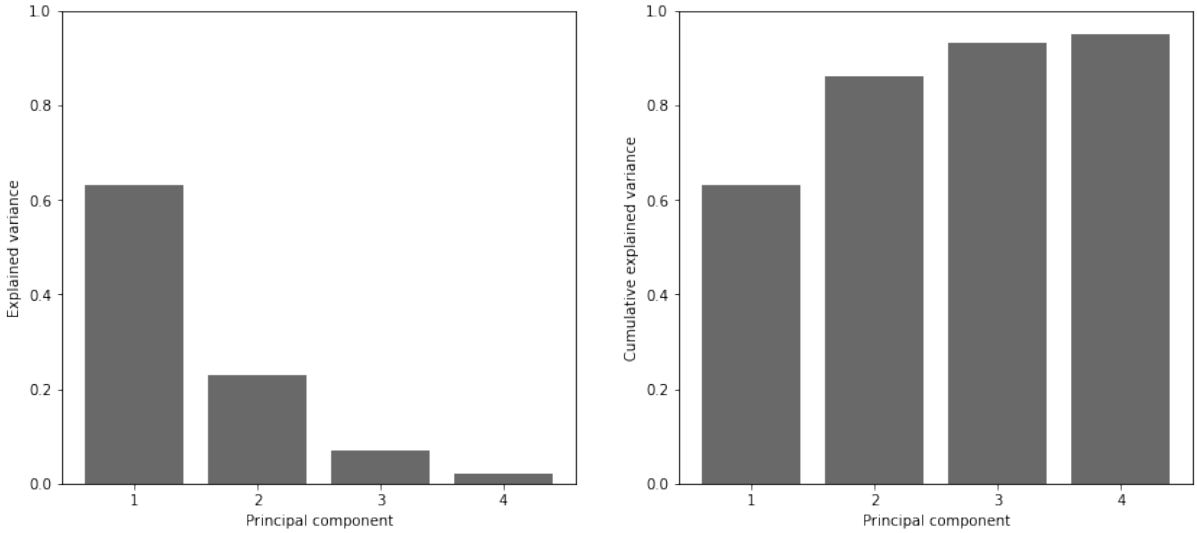


Figure 17: Variance Explained by Components after Normalization

anomalous observations, when features were standardized and PCA was not performed are presented in the Table 6.

In the best performed case, based on percentage of observations which had to be selected to obtain 100 % recall, that is k-NN model with 5 neighbors, randomly generated outlier were as low as only 1.3 % of observations had to selected. Precision in 100 % recall means, that outliers were assigned a correct label in the subsample when the lowest anomalous score belongs to synthetically generated outlier a given percentage of times. Based on this metric, k-NN also performed the best, as 79.2 % of artificial outliers were detected correctly. Matthews correlation coefficient values indicate how well model predicts values. It can take any values between -1 and +1, where -1 indicates that model in complete disagreement, that is does not predict correctly at all, value 0 is like a random prediction and +1 is a perfect prediction. Therefore the highest the score the better model predicts. The highest value also was achieved by k-NN model with 5 neighbors. ROC AUC value of 1 means, that the model separates two classes well, while value of 0 means, that model can not separate classes. Although the best precision was achieved in the case of 9 anomalous features among 31 of them, with normalization and without transformation, AUC did not exhibit best performance.

From the results in the Table 6 conclusion can be drown that under previously mentioned conditions, k-NN model with 5 neighbors performed the best according to all four computed metrics. Even the worst k-NN model with 20 neighbors were performing better than any other of two models isolation forest and autoencoder. Autoencoder showed the second best result among three models, leaving isolation forest behind.

In all other combinations, no matter the parameters, k-NN performed the best (See Appendix A).

3.3.2 Sector I - Accommodation and Catering Services

Altogether in sector I 53701 observations were included; therefore 537 anomalies were added. All steps of analysis were exactly the same as in sector Q.

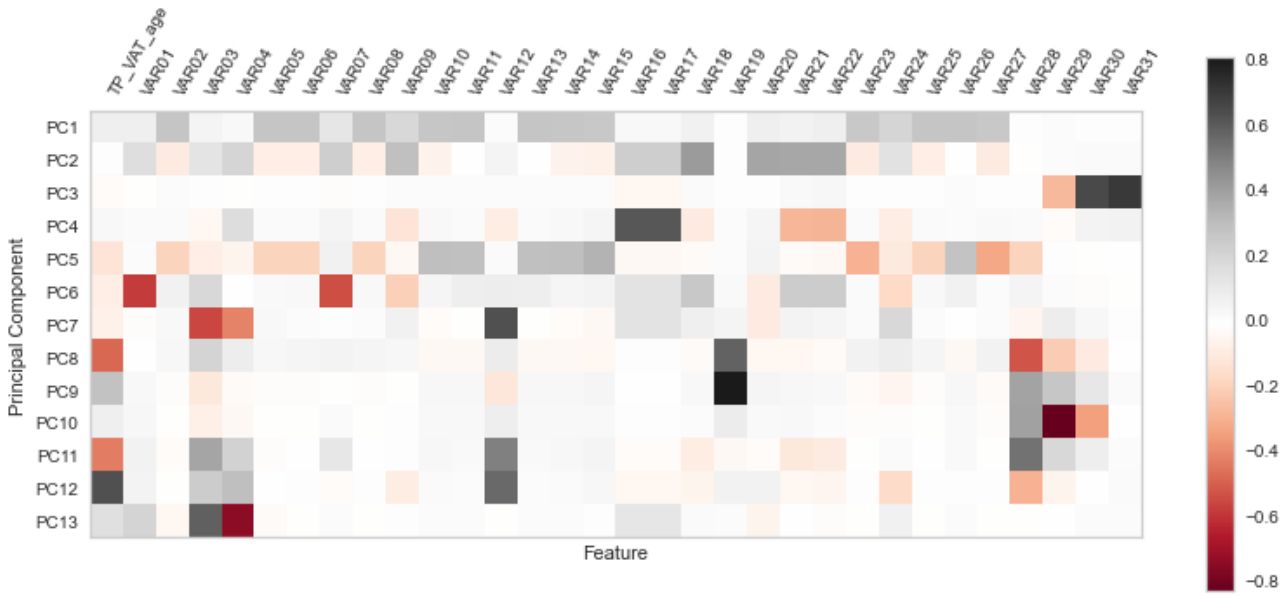


Figure 18: Impact of Original Features on Principal Components after Standardization

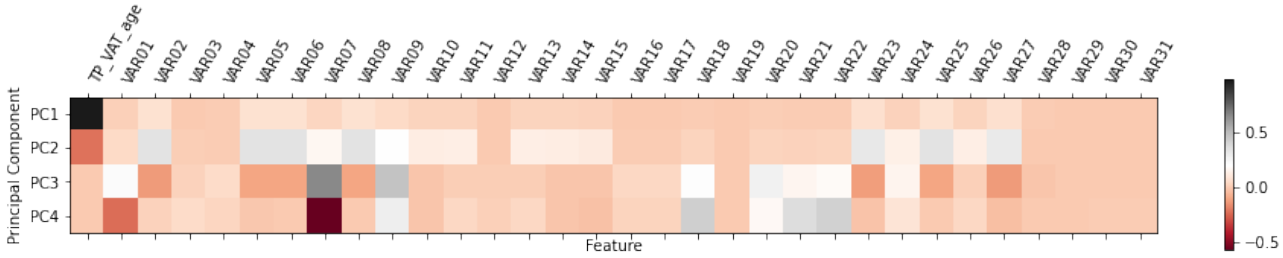
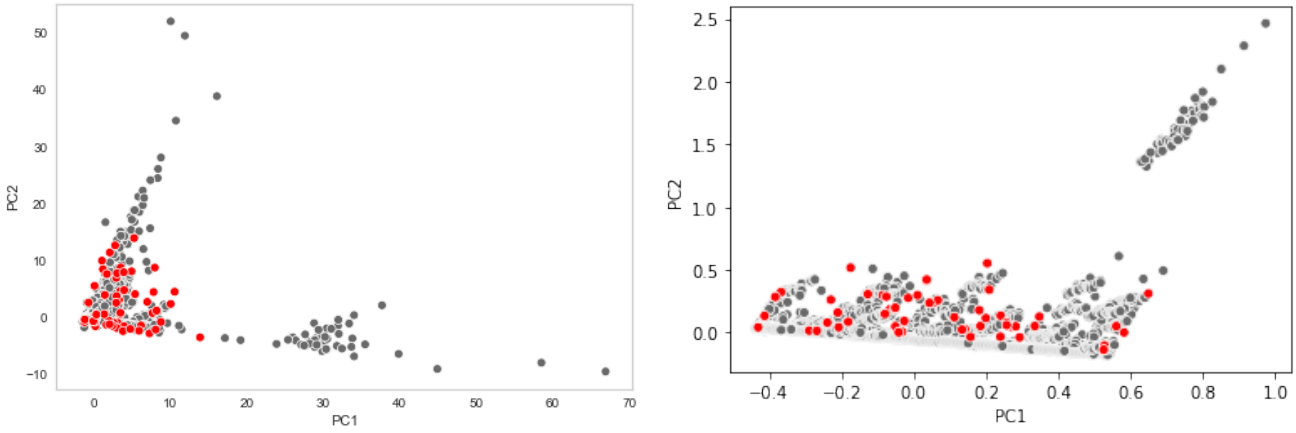


Figure 19: Impact of Original Features on Principal Components after Normalization

However, in case of data normalization by Min-max, only one principal component was needed to explain over 95 % of variance. By drawing heatmap of features affecting components, as shown on Figure 21, one feature TP_VAT_age was determined. In later analysis age was removed.

Similarly to Q sector, in I sector k-NN also showed the best performance. Performance measures can be found in Appendix B.



(a) On Standardized Data **(b)** On Normalized Data

Figure 20: Projections onto First Two Principal Components

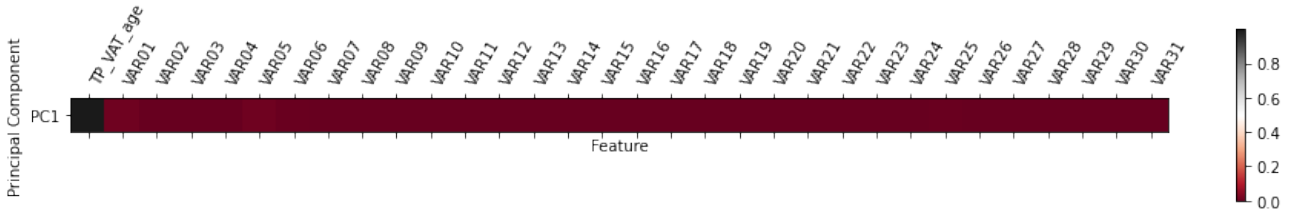


Figure 21: Impact of Original Features on Principal Components after Normalization

k-NN	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
k = 5	3.6254	0.6792	0.5595	0.7797
k = 10	4.5225	0.6226	0.5173	0.7586
k = 15	5.1579	0.6226	0.5333	0.7666
k = 20	5.3447	0.6037	0.5140	0.7570

iForest	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
features = 1.0	25.8830	0.0000	-0.0513	0.4833
features = 0.9	25.1354	0.0188	-0.0409	0.4845
features = 0.8	31.6015	0.0188	-0.0400	0.4838
features = 0.7	27.9573	0.0188	-0.0381	0.4818

Autoencoder	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
32-25-3-25-32	8.2975	0.4528	0.3788	0.6970

Table 6: Models performance (5 variables in 53 generated outliers are anomalous, standardization performed)

Conclusion

The focus of this paper was unsupervised learning methods, more specifically, anomaly detection models that can be applied to identify VAT fraud in declaration forms.

Unsupervised learning methods are superior to the supervised methods in the sense that labeled data is not required. In the tax domain, where data labeling is extremely difficult and time-consuming, such methods are of great importance. Another relevant aspect of unsupervised learning is that new, previously unseen patterns can be detected as models are not trying to learn from known instances but rather trying to discover hidden patterns in the data. This means that fraudsters, who are constantly looking for new schemes to exploit the VAT system, are always staying ahead of tax authorities if the latter is using supervised methods or, most commonly used at the moment, rules-based techniques. To both supervised machine learning and rules-based techniques, fraud cases have to be introduced so models could learn from them, and in the future, could alarm about detected similarities between new cases and learned from the historical ones.

Unsupervised learning is not commonly used in the tax domain because it is difficult to evaluate how well the model performs. It is very important not to falsely accuse compliant entities of committing fraud just because the model found some deviating from the norm structures. During this research, as labels were unknown and in such case to evaluate model is not possible, synthetic outliers were generated from original data features distributions and injected into the dataset during the test phase. Known outliers allowed to evaluate models with four metrics: 1. Percentage of observations needed to include all generated outliers that are to achieve 100 % recall. 2. Top-k precision, where k corresponds to a number of injected outliers. Precision is calculated in the top-k observations. 3. Matthews correlation coefficient that is a reliable statistic to measure model reliability. 4. Widely used in performance measures ROC AUC.

Based on computed values of the aforementioned metrics, were concluded that k-NN is a superior model to isolation forest and autoencoder in the anomaly detection field. Irrespective of chosen models' hyperparameters, performed data normalization, and reduction techniques, also count of features replaced by anomalous values in the generated outliers, k-NN anomaly detection method was always the best. Autoencoder performed worse; however, isolation forest was behind by a wide margin.

In the beginning ensemble method, by a combination of analyzed three methods, was considered. However, when results were obtained, and quite large differences were detected, the thought about ensemble models was abandoned.

Unsupervised learning methods have a great potential to help tax authorities in detecting anomalous events. However, due to difficulties related to evaluation of performance, and importance of possible mistakes, those methods are not widely used. Collaboration between data scientists and tax domain experts is required in order to achieve better results. More detailed investigations are needed in to put such methods into practice.

References

[1] Mtic (missing trader intra community) fraud. https://www.europol.europa.eu/crime-areas-and-trends/crime-areas/economic-crime/mtic-missing-trader-intra-community-fraud.

[2] Aktuali informacija gyventojams, pradedantiems verslą, apie pridėtinės vertės mokestį. https://www.vmi.lt/cms/pridetines-vertes-mokestis.

[3] Council directive 2010/24/eu of 16 march 2010 concerning mutual assistance for the recovery of claims relating to taxes, duties and other measures. *Official Journal of the European Union*, 2010.

[4] C. C. Aggarwal. An introduction to outlier analysis. In *Outlier analysis*, pages 1–34. Springer, 2017.

[5] M. Ahmed, A. N. Mahmood, and M. R. Islam. A survey of anomaly detection techniques in financial domain. *Future Generation Computer Systems*, 55:278–288, 2016.

[6] M. Albashrawi and M. Lowell. Detecting financial fraud using data mining techniques: A decade review from 2004 to 2015. *Journal of Data Science*, 14(3):553–569, 2016.

[7] R. Asadauskaitė. Tikrinamose odontologų įmonėse automobiliai – įmonės ar jos savininko? https://www.vmi.lt/evmi/-/tikrinamose-odontolog-c5-b3-c4-afmon-c4-97se-automobiliai-c4-afmon-c4-97s-ar-jos-
_ftn1, 2020.

[8] E. C. O. AUDITORS. Tackling intra-community vat fraud: More action needed, 2015.

[9] S. Basta, F. Fassetti, M. Guarascio, G. Manco, F. Giannotti, D. Pedreschi, L. Spinsanti, G. Papi, and S. Pisani. High quality true-positive prediction for fiscal fraud detection. In *2009 IEEE International Conference on Data Mining Workshops*, pages 7–12. IEEE, 2009.

[10] R. E. Bellman. *Adaptive control processes: a guided tour*, volume 2045. Princeton university press, 2015.

[11] I. Bobak. Hierarchical temporal memory for real-time anomaly detection. https://www.slideshare.net/ibobak/hierarchical-temporal-memory-for-realtime-anomaly-detection, 2017.

[12] F. Borselli. Organised vat fraud: features, magnitude, policy perspectives. *Policy Perspectives (October 31, 2011). Bank of Italy Occasional Paper*, (106), 2011.

[13] V. Chandola. Anomaly detection: A survey varun chandola, arindam banerjee, and vipin kumar, 2007.

[14] D. Chicco and G. Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):6, 2020.

[15] T. Dasgupta. Deep autoencoders using tensorflow. https://medium.com/themlblog/deep-autoencoders-using-tensorflow-4f68655c8d08, 2018.

[16] D. de Roux, B. Perez, A. Moreno, M. d. P. Villamil, and C. Figueroa. Tax fraud detection for under-reporting declarations using an unsupervised machine learning approach. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 215–222, 2018.

[17] R. Domingues, M. Filippone, P. Michiardi, and J. Zouaoui. A comparative evaluation of outlier detection algorithms: Experiments and analyses. *Pattern Recognition*, 74:406–421, 2018.

[18] S. Fedeli and F. Forte. Eu vat frauds. *European Journal of law and Economics*, 31(2):143–166, 2011.

[19] L. R. finansų ministerija. Valstybės biudžeto ir savivaldybių biudžetų konsoliduotos visumos pajamų struktūra. https://finmin.lrv.lt/lt/veiklos-sritys/biudzetas/patvirtinti-biudzetai/2020-m-biudzetas/issami-informacija-apie-2020-m-biudzeta, 2020.

[20] L. R. finansų ministerija. Pridėtinės vertės mokestis. https://finmin.lrv.lt/lt/veiklos-sritys/mokesciai/pagrindiniai-mokesciai/pridetines-vertes-mokestis, 2020.

[21] D.-G. for Taxation and F. . T. G. P. G. s. V. f. F. Customs Union (European Commission). The concept of tax gaps report iii : Mtic fraud gap estimation methodologies. Technical report, EUROPEAN COMMISSION, 2018.

[22] M. Frunza. Cost of the mtic vat fraud for european union members. *Available at SSRN 2758566*, 2016.

[23] M.-C. Frunza, D. Guegan, and A. Lassoudiere. Missing trader fraud on the emissions market. *Journal of financial crime*, 2011.

[24] O. G. Everything you need to know about neural networks and backpropagation - machine learning easy and fun. https://towardsdatascience.com/everything-you-need-to-know-about-neural-networks-and-backpropagation-machine-learning-made-easy 2019.

[25] S. García, J. Luengo, and F. Herrera. *Data preprocessing in data mining*. Springer, 2015.

[26] S. García, S. Ramírez-Gallego, J. Luengo, J. M. Benítez, and F. Herrera. Big data preprocessing: methods and prospects. *Big Data Analytics*, 1(1):9, 2016.

[27] M. Goldstein and S. Uchida. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PloS one*, 11(4):e0152173, 2016.

[28] M. Goldstein, S. Asanger, M. Reif, and A. Hutchison. Enhancing security event management systems with unsupervised anomaly detection. In *ICPRAM*, pages 530–538, 2013.

[29] P. C. González and J. D. Velásquez. Characterization and detection of taxpayers with false invoices using data mining techniques. *Expert Systems with Applications*, 40(5):1427–1436, 2013.

[30] F. E. Grubbs. Procedures for detecting outlying observations in samples. *Technometrics*, 11(1):1–21, 1969.

[31] N. Hangáčová and T. Strémy. Value added tax and carousel fraud schemes in the european union and the slovak republic. *European Journal of Crime, Criminal Law and Criminal Justice*, 26(2):132–159, 2018.

[32] E. Heinäluoma. Council directive 2006/112/ec of 28 november 2006 on the common system of value added tax. *Official Journal of the European Union*, 49:11.

[33] D. Hendrycks, M. Mazeika, and T. Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

[34] M. E. Hutton. *The Revenue Administration–Gap Analysis Program: Model and Methodology for Value-Added Tax Gap Estimation*. International Monetary Fund, 2017.

[35] G. M. J. Understanding k-means clustering in machine learning. https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1, 2018.

[36] M. Keen and S. Smith. Vat fraud and evasion: What do we know and what can be done? *National Tax Journal*, pages 861–887, 2006.

[37] S. E. Kohonen. Self-organizing maps. https://towardsdatascience.com/kohonen-self-organizing-maps-a29040d688da, 2019.

[38] Y. D. L. The complete guide to decision trees. https://towardsdatascience.com/the-complete-guide-to-decision-trees-28a4e3c7be14, 2019.

[39] M. Lamensch and E. Ceci. Vat fraud: Economic impact, challenges and policy issues. Technical report, 2018.

[40] K. Lind. Reverse charging: the best possible solution for preventing vat fraud. *World Journal of VAT/GST Law*, 2(2):97–115, 2013.

[41] F. T. Liu, K. M. Ting, and Z.-H. Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422. IEEE, 2008.

[42] J. M. Lobo, A. Jiménez-Valverde, and R. Real. Auc: a misleading measure of the performance of predictive distribution models. *Global ecology and Biogeography*, 17(2):145–151, 2008.

[43] K. Malaszczyk and B. M. Purcell. Big data analytics in tax fraud detection. *Northeastern Association of Business, Economics and Technology*, page 233, 2017.

[44] A. C. Müller, S. Guido, et al. *Introduction to machine learning with Python: a guide for data scientists*. " O'Reilly Media, Inc.", 2016.

[45] T. Nolle, S. Luettgen, A. Seeliger, and M. Mühlhäuser. Analyzing business process anomalies using autoencoders. *Machine Learning*, 107(11):1875–1893, 2018.

[46] A. Olofsdottir. Vat fraud. https://www.europarl.europa.eu/cmsdata/150601/5%20-%2003%20Alma%20Eurofisc%20Initial%20statement%20TAX3%20hearing%20on%20VAT%20fraud%20June%2028%202018%20(2).%20docx.pdf, 2018.

[47] J. Podlipnik. Missing trader intra-community and carousel vat frauds-ecj and ecthr case law. *Croatian yearbook of European law & policy*, 8(8):457–472, 2012.

[48] G. Poniatowski, M. Bonch-Osmolovskiy, and A. Śmietanka. Study and reports on the vat gap in the eu-28 member states: 2020 final report. 2020.

[49] M. Richardson. Principal component analysis. *URL: http://people. maths. ox. ac. uk/richardsonm/SignalProcPCA. pdf (last access: 3.5. 2013). Aleš Hladnik Dr., Ass. Prof., Chair of Information and Graphic Arts Technology, Faculty of Natural Sciences and Engineering, University of Ljubljana, Slovenia ales. hladnik@ ntf. uni-lj. si*, 6:16, 2009.

[50] D. Saxunova, R. Sulíkova, and R. Szarkova. Tax management hierarchy–tax fraud and a fraudster. In *Manag. Int. Conf*, pages 633–644, 2017.

[51] M. Schreyer, T. Sattarov, D. Borth, A. Dengel, and B. Reimer. Detection of anomalies in large scale accounting data using deep autoencoder networks. *arXiv preprint arXiv:1709.05254*, 2017.

[52] L. Sokanovic. Missing trader fraud as part of organised crime in eu. *Economic and Social Development: Book of Proceedings*, pages 160–168, 2017.

[53] Taxation and C. Union. What is vat? https://ec.europa.eu/taxation_customs/business/vat/what-is-vat_en.

[54] Taxation and C. Union. Modernising vat for cross-border e-commerce. https://ec.europa.eu/taxation_customs/business/vat/modernising-vat-cross-border-ecommerce_en, 2020.

[55] Taxation and C. Union. Vies (vat information exchange system) enquiries. https://ec.europa.eu/taxation_customs/business/vat/eu-vat-rules-topic/vies-vat-information-exchange-system-enquiries_en, 2020.

[56] J. Vanhoeyveld, D. Martens, and B. Peeters. Value-added tax fraud detection with scalable anomaly detection techniques. *Applied Soft Computing*, 86:105895, 2020.

A Sector Q Results

A.1 Sector: Q, Anomalous Features: 5, Scaling: Standardization, PCA: no

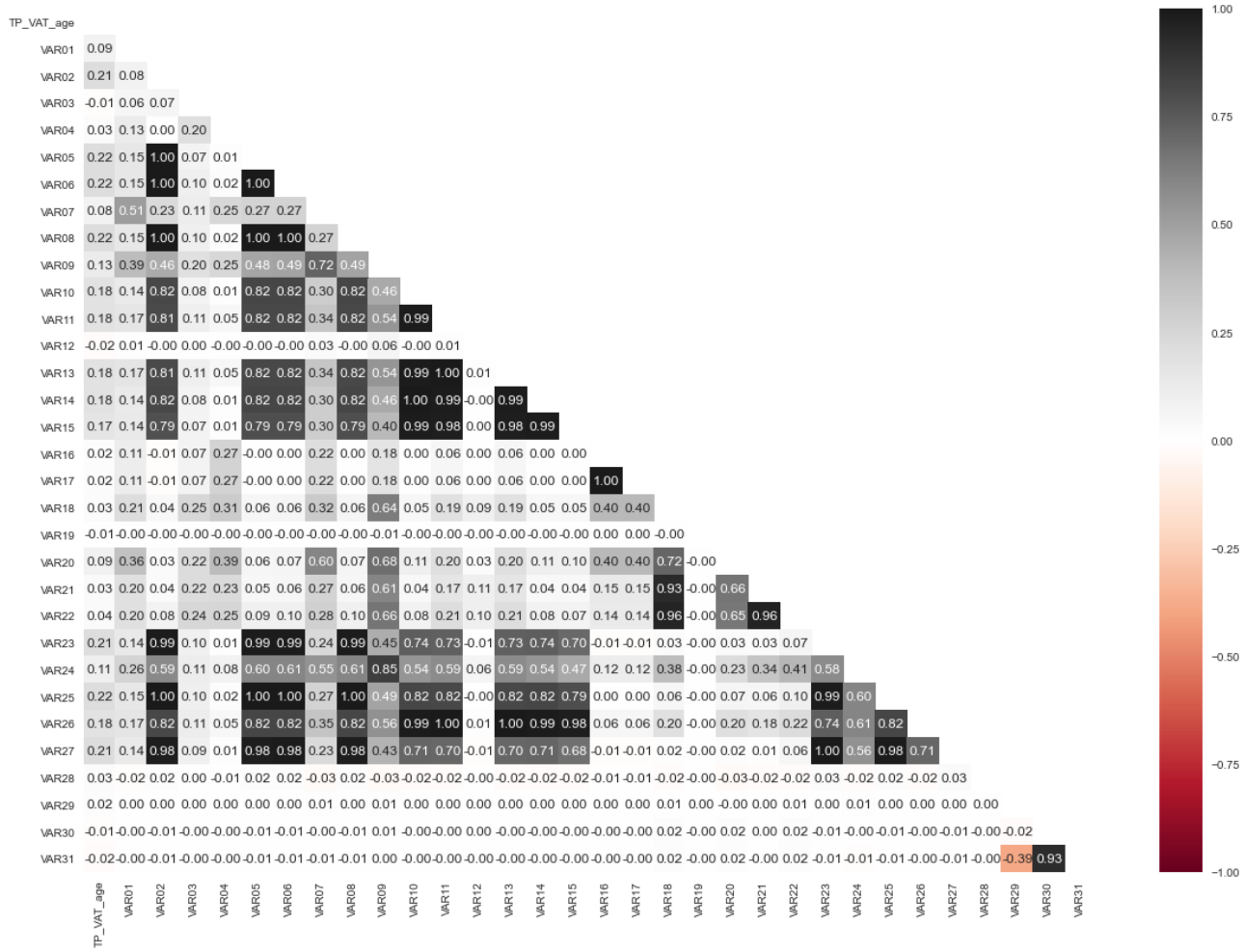


Figure 22: Correlation Matrix

k-NN	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
k = 5	3.6254	0.6792	0.5595	0.7797
k = 10	4.5225	0.6226	0.5173	0.7586
k = 15	5.1579	0.6226	0.5333	0.7666
k = 20	5.3447	0.6037	0.5140	0.7570

iForest	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
features = 1.0	25.8830	0.0000	-0.0513	0.4833
features = 0.9	25.1354	0.0188	-0.0409	0.4845
features = 0.8	31.6015	0.0188	-0.0400	0.4838
features = 0.7	27.9573	0.0188	-0.0381	0.4818

Autoencoder	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
32-25-3-25-32	8.2975	0.4528	0.3788	0.6970

Figure 23: Performance of the models

A.2 Sector: Q, Anomalous Features: 5, Scaling: Standardization, PCA: yes

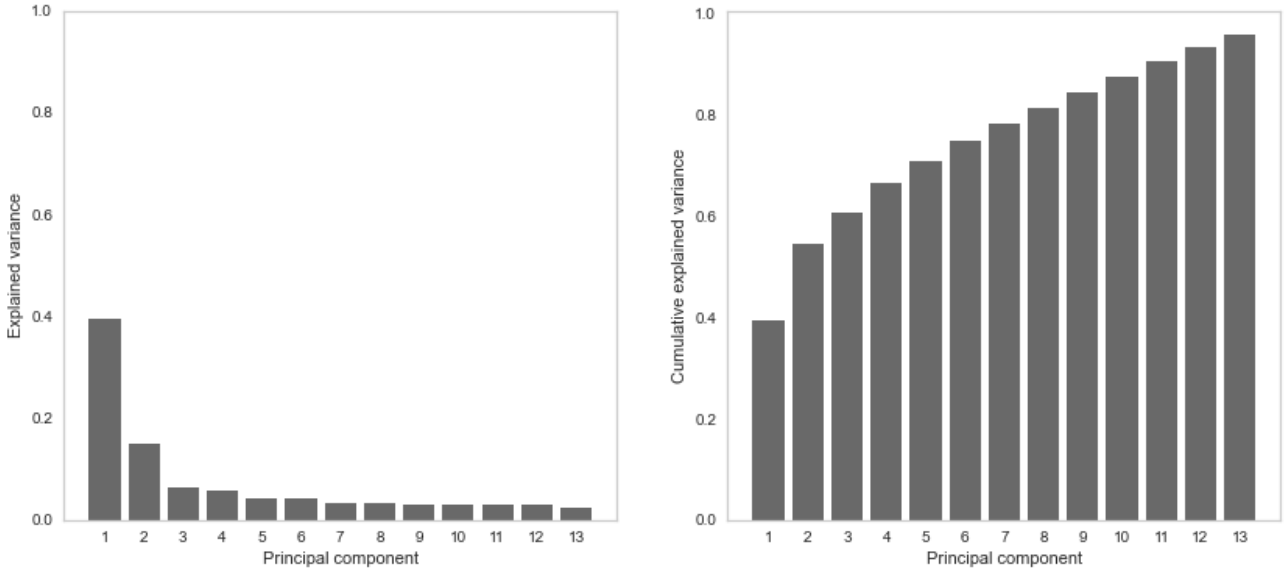


Figure 24: Explained Variance

Figure 25: PCA heatmap

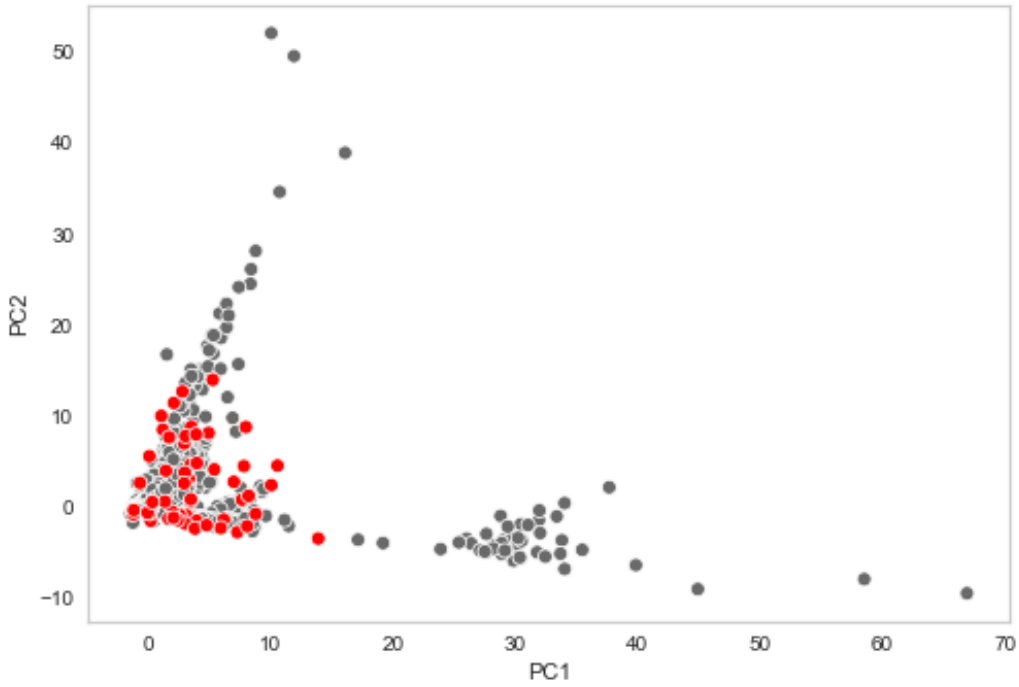


Figure 26: PC1 vs PC2

k-NN	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
k = 5	17.4173	0.4339	0.3998	0.6999
k = 10	17.9779	0.4150	0.3810	0.6905
k = 15	14.0160	0.3773	0.3300	0.6650
k = 20	14.9691	0.3773	0.3332	0.6666

iForest	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
features = 1.0	25.2102	0.3396	0.3126	0.6553
features = 0.9	18.5572	0.3018	0.2625	0.6735
features = 0.8	24.6122	0.2830	0.2529	0.6588
features = 0.7	27.8639	0.3396	0.3153	0.6763

Autoencoder	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
13-10-2-10-13	19.6411	0.3207	0.2895	0.6448

Figure 27: Performance of the models

A.3 Sector: Q, Anomalous Features: 5, Scaling: Normalization, PCA: no

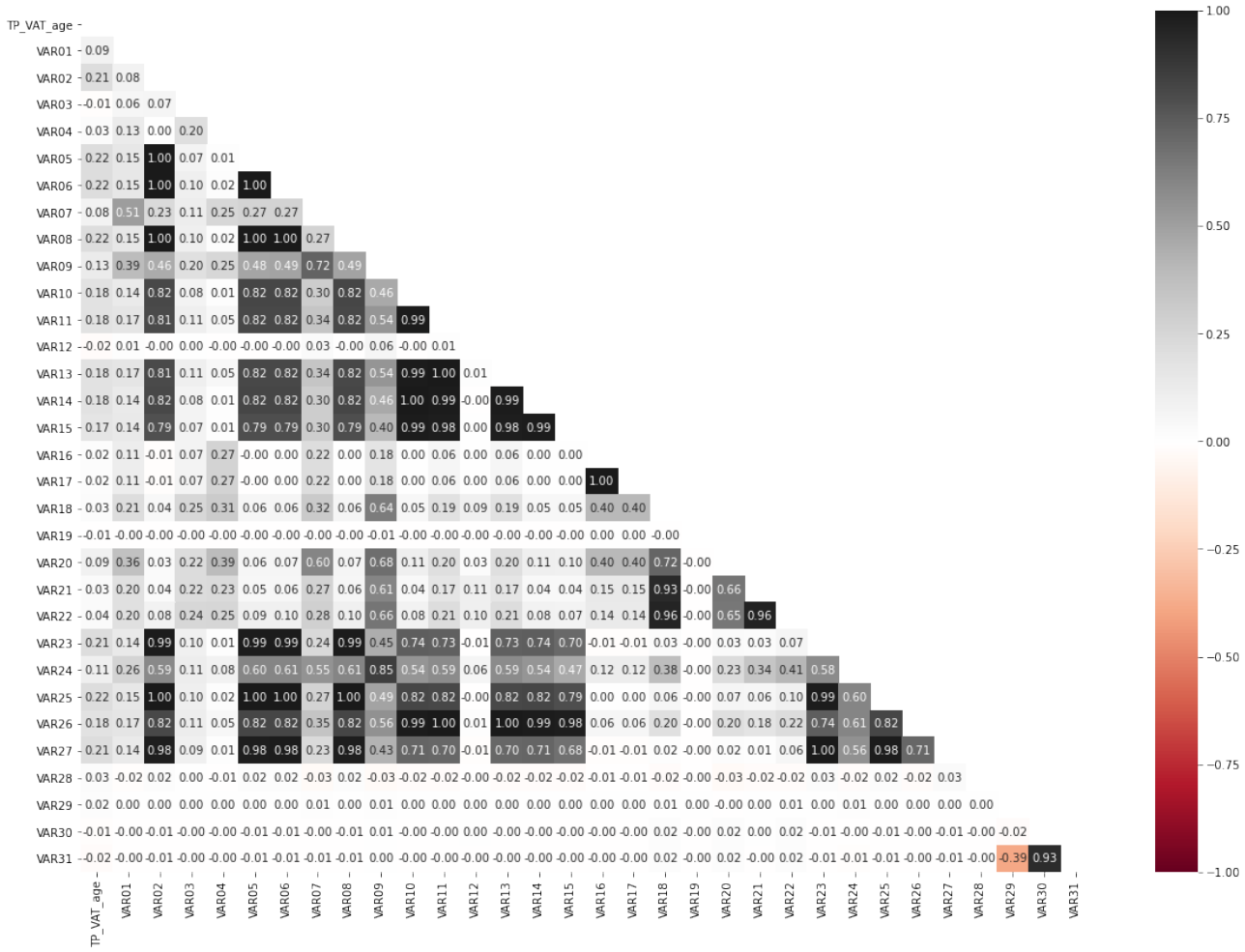


Figure 28: Correlation Matrix

k-NN	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
k = 5	4.8775	0.7358	0.6688	0.8344
k = 10	6.5968	0.6981	0.6449	0.8224
k = 15	8.0358	0.5849	0.5267	0.7633
k = 20	8.9142	0.6037	0.5543	0.7771

iForest	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
features = 1.0	30.7419	0.0000	-0.0332	0.4819
features = 0.9	32.9844	0.0000	-0.0309	0.4813
features = 0.8	31.5828	0.0000	-0.0323	0.4826
features = 0.7	28.2750	0.0000	-0.0362	0.4823

Autoencoder	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
32-25-3-25-32	7.5313	0.6792	0.6308	0.8154

Figure 29: Performance of the models

A.4 Sector: Q, Anomalous Features: 5, Scaling: Normalization, PCA: yes

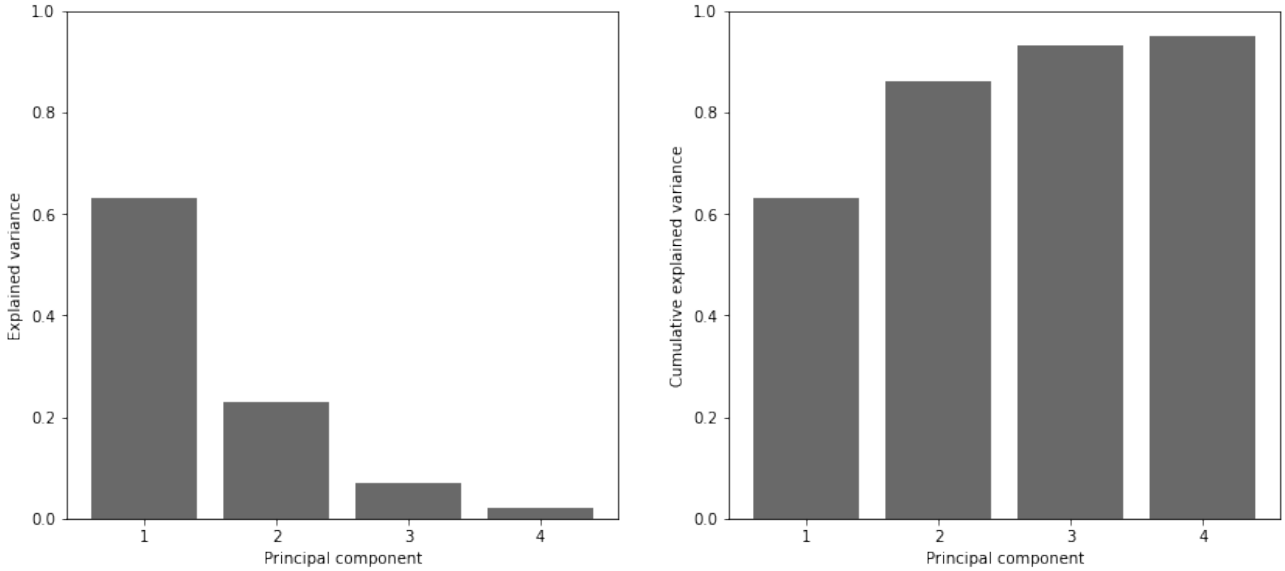


Figure 30: Explained Variance

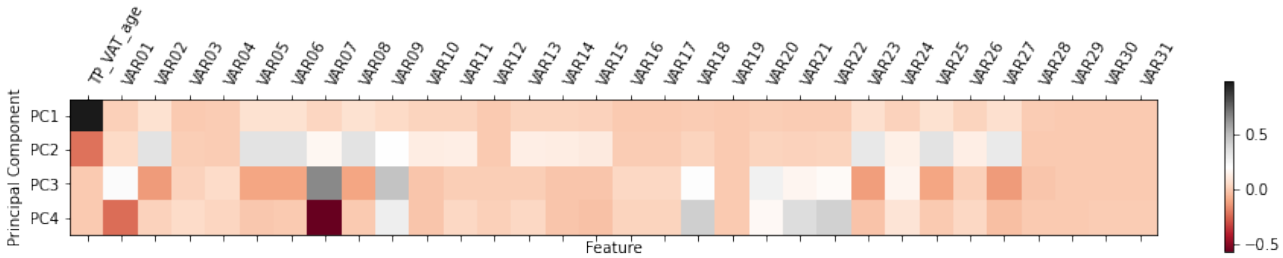


Figure 31: PCA heatmap

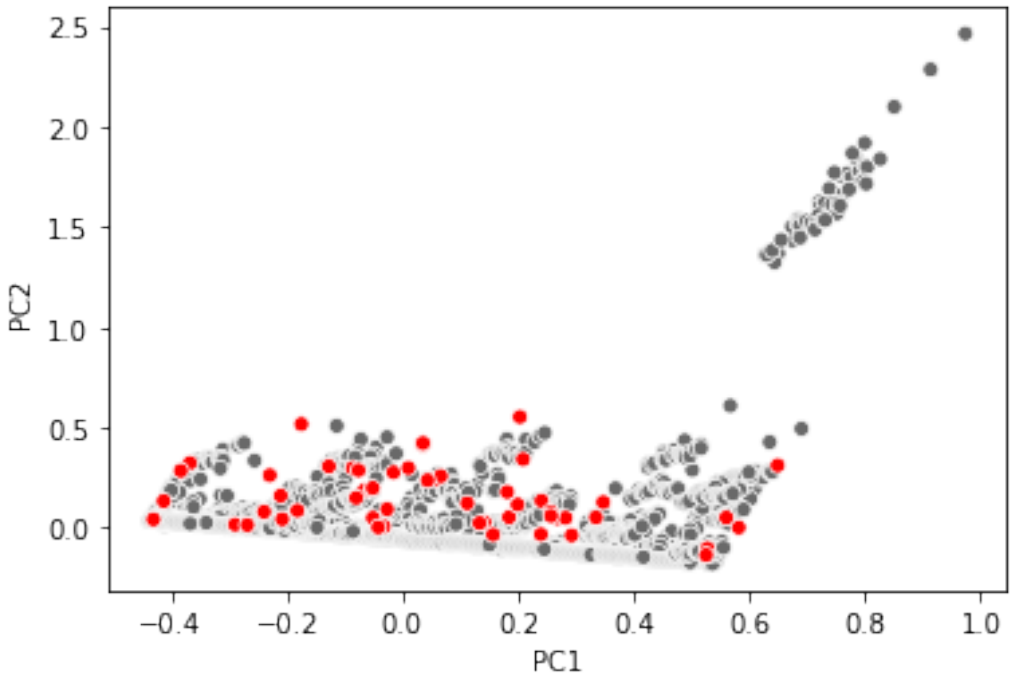


Figure 32: PC1 vs PC2

k-NN	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
k = 5	22.9863	0.2452	0.2113	0.6056
k = 10	30.2747	0.1698	0.1417	0.5708
k = 15	32.3490	0.1509	0.1241	0.5620
k = 20	34.9467	0.1509	0.1261	0.5630

iForest	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
features = 1.0	64.6421	0.0000	-0.0155	0.4922
features = 0.9	61.3156	0.0188	0.0027	0.5013
features = 0.8	61.3156	0.0188	0.0027	0.5013
features = 0.7	68.6039	0.0188	0.0044	0.5022

Autoencoder	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
5-4-2-4-5	90.1139	0.0943	0.0842	0.5421

Figure 33: Performance of the models

A.5 Sector: Q, Anomalous Features: 9, Scaling: Standardization, PCA: no

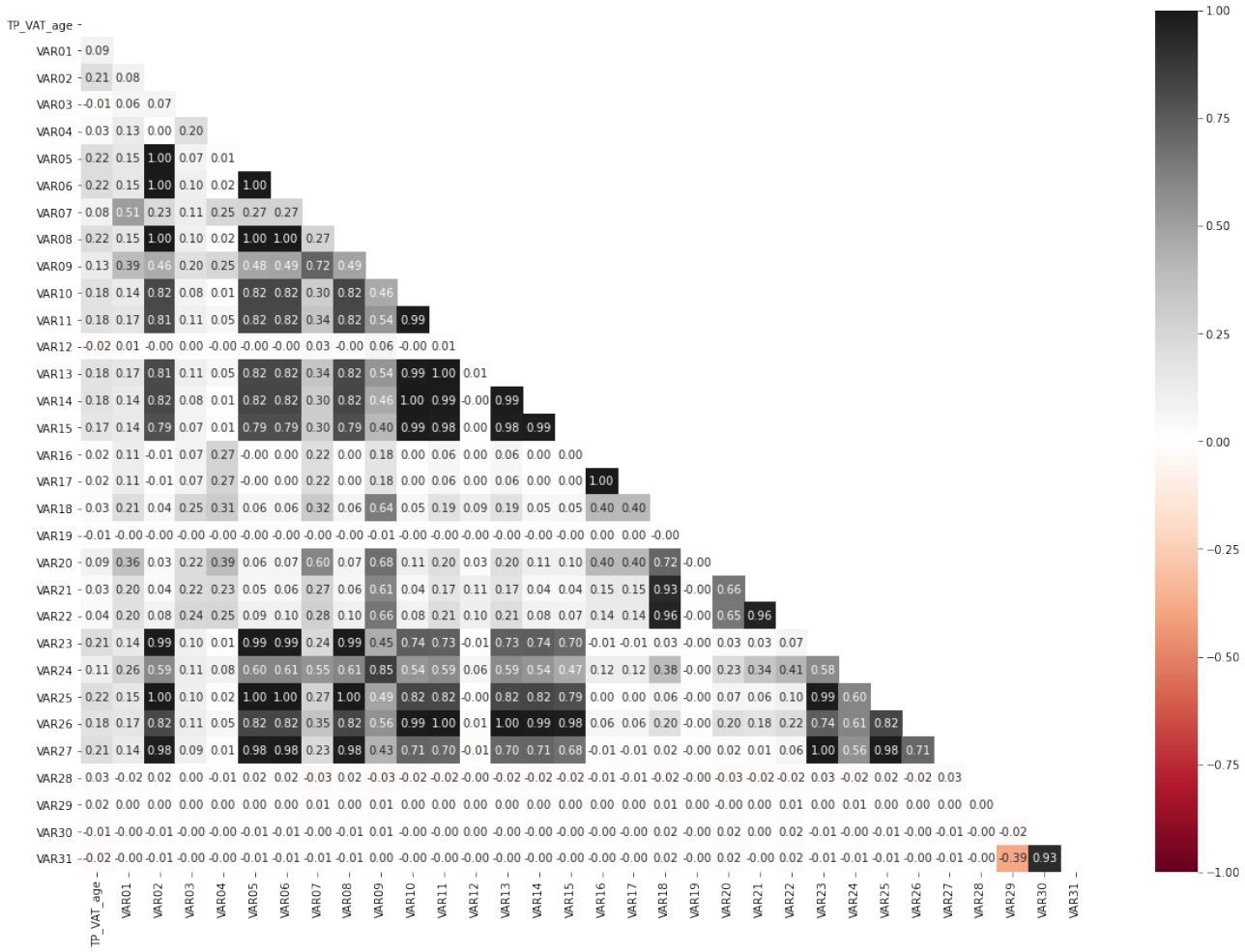


Figure 34: Correlation Matrix

k-NN	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
k = 5	1.3268	0.7735	0.1420	0.5710
k = 10	1.4950	0.7358	0.2358	0.6179
k = 15	1.9061	0.7358	0.4558	0.7279
k = 20	1.9248	0.7358	0.4613	0.7306

iForest	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
features = 1.0	13.6049	0.0000	-0.0784	0.4665
features = 0.9	14.5767	0.0000	-0.0728	0.4698
features = 0.8	15.0065	0.0000	-0.0705	0.4671
features = 0.7	14.9504	0.0000	-0.0708	0.4635

Autoencoder	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
32-25-3-25-32	2.9900	0.6603	0.4937	0.7468

Figure 35: Performance of the models

A.6 Sector: Q, Anomalous Features: 9, Scaling: Standardization, PCA: yes

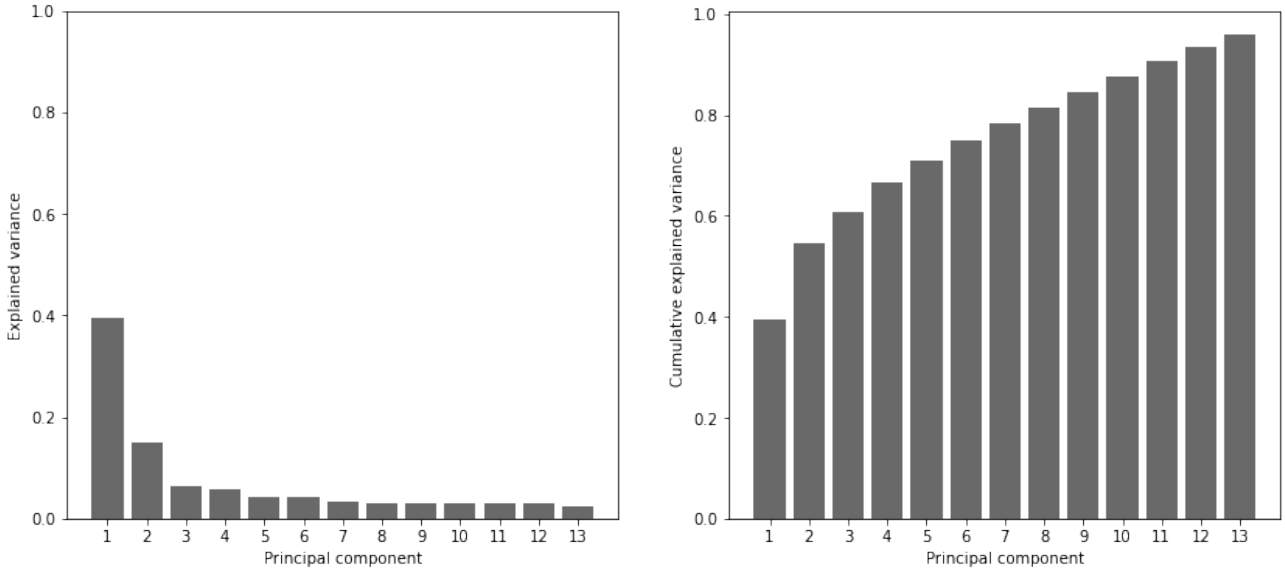


Figure 36: Explained Variance

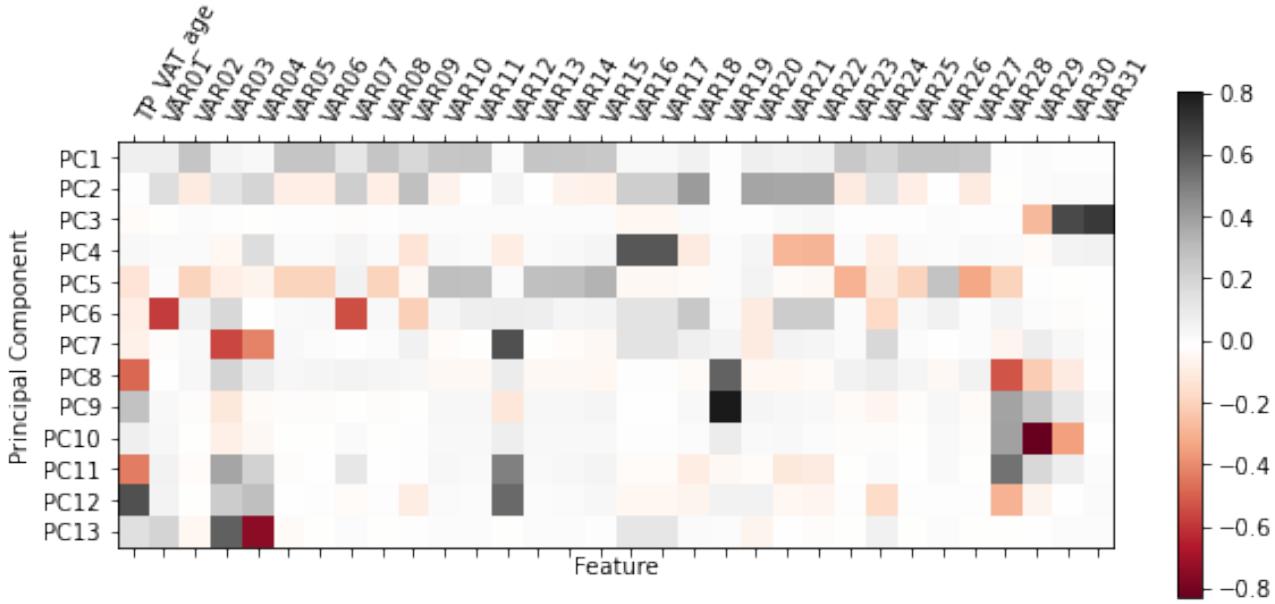


Figure 37: PCA heatmap

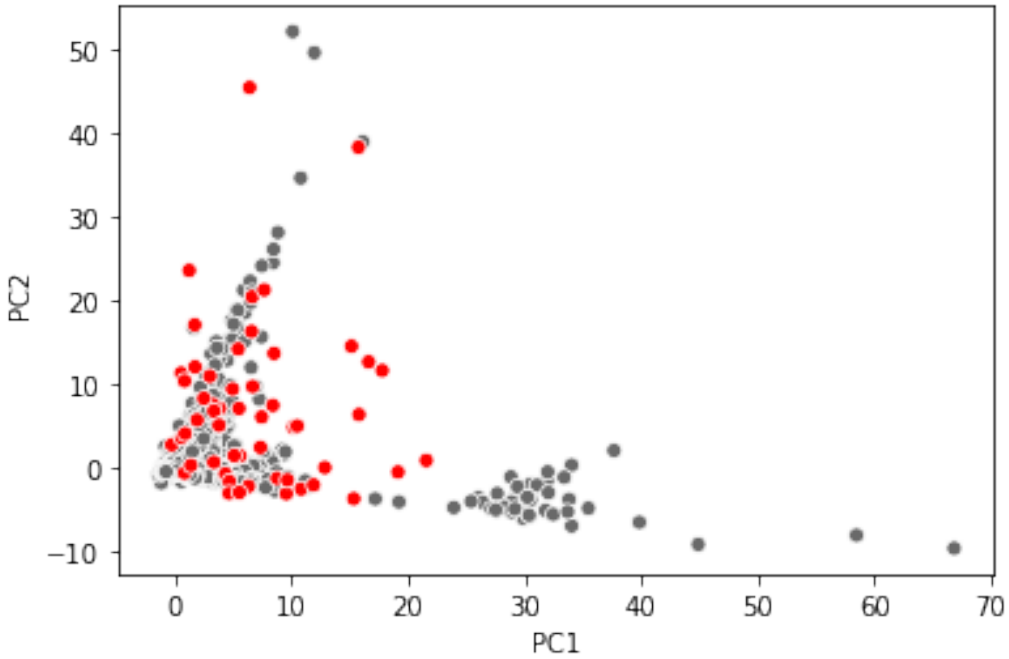


Figure 38: PC1 vs PC2

k-NN	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
k = 5	4.4478	0.6792	0.5878	0.7939
k = 10	5.5129	0.6603	0.5863	0.7931
k = 15	5.7559	0.6603	0.5900	0.7950
k = 20	6.3913	0.6792	0.6206	0.8103

iForest	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
features = 1.0	10.5774	0.5660	0.5212	0.7606
features = 0.9	9.6617	0.6037	0.5586	0.7793
features = 0.8	9.4001	0.6226	0.5782	0.7891
features = 0.7	12.4836	0.6037	0.5696	0.7848

Autoencoder	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
13-10-2-10-13	8.2040	0.4905	0.4207	0.7103

Figure 39: Performance of the models

A.7 Sector: Q, Anomalous Features: 9, Scaling: Normalization, PCA: no

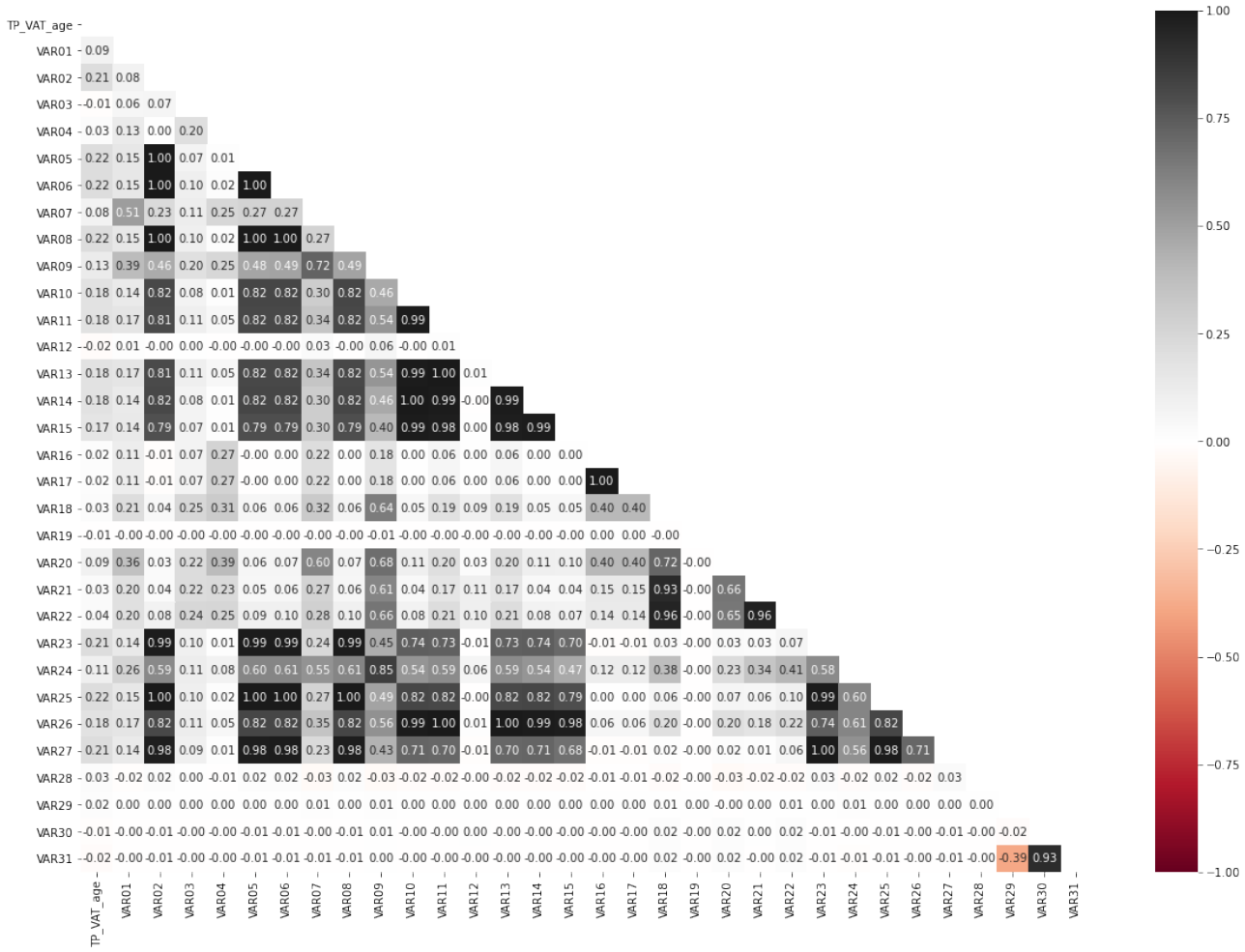


Figure 40: Correlation Matrix

k-NN	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
k = 5	1.3081	0.7924	0.1813	0.5906
k = 10	1.5698	0.7924	0.4487	0.7243
k = 15	1.9809	0.7358	0.4765	0.7382
k = 20	2.1117	0.7358	0.5063	0.7531

iForest	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
features = 1.0	14.3524	0.0000	-0.0740	0.4629
features = 0.9	13.8852	0.0000	-0.0767	0.4616
features = 0.8	16.4268	0.0000	-0.0640	0.4679
features = 0.7	16.0717	0.0000	-0.0650	0.4672

Autoencoder	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
32-25-3-25-32	1.8874	0.7924	0.5679	0.7839

Figure 41: Performance of the models

A.8 Sector: Q, Anomalous Features: 9, Scaling: Normalization, PCA: yes

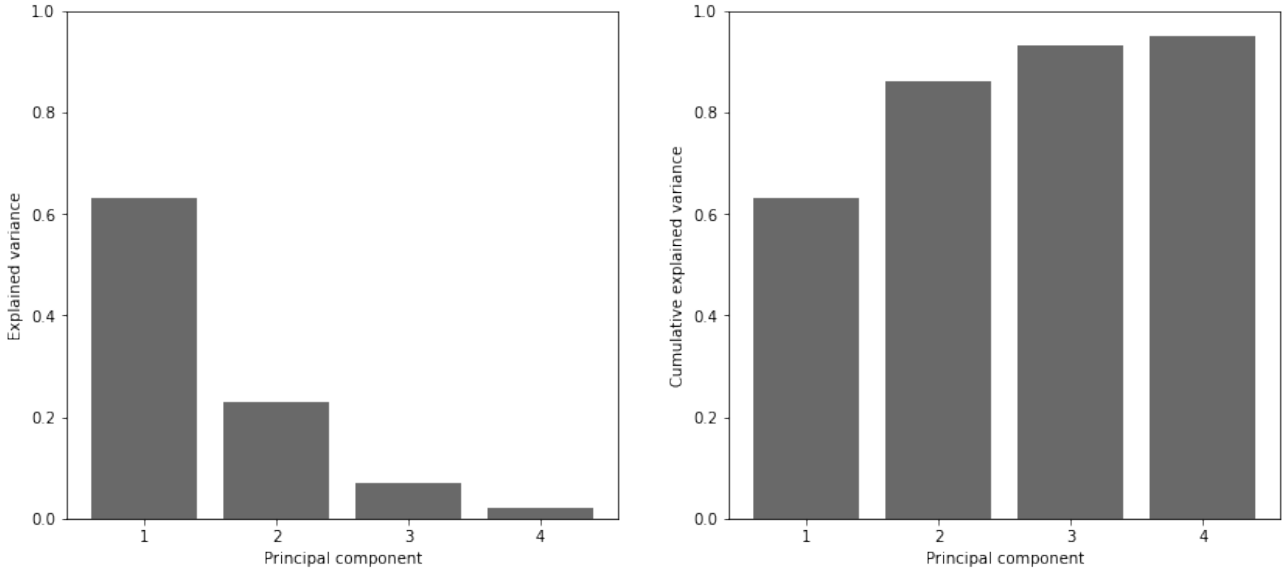


Figure 42: Explained Variance

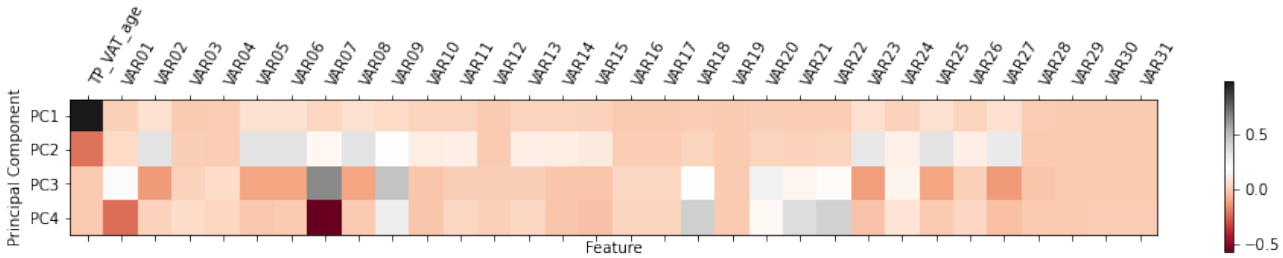


Figure 43: PCA heatmap

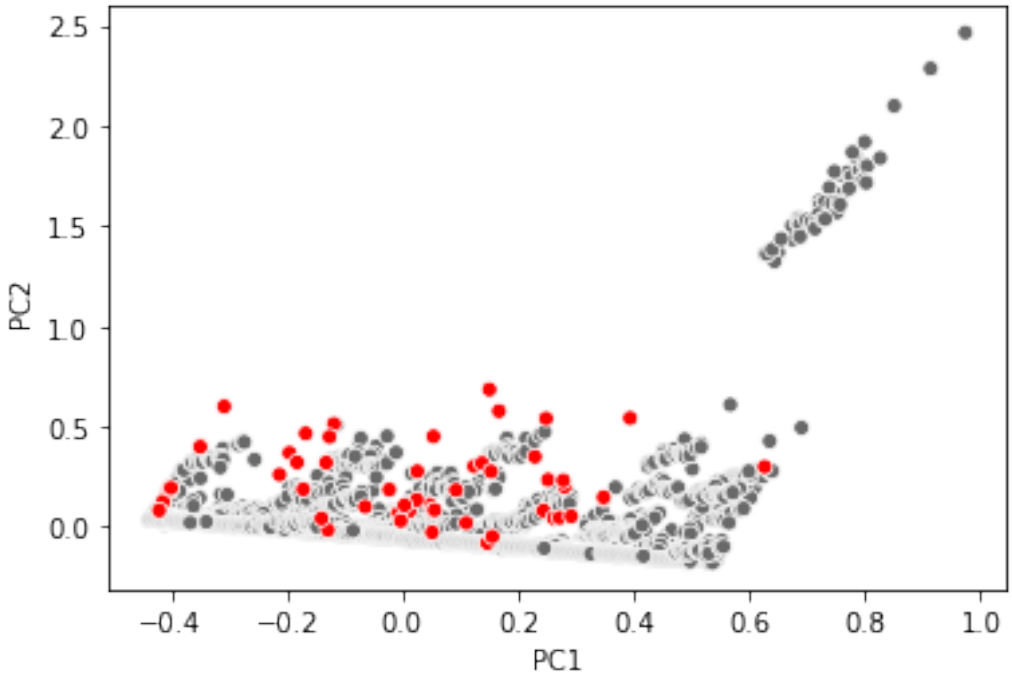


Figure 44: PC1 vs PC2

k-NN	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
k = 5	8.0732	0.4150	0.3335	0.6667
k = 10	10.8951	0.3396	0.2737	0.6368
k = 15	12.8947	0.3018	0.2438	0.6219
k = 20	15.3803	0.2830	0.2337	0.6168

iForest	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
features = 1.0	20.2765	0.0000	-0.0513	0.4743
features = 0.9	17.2117	0.0188	-0.0400	0.4795
features = 0.8	17.2117	0.0188	-0.0400	0.4795
features = 0.7	18.0153	0.0188	-0.0380	0.4809

Autoencoder	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
5-4-2-4-5	19.4169	0.1698	0.1252	0.5626

Figure 45: Performance of the models

B Sector I Results

B.1 Sector: I, Anomalous Features: 5, Scaling: Standardization, PCA: no

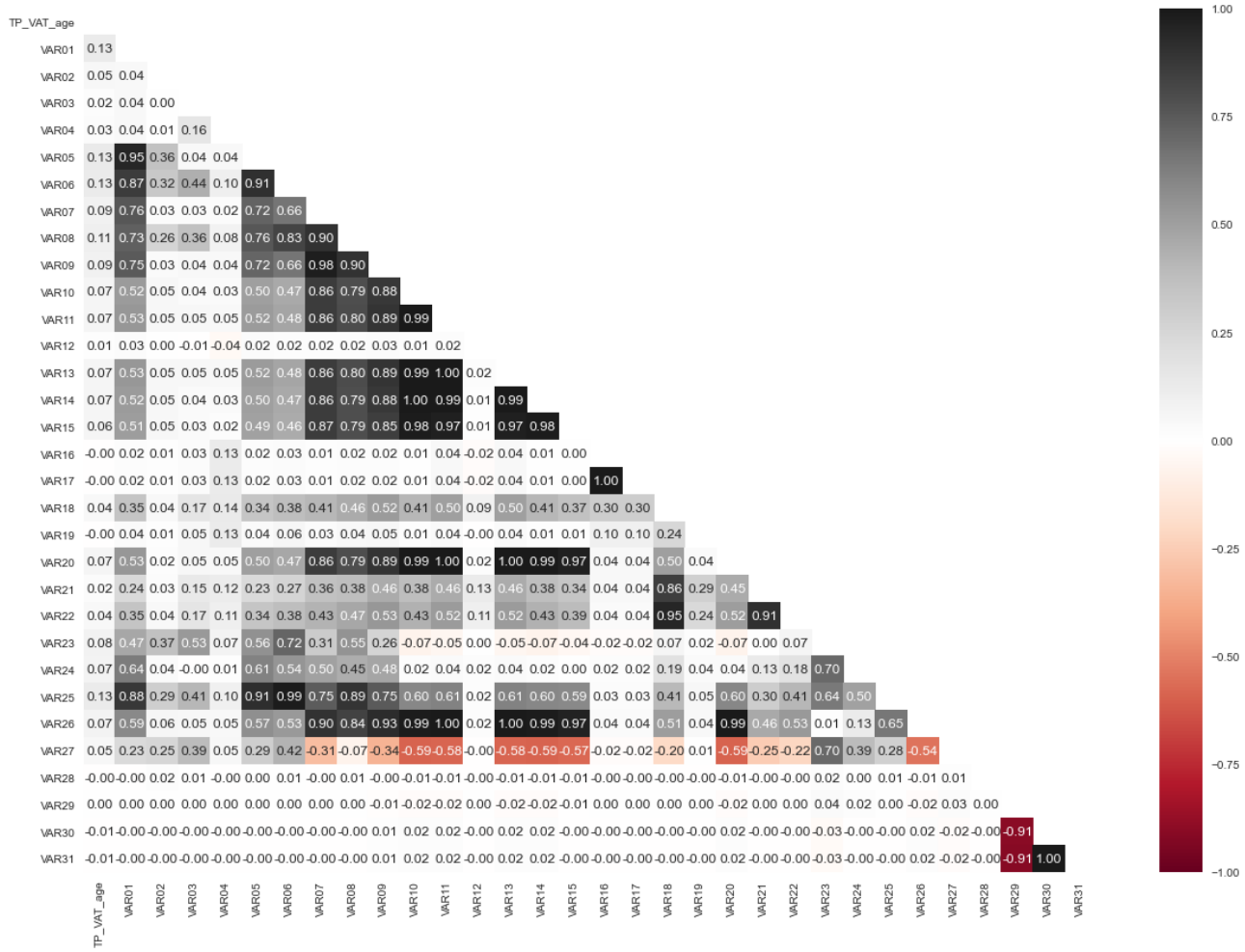


Figure 46: Correlation Matrix

k-NN	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
k = 5	8.2469	0.9255	0.9153	0.9576
k = 10	9.9339	0.9106	0.9007	0.9503
k = 15	11.0051	0.8975	0.8874	0.9437
k = 20	11.6541	0.8826	0.8717	0.9358

iForest	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
features = 1.0	48.2410	0.0000	-0.0209	0.4895
features = 0.9	51.3182	0.0000	-0.0196	0.4901
features = 0.8	57.9851	0.0000	-0.0173	0.4913
features = 0.7	65.8099	0.0000	-0.0152	0.4923

Autoencoder	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
32-25-3-25-32	20.8009	0.8417	0.8338	0.9169

Figure 47: Performance of the models

B.2 Sector: I, Anomalous Features: 5, Scaling: Standardization, PCA: yes

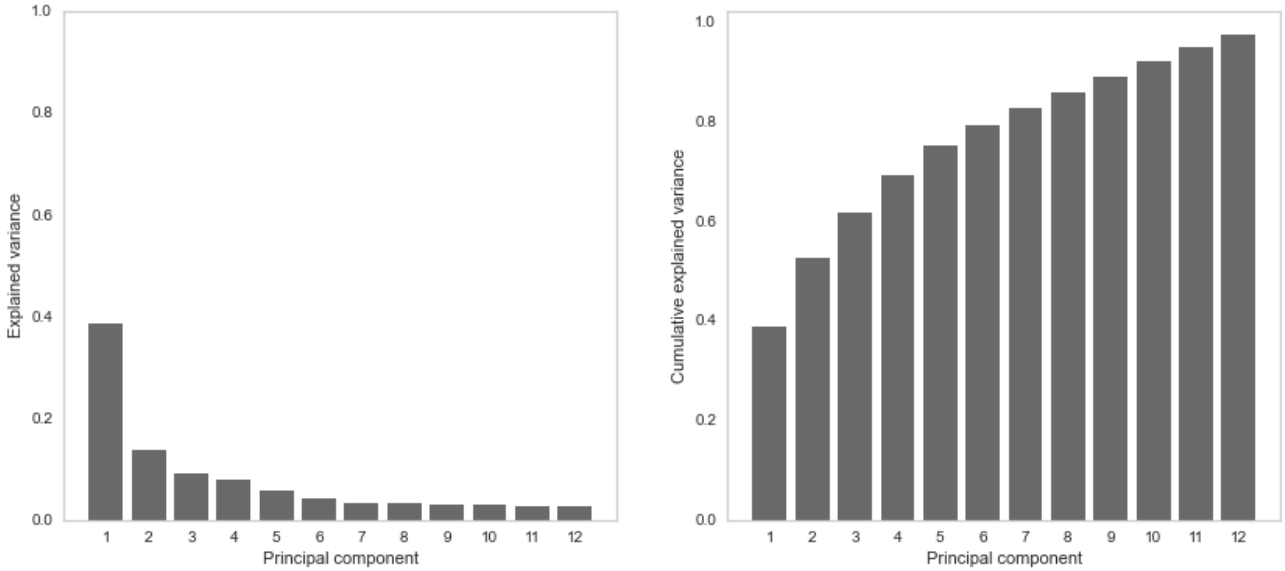


Figure 48: Explained Variance

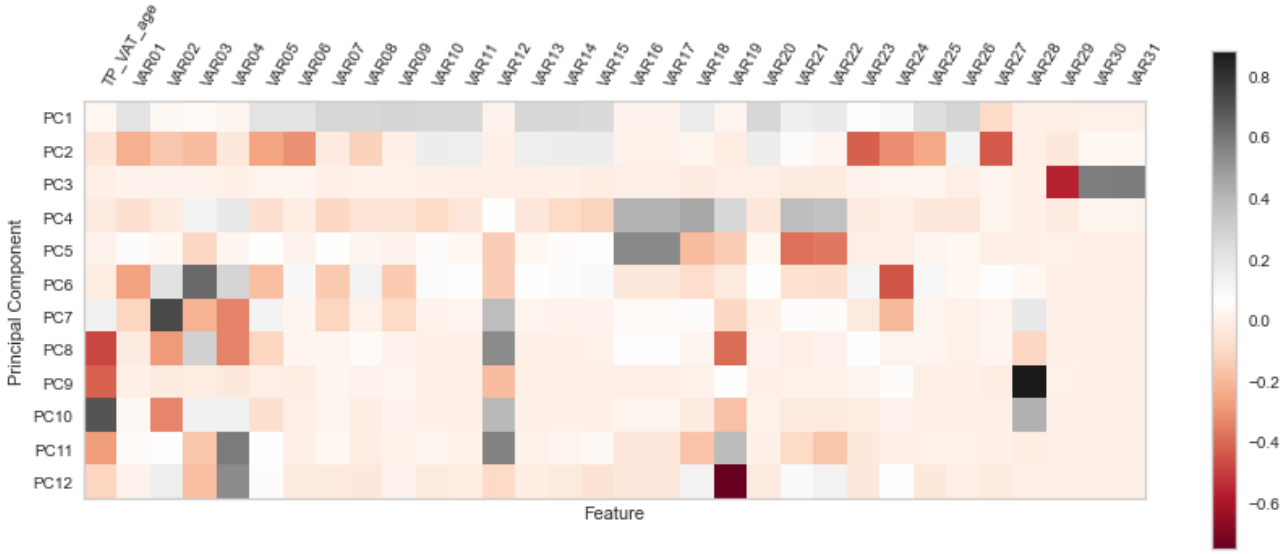


Figure 49: PCA heatmap

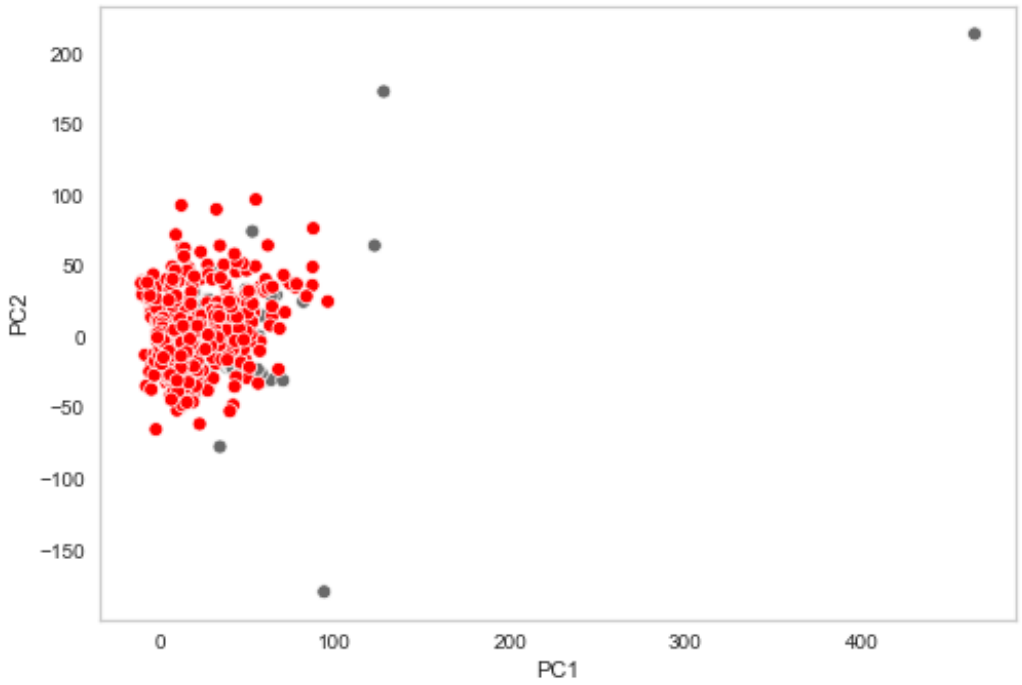


Figure 50: PC1 vs PC2

k-NN	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
k = 5	36.3416	0.8361	0.8315	0.9157
k = 10	39.7451	0.8137	0.8090	0.9045
k = 15	41.4893	0.8044	0.7996	0.8998
k = 20	41.2183	0.7951	0.7901	0.8950

iForest	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
features = 1.0	45.3279	0.5940	0.5849	0.7924
features = 0.9	42.1070	0.5865	0.5766	0.7883
features = 0.8	47.2749	0.5884	0.5796	0.7898
features = 0.7	50.5475	0.5679	0.5593	0.7796

Autoencoder	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	AUC
12-9-2-9-12	37.0146	0.6834	0.6747	0.8373

Figure 51: Performance of the models

B.3 Sector: I, Anomalous Features: 5, Scaling: Normalization, PCA: no

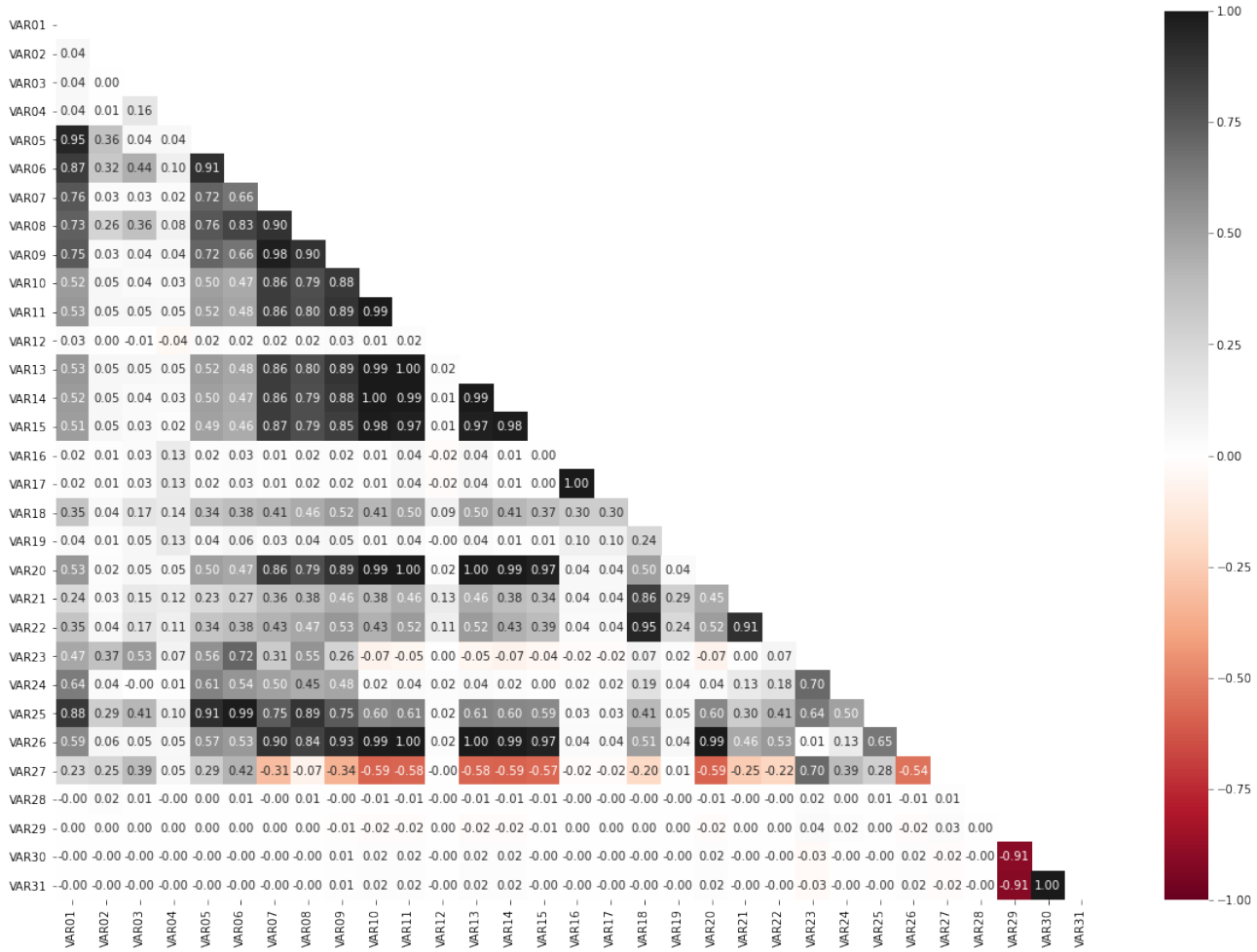


Figure 52: Correlation Matrix

k-NN	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
k = 5	19.3830	0.9031	0.8979	0.9489
k = 10	22.2924	0.8864	0.8811	0.9405
k = 15	24.7686	0.8733	0.8680	0.9340
k = 20	26.7672	0.8659	0.8607	0.9303

iForest	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
features = 1.0	54.1557	0.0000	-0.0186	0.4906
features = 0.9	51.3846	0.0000	-0.0196	0.4901
features = 0.8	57.8837	0.0000	-0.0170	0.4912
features = 0.7	65.9224	0.0000	-0.0152	0.4923

Autoencoder	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
32-25-3-25-32	42.4370	0.8715	0.8684	0.9342

Figure 53: Performance of the models

B.4 Sector: I, Anomalous Features: 5, Scaling: Normalization, PCA: yes

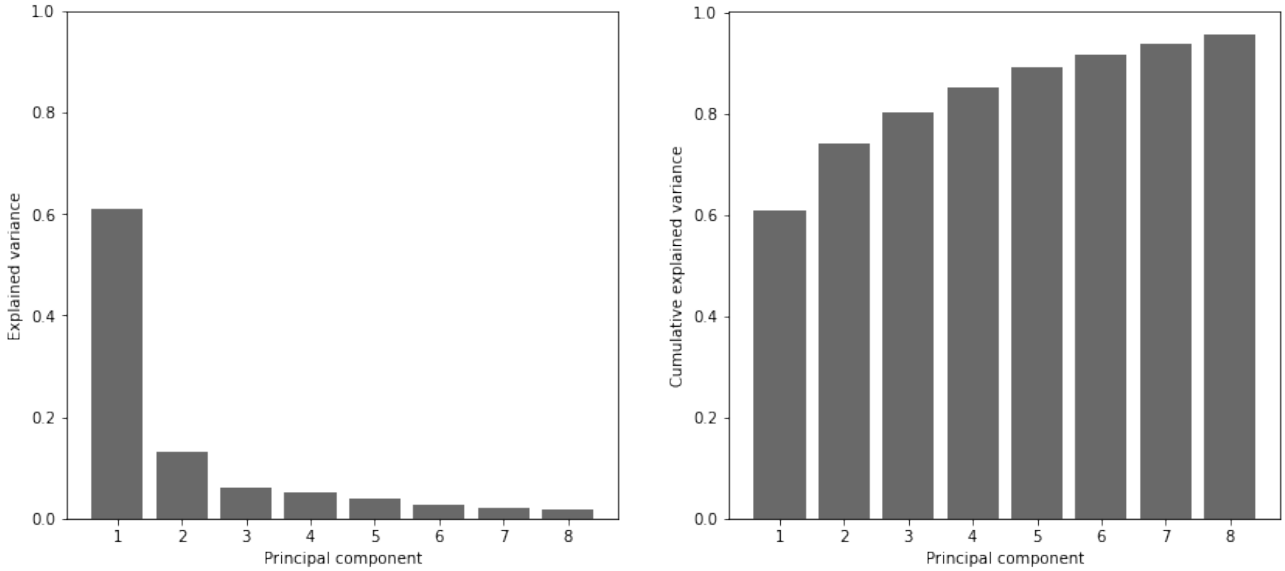


Figure 54: Explained Variance

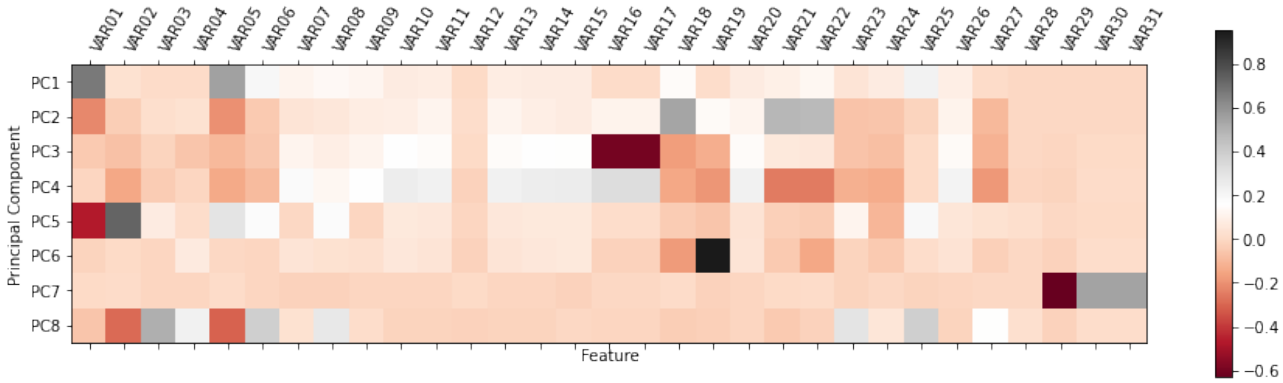


Figure 55: PCA heatmap

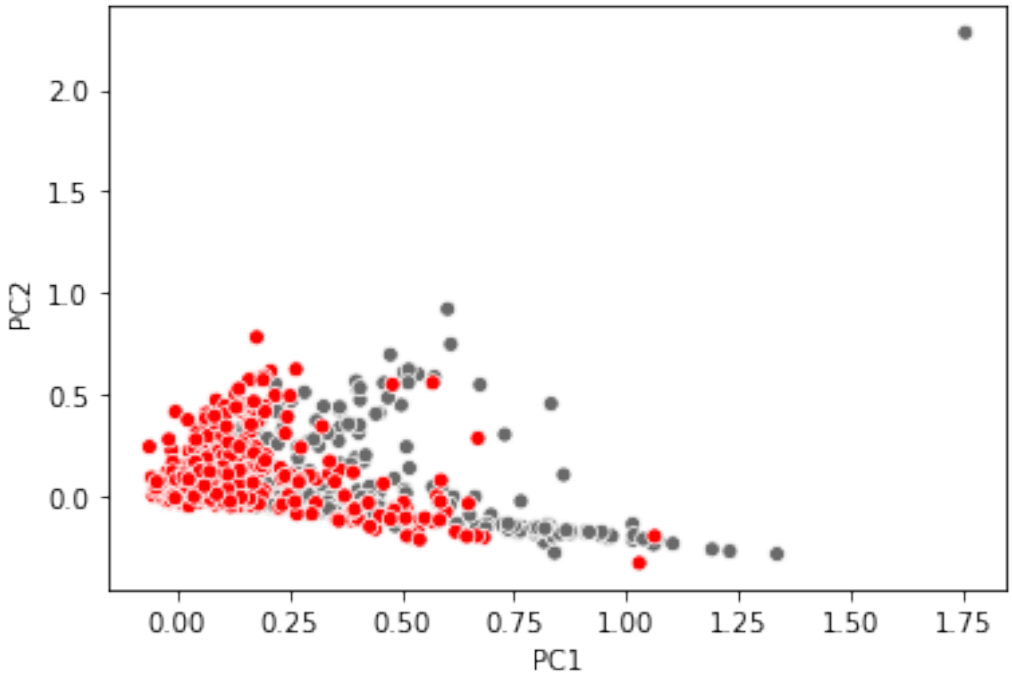


Figure 56: PC1 vs PC2

k-NN	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
k = 5	8.0017	0.8193	0.7938	0.8969
k = 10	8.6083	0.8026	0.7769	0.8884
k = 15	8.9199	0.7839	0.7570	0.8785
k = 20	9.1061	0.7653	0.7367	0.8683

iForest	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
features = 1.0	15.9556	0.5698	0.5413	0.7706
features = 0.9	18.8852	0.5363	0.5106	0.7553
features = 0.8	17.8196	0.5754	0.5504	0.7752
features = 0.7	19.6946	0.6461	0.6274	0.8137

Autoencoder	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
8-6-2-6-8	43.7700	0.6703	0.6627	0.8313

Figure 57: Performance of the models

B.5 Sector: I, Anomalous Features: 9, Scaling: Standardization, PCA: no

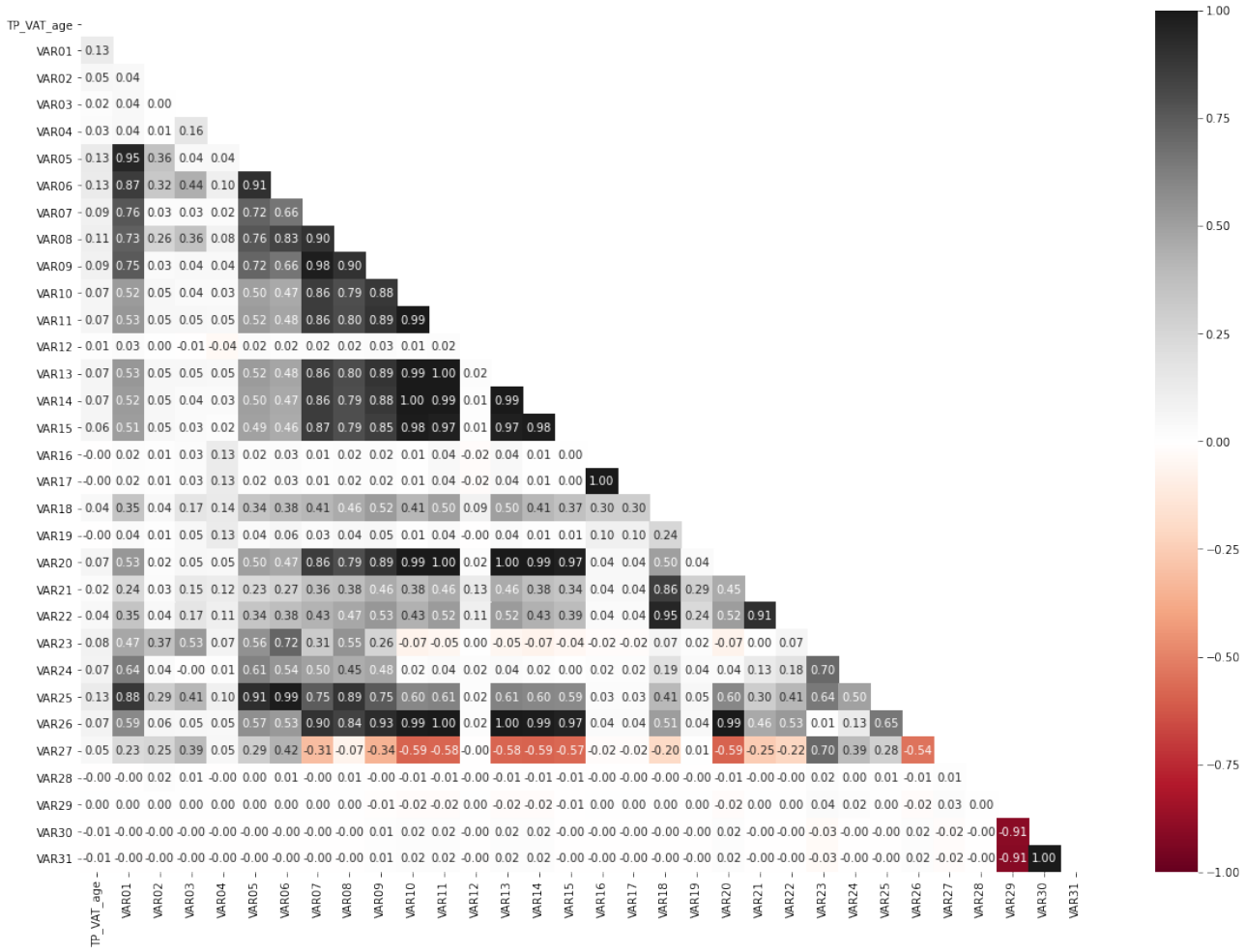


Figure 58: Correlation Matrix

k-NN	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
k = 5	1.3496	0.9739	0.9025	0.9512
k = 10	1.4657	0.9702	0.9084	0.9542
k = 15	1.5948	0.9571	0.8872	0.9436
k = 20	1.6648	0.9515	0.8807	0.9403

iForest	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
features = 1.0	18.8815	0.0000	-0.0553	0.4723
features = 0.9	18.1293	0.0018	-0.0557	0.4721
features = 0.8	15.5020	0.0000	-0.0680	0.4658
features = 0.7	17.9062	0.0000	-0.0585	0.4707

Autoencoder	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
32-25-3-25-32	2.9333	0.9404	0.9100	0.9550

Figure 59: Performance of the models

B.6 Sector: I, Anomalous Features: 9, Scaling: Standardization, PCA: yes

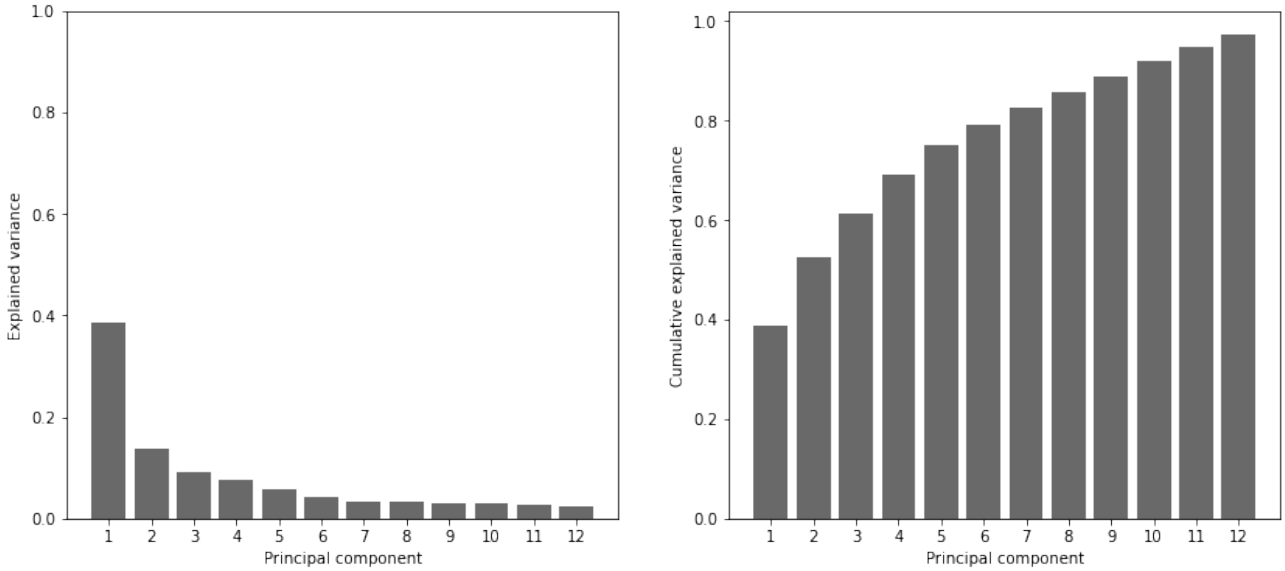


Figure 60: Explained Variance

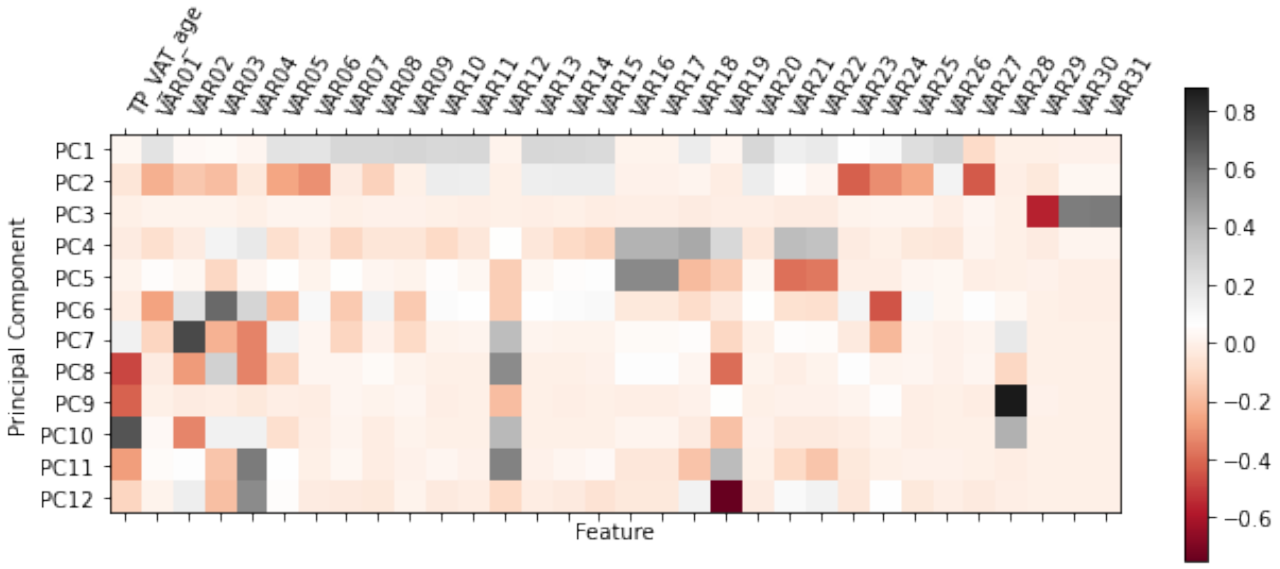


Figure 61: PCA heatmap

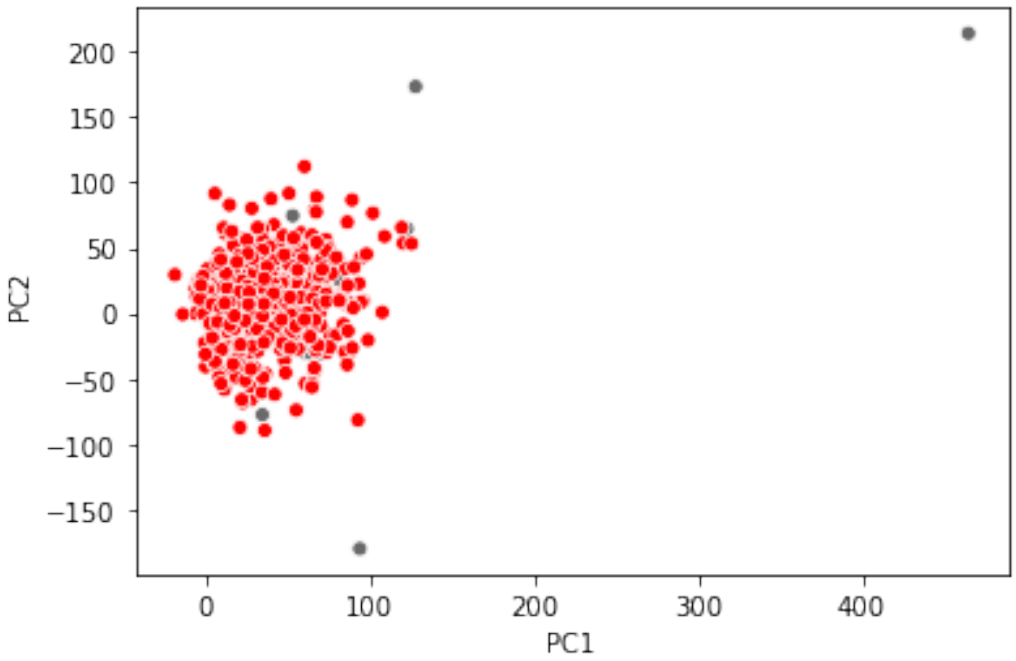


Figure 62: PC1 vs PC2

k-NN	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
k = 5	1.8511	0.9255	0.8400	0.9200
k = 10	2.1442	0.9143	0.8409	0.9204
k = 15	2.2309	0.8957	0.8126	0.9063
k = 20	2.3691	0.8789	0.7921	0.8960

iForest	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
features = 1.0	7.8870	0.6890	0.6443	0.8221
features = 0.9	6.5157	0.7448	0.6991	0.8495
features = 0.8	6.4862	0.7113	0.6593	0.8296
features = 0.7	6.5286	0.7206	0.6707	0.8353

Autoencoder	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
12-9-2-9-12	4.7180	0.8435	0.8020	0.9010

Figure 63: Performance of the models

B.7 Sector: I, Anomalous Features: 9, Scaling: Normalization, PCA: no

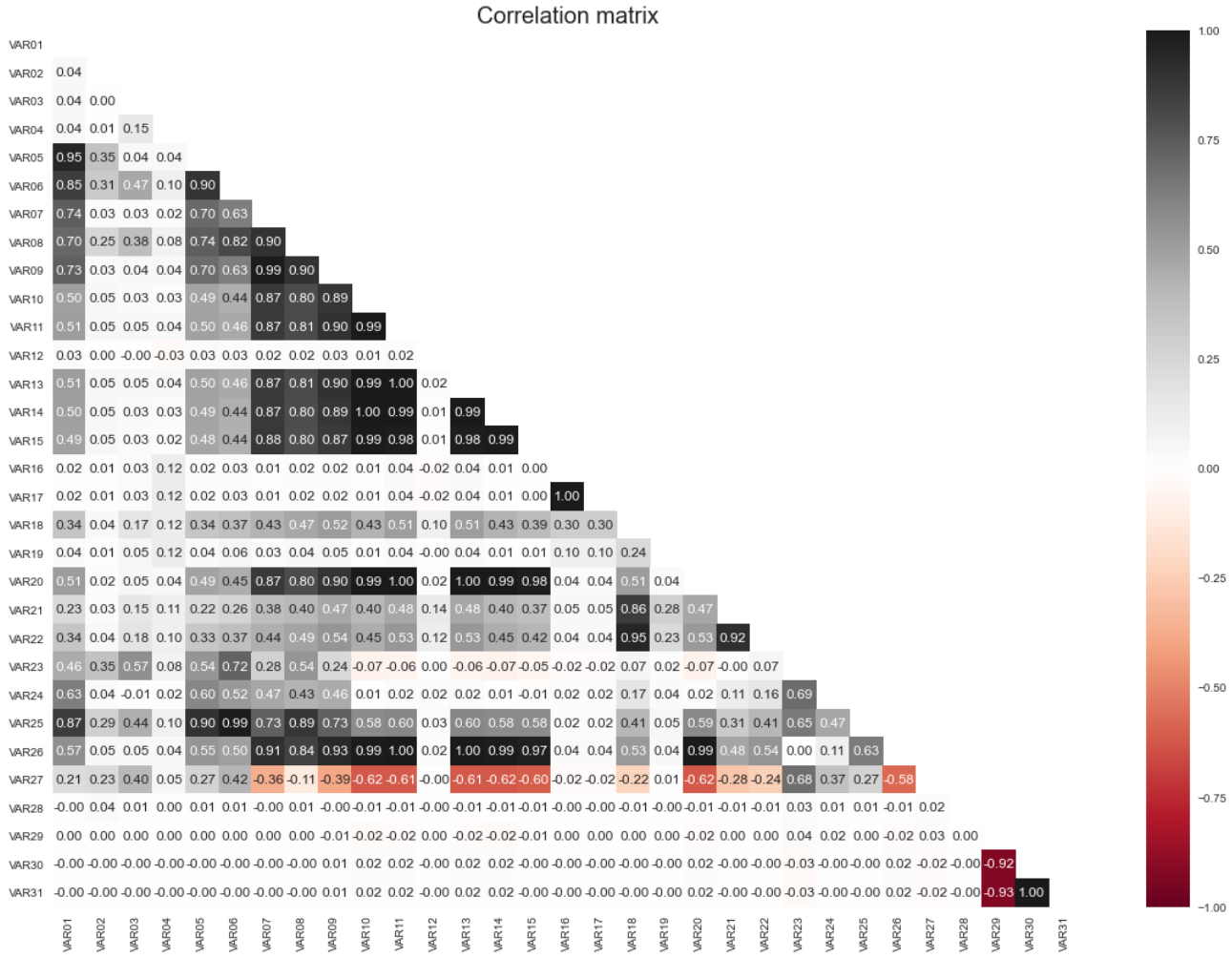


Figure 64: Correlation Matrix

k-NN	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
k = 5	1.2445	0.9757	0.8822	0.9411
k = 10	1.3367	0.9702	0.8855	0.9427
k = 15	1.4049	0.9571	0.8553	0.9276
k = 20	1.5100	0.9478	0.8489	0.9244

iForest	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
features = 1.0	19.5748	0.0000	-0.0532	0.4733
features = 0.9	22.2924	0.0000	-0.0460	0.4767
features = 0.8	19.5895	0.0037	-0.0493	0.4753
features = 0.7	20.4930	0.0000	-0.0507	0.4746

Autoencoder	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
31-24-3-24-31	1.5708	0.9627	0.8990	0.9497

Figure 65: Performance of the models

B.8 Sector: I, Anomalous Features: 9, Scaling: Normalization, PCA: yes

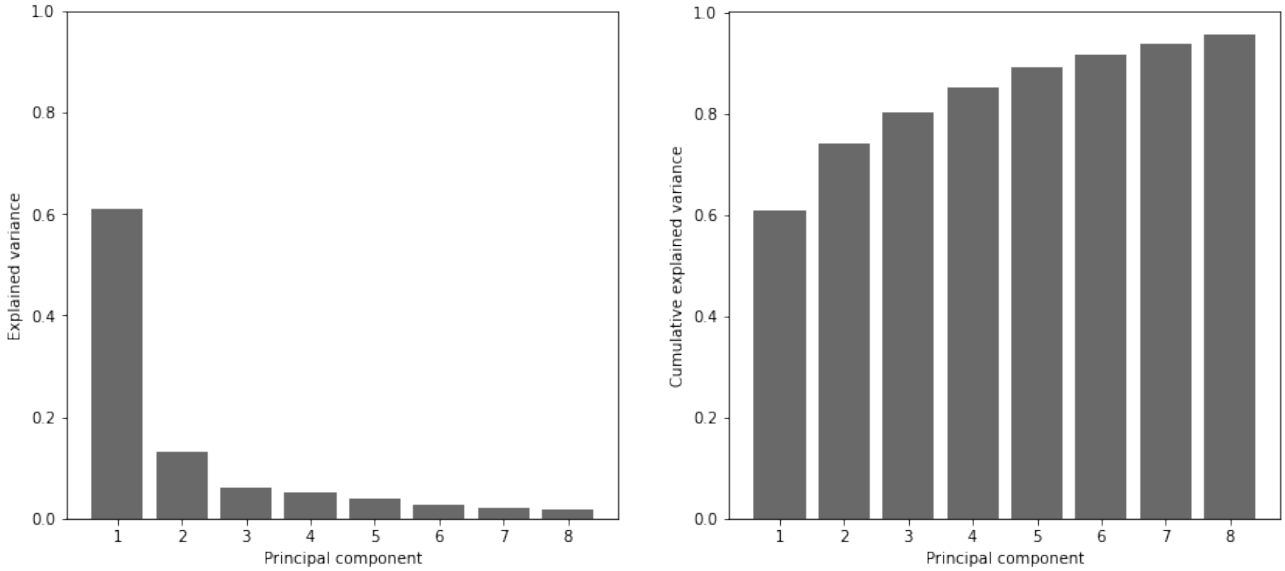


Figure 66: Explained Variance

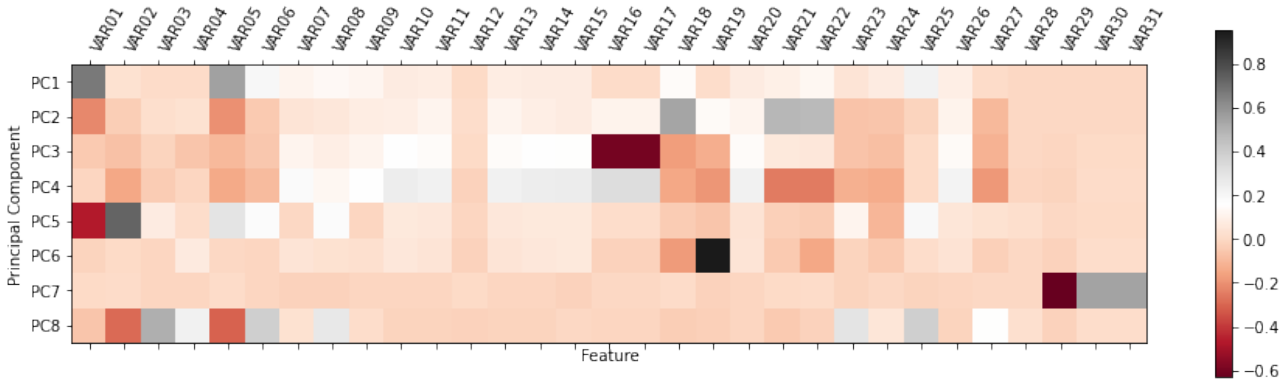


Figure 67: PCA heatmap

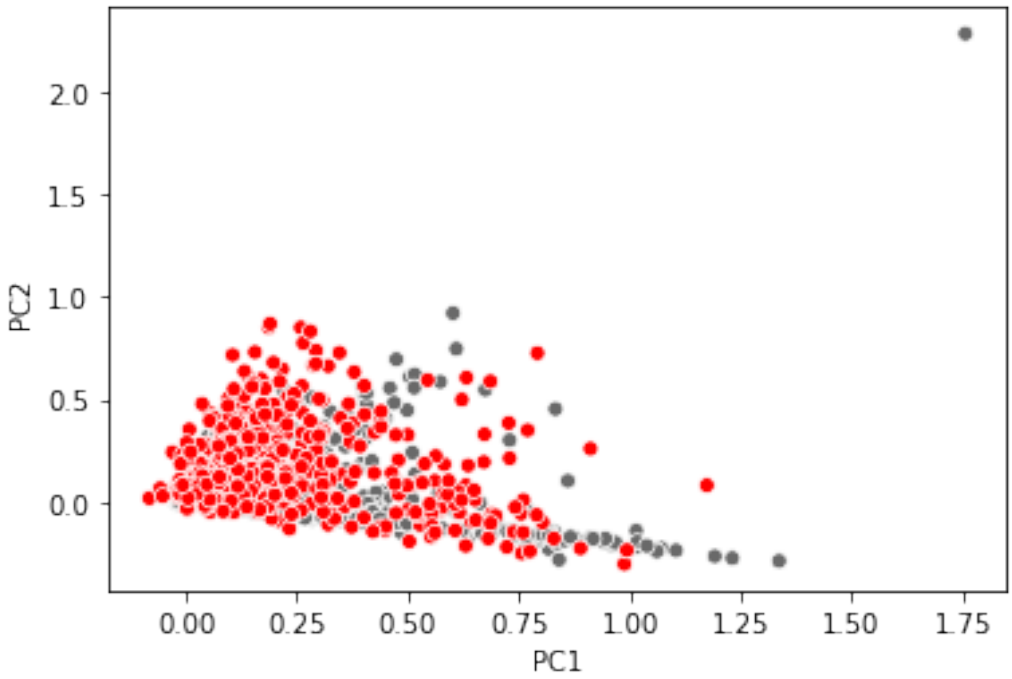


Figure 68: PC1 vs PC2

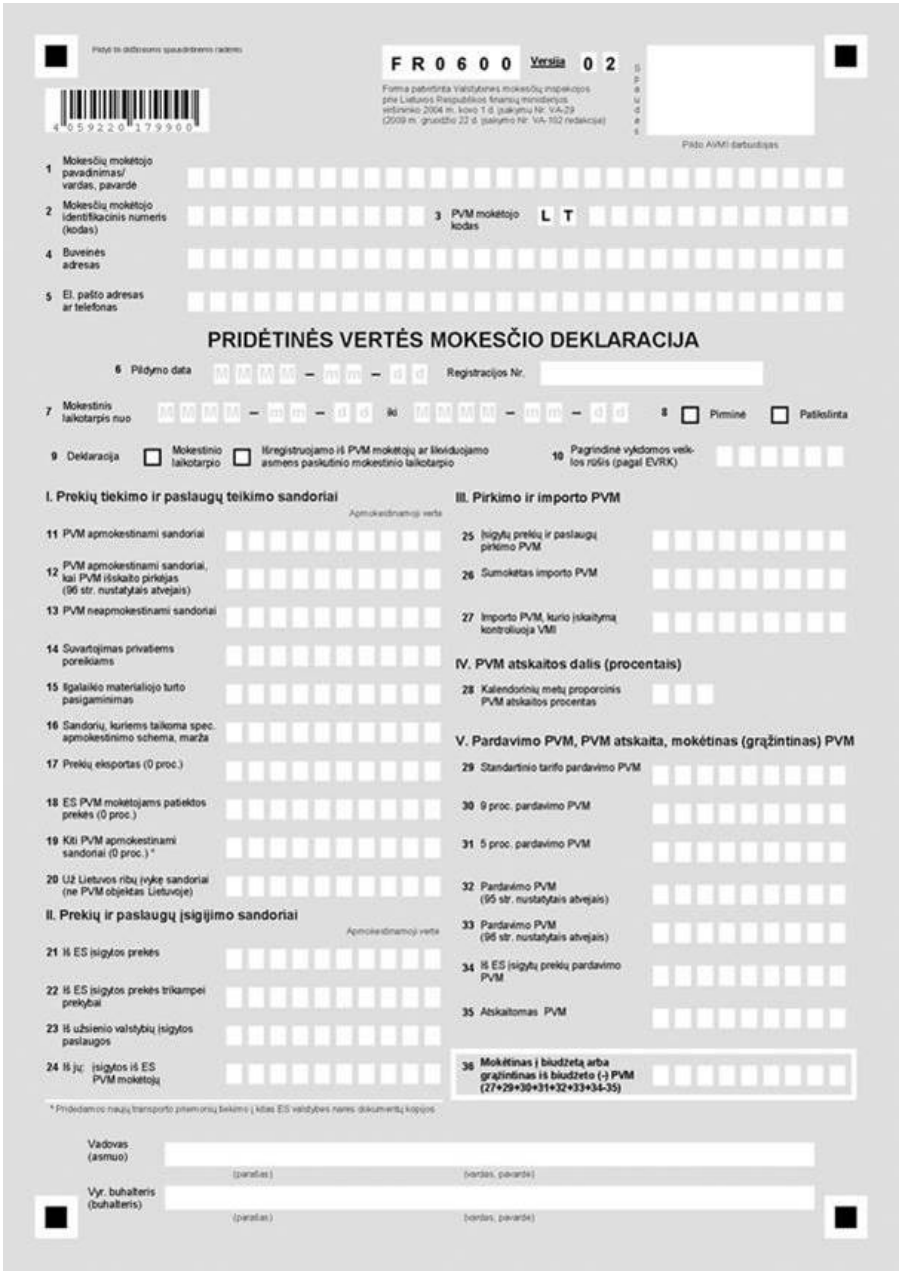
k-NN	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
k = 5	1.6538	0.9143	0.7869	0.8934
k = 10	1.8142	0.8864	0.7502	0.8751
k = 15	1.8160	0.8770	0.7301	0.8650
k = 20	1.9248	0.8752	0.7433	0.8716

iForest	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
features = 1.0	3.2136	0.7541	0.6448	0.8224
features = 0.9	3.0845	0.7076	0.5695	0.7847
features = 0.8	2.7010	0.7206	0.5592	0.7796
features = 0.7	2.5424	0.7635	0.6128	0.8064

Autoencoder	Percentage of Synthetic Outliers in 100 % recall	Precision in 100 % Recall	Matthews Correlation Coefficient in 100 % Recall	ROC AUC
8-6-2-6-8	2.2014	0.8454	0.7192	0.8596

Figure 69: Performance of the models

C FR0600 VAT declaration form



FR0600 Versija 02

PRIDĖTINĖS VERTĖS MOKESČIO DEKLARACIJA

1 Mokesčių mokėtojo pavadinimas/ vardas, pavardė

2 Mokesčių mokėtojo identifikacinis numeris (kodas)

3 PVM mokėtojo kodas LT

4 Buveinės adresas

5 El. pašto adresas ar telefonas

6 Pildymo data M M M M – m m – d d Registracijos Nr.

7 Mokestinis laikotarpis nuo M M M M – m m – d d iki M M M M – m m – d d

8 Pirminė Patikslinta

9 Deklaracija Mokestinio laikotarpio Išregistruojamo iš PVM mokėtojų ar likviduojamo asmens paskutinio mokestinio laikotarpio

10 Pagrindinė vykdomos veiklos rūšis (pagal EVRK)

I. Prekių tiekimo ir paslaugų teikimo sandoriai

Apmokestinamoji vertė

11 PVM apmokestinami sandoriai

12 PVM apmokestinami sandoriai, kai PVM išskaito pirkėjas (96 str. nustatytais atvejais)

13 PVM neapmokestinami sandoriai

14 Suvartojimas privatiems poreikiams

15 Ilgalaikio materialiojo turto pasigaminimas

16 Sandorių, kuriems taikoma spec. apmokestinimo schema, marža

17 Prekių eksportas (0 proc.)

18 ES PVM mokėtojams patiektos prekės (0 proc.)

19 Kiti PVM apmokestinami sandoriai (0 proc.) *

20 Už Lietuvos ribų įvykę sandoriai (ne PVM objektas Lietuvoje)

II. Prekių ir paslaugų įsigijimo sandoriai

Apmokestinamoji vertė

21 Iš ES įsigytos prekės

22 Iš ES įsigytos prekės trikampei prekybai

23 Iš užsienio valstybių įsigytos paslaugos

24 Iš jų: įsigytos iš ES PVM mokėtojų

III. Pirkimo ir importo PVM

25 Įsigytų prekių ir paslaugų pirkimo PVM

26 Sumokėtas importo PVM

27 Importo PVM, kurio įskaitymą kontroliuoja VMI

IV. PVM atskaitos dalis (procentais)

28 Kalendorinių metų proporcinis PVM atskaitos procentas

V. Pardavimo PVM, PVM atskaita, mokėtinas (grąžintinas) PVM

29 Standartinio tarifo pardavimo PVM

30 9 proc. pardavimo PVM

31 5 proc. pardavimo PVM

32 Pardavimo PVM (95 str. nustatytais atvejais)

33 Pardavimo PVM (96 str. nustatytais atvejais)

34 Iš ES įsigytų prekių pardavimo PVM

35 Atskaitomas PVM

36 **Mokėtinas į biudžetą arba grąžintinas iš biudžeto (-) PVM (27+29+30+31+32+33+34-35)**

* Pridedamos naujų transporto priemonių tiekimo į kitas ES valstybės narės dokumentų kopijos

Vadovas (asmuo) (parašas) (vardas, pavardė)

Vyr. buhalteris (buhalteris) (parašas) (vardas, pavardė)

D VAT Gap as a percent of the VTTL in EU-28 Member States[48]

	2017				2018				VAT Gap Change (pp)
MS	Revenues	VTTL	VAT Gap	VAT Gap (%)	Revenues	VTTL	VAT Gap	VAT Gap (%)	
BE	29,763	33,619	3,856	11.5%	31,053	34,670	3,617	10.4%	-1.0
BG	4,664	5,313	649	12.2%	5,097	5,711	614	10.8%	-1.5
CZ	14,703	16,694	1,991	11.9%	16,075	18,261	2,187	12.0%	0.0
DK	27,966	30,475	2,509	8.2%	29,121	31,369	2,248	7.2%	-1.1
DE	226,582	248,382	21,800	8.8%	235,130	257,207	22,077	8.6%	-0.2
EE	2,149	2,286	137	6.0%	2,331	2,458	127	5.2%	-0.8
IE	13,060	14,652	1,592	10.9%	14,175	15,857	1,682	10.6%	-0.3
EL	14,642	21,898	7,256	33.1%	15,288	21,858	6,570	30.1%	-3.1
ES	73,970	79,003	5,033	6.4%	77,561	82,470	4,909	6.0%	-0.4
FR	162,011	173,840	11,829	6.8%	167,618	180,406	12,788	7.1%	0.3
HR	6,465	6,843	378	5.5%	6,946	7,198	252	3.5%	-2.0
IT	107,576	142,939	35,363	24.7%	109,333	144,772	35,439	24.5%	-0.3
CY	1,765	1,859	93	5.0%	1,951	2,028	77	3.8%	-1.2
LV	2,164	2,512	348	13.9%	2,449	2,705	256	9.5%	-4.4
LT	3,310	4,422	1,111	25.1%	3,522	4,754	1,232	25.9%	0.8
LU	3,433	3,525	92	2.6%	3,729	3,928	199	5.1%	2.5
HU	11,729	13,564	1,835	13.5%	12,950	14,140	1,190	8.4%	-5.1
MT	810	984	174	17.7%	920	1,084	164	15.1%	-2.5
NL	49,833	52,329	2,496	4.8%	52,619	54,897	2,278	4.2%	-0.6
AT	28,304	30,949	2,645	8.5%	29,323	32,231	2,908	9.0%	0.5
PL	36,330	42,374	6,044	14.3%	40,411	44,862	4,451	9.9%	-4.3
PT	16,810	18,872	2,062	10.9%	17,865	19,754	1,889	9.6%	-1.4
RO	11,650	17,727	6,077	34.3%	12,890	19,485	6,595	33.8%	-0.4
SI	3,482	3,640	159	4.4%	3,765	3,913	148	3.8%	-0.6
SK	5,919	7,362	1,443	19.6%	6,319	7,899	1,579	20.0%	0.4
FI	20,404	21,510	1,106	5.1%	21,364	22,171	807	3.6%	-1.5
SE	44,115	44,987	872	1.9%	43,433	43,739	306	0.7%	-1.2
UK	162,724	184,706	21,982	11.9%	168,674	192,126	23,452	12.2%	0.3
Total EU-28	**1,086,332**	**1,227,266**	**140,935**	**11.5%**	**1,131,912**	**1,271,953**	**140,042**	**11.0%**	**-0.5**
Median				**10.9%**				**9.2%**	

Source: own calculations.