10-2023

Predicting Payment Defaults in Customs Authorities

Ahmed Fayed Elnahel
afe6589@rit.edu

Predicting Payment Defaults in Customs Authorities

by

Ahmed Fayed Elnahel

A Thesis Submitted in Partial Fulfilment of the Requirements for the

Degree of Master of Science in Professional Studies:

Data Analytics

Department of Graduate Programs & Research

Rochester Institute of Technology (RIT DUBAI)
Oct. 2023

RIT Dubai

Master of Science in Professional Studies:
Data Analytics

Graduate Thesis Approval

Student Name**:** **Ahmed** **Fayed** **Elnahel**

Thesis Title**: Predicting Payment Defaults in Customs**

Graduate Committee:

Name: Dr. Sanjay Modak **Date:**
Chair of the committee

Name: Dr. Khalil Al Hussaeni **Date:**
Member of committee

Acknowledgments

I would like to express my sincere gratitude to all those who have contributed to the completion of this thesis on predicting payment defaults in customs.

Firstly, I would like to thank my thesis advisor, Dr. Khalil Al Hussaeni, for his guidance, support, and insightful feedback throughout the entire research process. Dr. Ernest Fokoue, and Dr. Yannis Karamitsos for the valuable course materials and sample code. Their expertise and knowledge in the field of machine learning have been invaluable to the success of this study.
I would also like to extend my thanks to the staff and officials at the customs head office who provided the data for this study and for their cooperation and assistance throughout the data collection process.

I am grateful to my family, friends, and colleagues who provided their encouragement, support, and understanding during this challenging but fulfilling journey. Their belief in my abilities has been a constant source of motivation and inspiration.

Lastly, I would like to acknowledge the many researchers, scholars, and professionals whose work has contributed to the development of the field of customs and trade finance. Their insights and knowledge have been instrumental in shaping this thesis and advancing the field as a whole.

Abstract

As the global economy continues to expand, international trade is growing at an unprecedented rate, resulting in a surge of cargo movements across borders. With so much activity taking place, it's imperative that customs systems operate intelligently to identify potential risks in every transaction and highlight any potential fraud or manipulation that may be occurring.

Customs play a crucial role in supporting legitimate trade, protecting society, and promoting sustainable economic development. However, it is unfortunate that some companies take advantage of the facilities provided by customs departments to avoid paying their duties or engaging in smuggling from the free zone to local market activities.

According to data provided by the customs departments in the UAE, there has been a concerning increase in the number of such payment defaults over the years. This not only undermines the fairness and integrity of the customs system but also has a negative impact on the broader economy and society at large. Therefore, customs authorities must remain vigilant and take proactive measures to prevent and deter such illicit activities. By doing so, they can ensure that legitimate businesses are not put at a disadvantage and that the benefits of international trade are shared fairly and equitably.

CRISP-DM (Cross-Industry Standard Process for Data Mining) is a widely-used framework for conducting data mining and machine learning projects which will be utilized here. This research is comparing the performance of multiple machine learning techniques including Linear Discriminant Analysis LDA, Random Forest (RF), Quadratic Discriminant Analysis (QDA), Logistic Regression (LR), Naive Bayes (NB), Nearest Neighbours Learning Machine (KNN), Support Vector Machines (SVM), Stochastic Adaptive Boosting, Gaussian Processes, and Bagging to examine if we can predict the payment default in the transactional historical data of the free-zone companies.

The research proves that machine learning can be used to predict the payment defaults behaviour, and that the Random Forest model consistently performed better than other evaluated models.

Keywords: Payment default, delinquent prediction, default prediction, machine learning, customs, CRISP-DM, Logistic regression, Random Forest, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Support Vector Machines, Nearest Neighbours Learning Machine, Naive Bayes, Bagging, Boosting.

TABLE OF CONTENTS

TABLE OF FIGURES

Chapter 1 – Introduction

1- Background Information

Freezone companies in the UAE are offered a valuable opportunity to claim exemption from Custom Duty when re-exporting cargo from the freezone to the rest of the world. This service provides significant facilitation for businesses operating within the Freezone. However, it has come to customs' attention that some clients are claiming duty exemption without providing sufficient proof of the actual cargo exiting through the customs border. It is important to note that in order to fully comply with regulations and avoid penalties, it is essential for companies to provide proper documentation and evidence of cargo exit when claiming duty exemption. Therefore, it is imperative for companies to maintain proper records and ensure they are in full compliance with customs regulations when utilizing this service.

Figure 1 Fenced Freezone area with import and export activities

Figure 1 shows an example UAE freezone warehouses to store different commodities temporary for re-export or to consume in the local market.

2- Problem Statement

The topic of payment defaults in the customs domain has not received as much attention as payments defaults related to other tax authorities and bank loans. Even in customs world, some clients regularly default on their obligation to pay the deferred duty invoices, disregarding customs requirements. especially in the cases for freezone import and re-export scenarios.

In situations where clients fill a re-export customs declaration but fail to provide evidence of the actual exports, customs rules allow for a two-month grace period to comply. However, if no evidence of goods exit is provided within this period, customs will issue a forfeiture for the non-remittance (NR) virtual deposit, and customs will demand a duty invoice based on the value of the exported goods. This is the point at which some clients may default on their obligations. Customs authorities require a mechanism to avoid these defaults that lead to revenue losses.

3- Research Aim and Objectives

Over the past five years, the problem of payment defaults has steadily increased, as demonstrated by the following, Figure 2, which depicts a growing number of defaulting transactions over time.

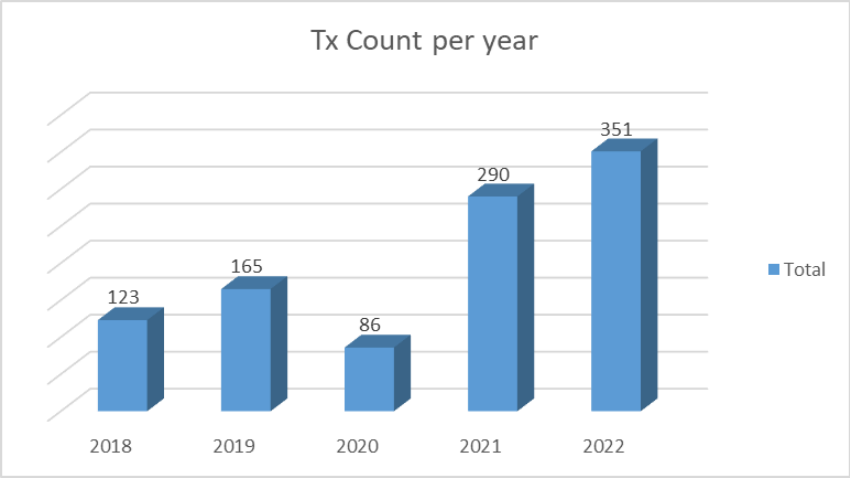

Figure 2 payment defaults, graph is generated from the customs dataset

The research results are expected to have a significant impact on the duty collection process within the customs domain, particularly in relation to the aged invoices where strict controls are currently lacking. As far as my knowledge extends, no previous studies have investigated this specific financial area of customs, making this research a pioneering contribution to the field. The insights gained from this study have the potential to benefit other customs organizations and advance knowledge in this important area.

The aim of the research is to know whether we can effectively predict which companies are more likely to default using the available data. By accurately predicting payment defaults, we can take appropriate measures such as requiring the company to provide a cash deposit when declaring their exports, rather than relying on the current deposit deferring facility. This can help mitigate the risks associated with non-payment and ensure that the exporting company fulfils its financial obligations.

4- Research Questions

The research tries to test the following hypothesis and answer the main business questions:

- Whether the defaulting companies are avoiding the customs' closing audit period, the related feature is the remaining period to the license expiry date.
- Whether there is possible smuggling to local market activity, the related feature is the "Gatepass" count which indicates the coexistence of a gate-passes to local market at the same time of the export declaration.
- The main hypothesis is the feasibility to predict the defaulting companies during the time of submitting a re-export transaction, which will be tested by evaluation of 10 different machine learning models.

5- Limitations of the study

The study of predicting payment defaults in customs may has several limitations that should be taken into consideration.

- The data used for analysis might not be entirely representative of the population of interest, as it may contain biases or errors.
- The model used for prediction might not be accurate enough due to limitations in the algorithms or libraries used.
- The study might only consider a limited number of variables, neglecting important factors that could have an impact on the prediction of payment defaults but were not available in the datasets.
- The study may have limited generalizability, as it only focuses on a specific industry or geographical region.

Lastly, the study may not be able to take into account unexpected events or changes that could affect payment defaults, such as economic crises or global pandemics.

6- Structure of the Thesis

The thesis is organized into several chapters, each dedicated to exploring different aspects of predicting payment defaults.

In the first chapter, an introduction is provided, outlining the research question and objectives. This sets the foundation for the study and establishes its scope.

Moving on to the second chapter, a comprehensive review of the relevant literature is conducted. This includes an examination of previous research conducted on the topic, exploration of the underlying theories and concepts, and an assessment of the models and methods employed for prediction.

The third chapter delves into the methodology employed for the study. It covers the process of data collection and analysis, detailing the sources of data utilized, the methods employed for data cleaning and preparation, and the statistical techniques applied for analysis.

In the fourth chapter, the results of the analysis are presented. This includes the development of a predictive model and the key findings derived from the study's investigation.
The fifth chapter engages in a discussion of the implications of the study for relevant stakeholders such as customs, and policymakers. This highlights the practical implications and potential applications of the research findings.

Finally, the last chapter serves as a summary, consolidating the main findings of the study. It also addresses the limitations encountered during the research process and provides recommendations for future research. The chapter concludes the thesis, wrapping up the entire study.

Chapter 2 – Literature Review

The issue of payment defaults in customs is a significant concern for both importers and customs authorities. This literature review will explore various studies on this topic and provide an overview of the current state of knowledge.

One study by (Abedin et al., 2021) investigated the factors that contribute to payment default in tax authorities. His methodology focuses on feature transformation by various approaches, and evaluated 13 state-of-the-art machine learning methods from business domains, The author found that Extreme Gradient Boosting (XGBoost) and Random Forest performed better than other machine learning algorithms in classification accuracy and other performance measures. However, the dataset had information about the company's financial statements which is not available in the case of customs data.

Another study (Seify et al., 2022) developed CNN model to detect fraudulent invoices and inflated prices, which were able to scan monthly 12 million invoices just in 5 days that manually took about 60 months. However, the details of the model configuration and accuracy were not mentioned in the research. Although detecting undervalued invoices is valuable for customs, the main focus here is detecting the duty payment default.

In a similar vein, (Vanhoeyveld et al., 2020) examined unsupervised anomaly detection, and the author found that it shows high predictive power for VAT in the area of tax evasion, however, no other methods were evaluated to compare the results, the study analyses a unique dataset containing the VAT declarations and client listings of all Belgian tax administration, the data includes companies sales and purchases unlike customs data which contains only companies cross border trade data.

The study (Vanhoeyveld, J., Martens, D. and Peeters, B. 2020) employed anomaly detection techniques to detect tax fraud cases among the companies in similar sectors. It shows that the predictive performance and the optimal method are sector dependent. The research studied the data sets provisioned by Belgian tax administration. Although the research yields positive findings, it requires different models and effort for optimization for different business sectors, which is a lot of effort to build such framework, besides the limitation in the unsupervised methods because it's not benefiting from the previous labelled audit results.

Another study by (Alsadhan, N. 2023), the study utilized three different models to detect frauds in tax payments, a tree-based model, unsupervised anomaly scoring model, and a behavioral module, which assigns a compliance score for each taxpayer. The three outputs are processed by a prediction module, which calculates the likelihood of fraud for each tax return. The framework was tested successfully on existent tax returns provided by the Saudi ZAkat, Tax, and Customs Authority (ZATCA). However, the study focused on the fraud related to a fraudulent increase in the purchase or an intentional decrease in companies' sales, which is basically different from cross borders imports scenarios.

A study by (Murorunkwere, B.F. et al. 2023) on large-scale data from the Rwanda Revenue Authority (RRA) resulted in a supervised neural network model to detect the tax evasion with 84% accuracy. The study reveals that African countries lose 10 times the international aids in the form of tax evasion. The research aims to identify factors related to tax fraud in both domestic and customs (imports and exports) businesses. Logistic Regression, Support Vector Machine (SVM),

Decision Tree, Random Forests, and eXtreme Gradient Boosting (XGBoost) were the supervised machine learning models used. The Artificial Neural Network identified fraud cases better than the Logistic Regression model, while still performing well on non-fraudulent cases.

Another study by (Murorunkwere, B.F. *et al.* 2022), Fraud Detection with Neural Networks used on income tax-related data from the Rwanda Revenue Authority (RRA). The study compared Artificial Neural Networks (ANN), support vector machines (SVM), and K-Nearest Neighbours. The results indicate that Artificial Neural Networks outperform other models in identifying tax fraud with 92% accuracy, 85% precision, 99% recall score, and a 95% AUC-ROC.

Another research by (González García & Mateos Caballero, 2022). This research utilized transaction-level import data from Spain between 2017 and 2019, with 20,000,241 import declarations. The author measures the Bayesian probabilities for the goods classification (HS Code) with the country of origin and highlight the suspected fraudulent cases. Gradient boosting was used on 43 features resulting in a significantly improved machine learning solution with an accuracy of ROC = 0.985 when compared to the actual system data.

An earlier study by the same authors (González García & Mateos Caballero, 2021), they used multi-objective Bayesian with dynamic optimization, the author studied the company's ownership, resources, and hidden wealth in Spain, and were able to highlight the risks related to those elements. However, there is no much details about the machine learning part of the study.

Furthermore, in a study by (Mojahedi et al., 2022) working on detecting payment defaults in tax domain. The dataset consists of 1500 samples with nine features, collected from the General Administration of Tax Affairs in West Azerbaijan Province. The research assessed the accuracy of different algorithms - Particle Swarm Optimization (IPSO) with support vector machine (SVM), Naive Bayes, k-nearest neighbour, C5.0 decision tree, and AdaBoost. IPSO-MLP and IPSO-SVM models achieved 93.68% and 92.24% accuracy, respectively. The data was tax related and has no companies trade information.

A further study by (Y. Wu et al., 2019) is about tax evasion detection method based on positive and un-labeled learning (TEDM-PU), the model has three stages the first one relies on Random Forest model to identify the key features, the final classification model is based on gradient boosting algorithm. The tax data is obtained from tax authorities in China for 20,444 taxpayers. After evaluation, the study demonstrates that TEDM-PU yields error rates that are on average 12.7% lower than other machine learning methods.

Another study by (Zumaya et al., 2021) conducted on a set of 81,511,015 taxpayers in Mexico, trying to detect tax evasion in Mexican companies, using a dynamic recurrent neural network (DRNN), evaluating deep neural networks and random forests, achieving an accuracy of more than 0.9 is possible with both methods.

Another study conducted by (Triepels et al., 2018) tackles the problem trying to detect smuggling actions across the border, the study confirms that data-driven fraud detection provides higher quality fraud alarms than random audits. Probabilistic discriminative models were derived from the Bayesian network, the former is trained to detect miscoding and smuggling from the shipments data. The model achieved a 35% average precision and 99% recall for miscoding, and 51% precision and 69% recall for smuggling.

Moreover, a study by (Fan Yu et al., 2003) About building a decision tree classification algorithm to detect fraudulent tax declarations, achieving an accuracy rate between 85-90%. The dataset

comprises one year's worth of data from 500 commercial enterprises. Each record contains 100 attributes sourced from the taxpayer's declaration form, related fiscal reports, and third-party departments.

Another study by (Höglund, 2017) has collected data source from Finnish tax authorities, data includes information about the company's financial statement and sales performance. The study identified solvency, liquidity, and payment period of trade payables as key predictors of tax defaults. Random Forest with 0.91 ROC, and Neural Network with F1-score of 0.87 were examined for tax evasion prediction.

One study by (Hua Shao et al., 2002), the main objective was to detect fraudulent behavior in customs declarations. The researchers utilized both star schema and snowflake schema to obtain various statistical measures about the data attributes. The study involved training a decision tree algorithm to predict instances of fraud. However, the authors did not report on the accuracy of the model or the specific types of fraud that were predicted.

Another research by (Bonchi et al., 1999) intended to detect tax evasion and fraudulent claims, the author tried to a posteriori detect frauds for the purpose of audit planning. The decision trees-based model is trained to distinguish between positive cases (fruitful audits) and negative cases (unfruitful audits), a 10-trees adaptive boosting gave 78% prediction accuracy. The research demonstrates how classification techniques support audit strategy planning.

An interesting study by (Zhu et al., 2018) proposed an Inter-Region Tax Evasion Detection method based on Transfer Learning to be used when sufficient tax data are not readily available. It integrates Transfer Adaboost (TrAdaBoost), Transfer Component Analysis (TCA), and LightGBM models. Data was obtained from tax authorities in China. The author was able to get acceptable accuracy and error rates using the transfer learning methods between regions.

In a study by (R.-S. Wu et al., 2012) The author analyzed VAT data using association rules to detect patterns of relationships between attributes and VAT evasion. The proposed data mining technique is a more scientific and resource-saving approach compared to manual screening. However, more advanced data mining models were not evaluated during the study.

Another study by (Baghdasaryan, V. et al. 2022) using data on prevalence of fraud within the supplier and buyer network of the taxpayers, enriching the data whenever there is no enough historical audit or fraud cases available. The study used machine learning algorithm for detection of tax fraud or evasion. Tree-based ensemble model-gradient boosting machines used on a tax returns data set from business entities in Armenia. Gradient boosting performance was compared to Random Forest, Decision tree, and Logistic Decision and found superior in performance.

A study by (Hooda, N., Bawa, S. and Rana, P.S. 2020) uses Ensemble learning as a powerful tool in machine learning to classify the Fraudulent Firm Prediction. It was tested on 776 firms out of 46 different cities from the Audit General Office (AGO) of India. Ten state-of-the-art classification methods namely decision tree (DT), adaboost (AB), random forest (RF), support vector machine (SVM), probit linear model (PLM), neural network (NN), decision stump (DSM), J48, Naive Bayes (NB), and Bayesian (BN) are employed to make the pool of classifiers called model-pool for ensemble building. Tehe accuracy of the Ensemble with majority voting exceeded 94%. The study didn't give much details about the used data set as input to the models.

Another study by (Ben Ismail, M.M. and AlSadhan, N. 2023) for Zakat Under-Reporting Detection which is a king of tax evasion. The study used a dataset including 51,919 Zakat declarations to evaluate the model. A deep neural network is designed to classify Zakat declarations into “underreporting” or “actual declaration” classes, and predict the expected tax gap. The results show that using SMOT technique to balance the classes in the training data yields a better accuracy than training the model on unsampled data. A classification accuracy of 99% reported in the study may indicates that the model is overfitting, we cannot validate that point as cross validation was not utilized in the research.

In another study by (Castellón González, P. and Velásquez, J.D. 2013) to detect the false invoices depending on the taxpayer information and historical payments. The study started by clustering the taxpayers into similar categories, then applied different machine learning techniques on each group. The study tested Decision Tree, Bayesian Networks, and Neural network model which performed better with accuracy between 89% and 92%. However, the accuracy of the decision tree and Bayesian models were not reported.

One study by (Ngah, Z.A., Ismail, N. and Abd Hamid, N. 2022) tried to find relation between interesting features like family-owned company, presence of tax professionals, company’s duration in business and frequency of tax audits, with the probability of fraudulent financial information and tax evasion. The study used least square regression analysis in predicting tax evasion through fraudulent financial reporting using a Malaysian tax data. However, there is not reported information about the model output, statistical significance of the features, and accuracy.

Another study by (González-Martel, C., Hernández, J.M. and Manrique-de-Lara-Peñate, C. 2021) to detect the mismatch in declarations of the Value added tax (VAT) by companies in the same network. The study tested random Forest model to classify four types of errors using the information of such declarations for a region in Spain during year 2002. The model was not compared to other potential models for evaluation.

In a study by (Alejandrino, J.C., P. Bolacoy, J.Jr. and Murcia, J.V.B. 2023) to compare different data mining approaches in loan default prediction. The author examines both supervised and unsupervised techniques including k-nearest neighbours (k-NN) 3, naïve Bayes and logistic regression. The study uses the company information and previous history to predict the defaults. Although the study mentioned the use of unsupervised methods, there was no mention of any specific unsupervised technique used during the study.

Another study in banking sector by (Madaan, M. et al. 2021), the researchers compare the use of decision tree and random forest to predict loan defaults. The author worked on publicly available Lending Club dataset from Kaggle. As expected, the Random Forest model performed with a higher accuracy than Decision Tree. The author should evaluate more models, such as boosting and bagging, for a better comparison.

Another study by (28. Aslam, U. et al. 2019), the research studied other past researches techniques in predicting the loan defaults, the reviewed studies examined multiple machine learning technique such as Decision trees, Neural Networks, logistic regression and support vector machine. The results showed that Decision tree performed the best with accuracy 96.60%, even better than neural network. However, the comparison may not be accurate or reliable as the various research studies used different data sets and techniques that are likely not comparable.

Another study by (Zhu, Q. et al. 2022), the authors tried a new innovative technique to combine Convolutional Neural Network CNN with LightGBM model for Loan Default Prediction. The CNN process the load data features and produce a transformed features that are consumed by the LightGBM model for training and prediction. The study compared this technique with four other traditional machine learning models and found that this approach outperformed the other tested methods.

A final most recent study by (Owusu, E. et al. 2023), the study examined loan default in online peer-to-peer lending activities. The dataset was imbalanced like the case at hand, and the study used a synthetic data to balance the classes as this research will review as well. The researchers examined a deep learning approach for loan default prediction with resulted accuracy of 94.1%. However, the study hasn't evaluated this model with other machine learning approaches for comparison.

Moreover, customs authorities can play a critical role in preventing payment defaults by implementing effective risk management strategies and providing support and training to importers and customs brokers.

Overall, these studies highlight the following,

- The successful use of machine learning in customs and tax authorities has enabled efficient prediction of delinquent payments and fraud, leading to increased compliance and reduced financial losses.
- While the tax and customs domains share certain similarities, the unique data features of each domain create significant differences that can hinder the direct application of tax-related research as a solution for customs use cases. Therefore, it's important to approach each domain separately and develop customized solutions tailored to their specific data characteristics.
- Recent research has shown that machine learning models can predict payment defaults both before and after transactions occur. A priori models can help prevent fraud, while a posteriori models can aid in audit planning.
- Tree-based models such as Random Forest, Boosting, and Bagging have been found to be particularly effective in predicting duty and payment fraud cases, providing a reliable and accurate tool for customs and tax authorities to prevent financial loss and enhance compliance.
- Based on the findings of numerous studies, it appears that payment history and past defaults are among the most important variables, indicating that these factors are likely to be equally important in the context of predicting customs defaults.

Chapter 3 – Research Methodology

The objective of this research is to leverage the historical payments data that is accessible to the customs department and use it to train machine learning models that produce accurate results. Unlike other studies that rely on client's company's financial statements and external information that may not be accessible to customs departments, this research makes optimal use of payments historical data and transaction information to generate comparative results with high prediction accuracy.

This study employs the CRISP-DM (Cross-Industry Standard Process for Data Mining) as a methodology. The CRISP-DM is a widely-used framework for conducting data mining and machine learning projects. As depicted in Figure 3, its structured and iterative approach provides a clear roadmap for researchers to follow, helping them to better plan, execute, and manage their research projects. Using CRISP-DM can help the research to stay organized, ensure collecting and analysing the right data, and make better-informed decisions based on the results of the analysis. By following the CRISP-DM framework, we can reduce the risk of failure, save time and resources, and increase the likelihood of producing valuable insights and outcomes.

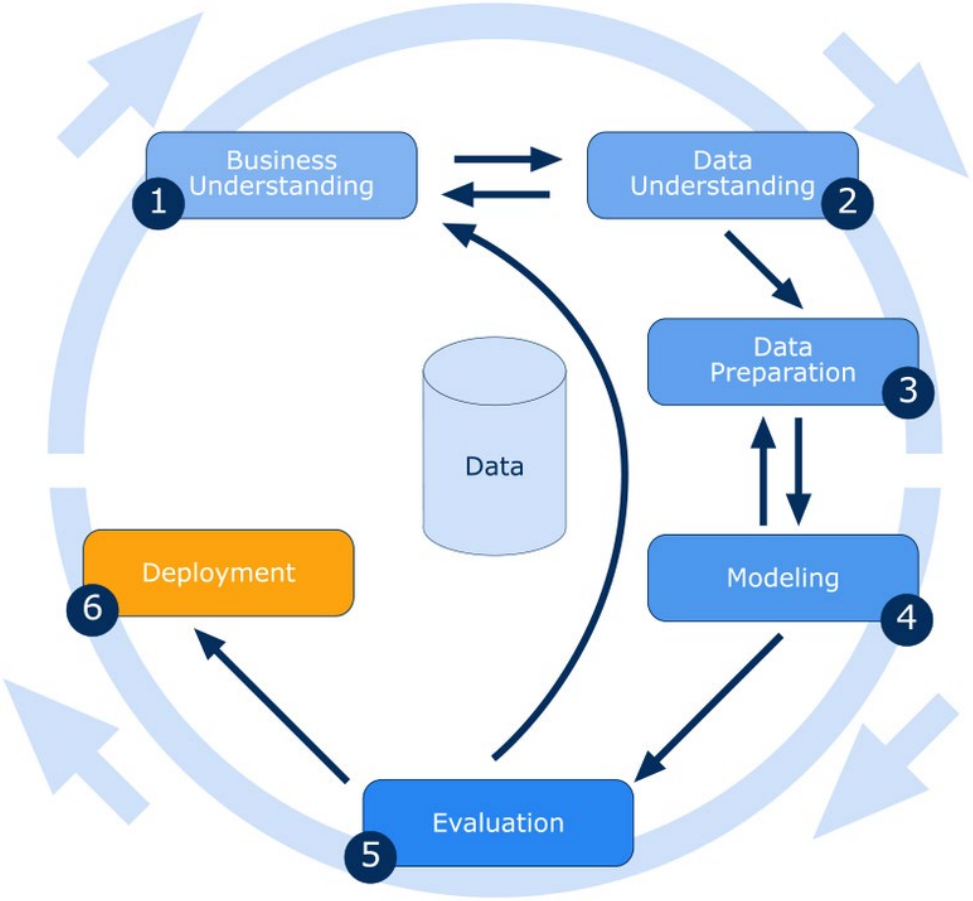

Figure 3 CRISP-DM, Image Source https://www.konato.be/crisp-dm/

1- Solution Design

The data is collected from customs financial ERP systems, the Enterprise Client Registration system (ECR), and the declaration system to be joined later for further processing, including data cleansing and imputation, and numeric attributes scaling, later the data is fed into ten different machine learning models for prediction accuracy evaluation. Figure 4 shows the data flow between the different components in the design.

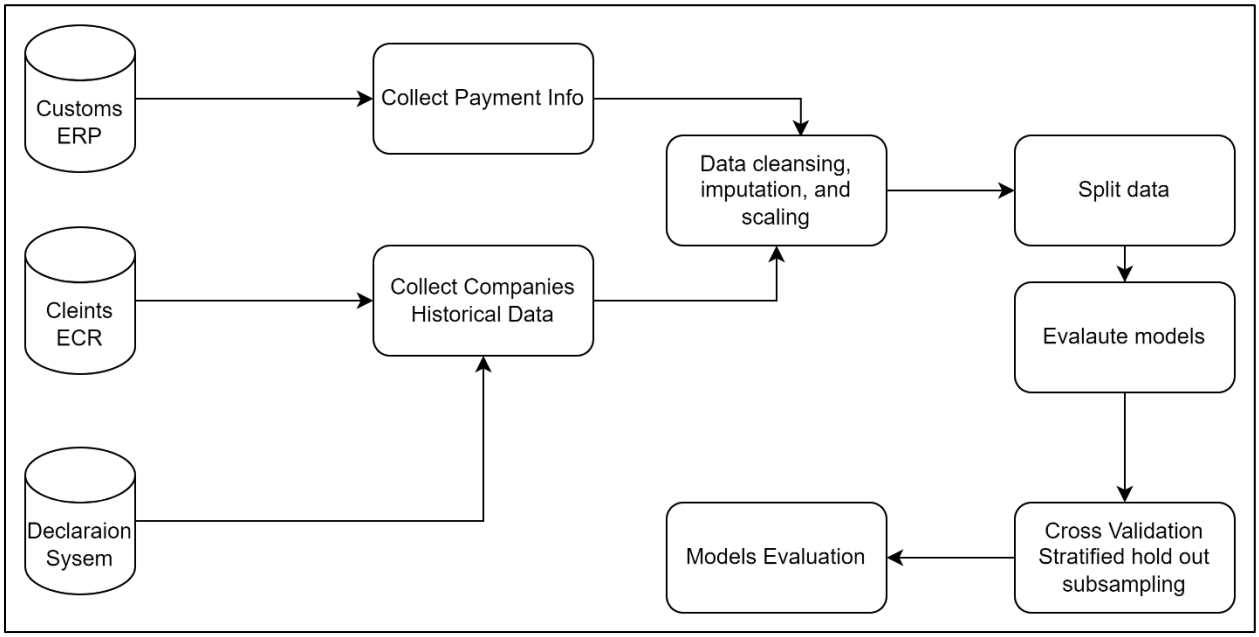

Figure 4 Solution Design

2- MLOps Platform and IDE

MLRun platform is used for data collection and processing pipeline, MLRun is an open source MLOps framework for quickly building and managing continuous ML applications across their lifecycle. For further processing of the data and for building the machine learning models, RStudio will be used which is an integrated development environment for R.

3- Machine Learning Techniques

When one has to deal with classification tasks for which the prior probabilities of class membership are very different, it becomes important to choose training and test sets that faithfully represent the original label distribution. This is where stratification of the subsampling becomes essential. The solution utilizes a function which performs stratified hold out subsampling which will be used for Cross validation (bootstrap) to obtain the models predictive performances.

Machine learning techniques including Linear Discriminant Analysis LDA, Random Forest (RF), Quadratic Discriminant Analysis (QDA), Logistic Regression (LR), Naive Bayes (NB), Nearest Neighbours Learning Machine (KNN), Support Vector Machines (SVM), Stochastic Adaptive Boosting, Gaussian Processes, and Bagging.

4- Performance Measurement

The accuracy is measured using ROC (Receiver Operating Characteristic) curve and confusion matrix measures to compare the different model's prediction capabilities. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) for all possible classification thresholds. TPR represents the proportion of actual positive instances that are correctly classified

as positive, while FPR represents the proportion of actual negative instances that are incorrectly classified as positive. A good classification model will have a high TPR and a low FPR across a range of thresholds, resulting in a curve that hugs the upper-left corner of the ROC plot.

The area under the ROC curve (AUC) is often used as a summary metric to compare the performance of different classification models. AUC ranges from 0.0 to 1.0, with a value of 0.5 indicating a random classification model, and a value of 1.0 indicating a perfect classification model. In general, a higher AUC indicates better performance.

The machine used for running the models has the following specifications: an 11th generation Intel(R) Core(TM) i7-11800H processor @ 2.30GHz, 2304 Mhz, 8 Core(s), 16 Logical Processor(s), 64.0 GB of RAM, and an NVIDIA® GeForce RTX™ 3050 TI Laptop GPU.

Chapter 4- Findings and Data Analysis

1- Dataset

The data collected from the customs department includes 18 attributes describing the company transactions information and the payment history, there are 4699 observations in the dataset labelled as settled or aged payments.

Feature	Type	Description
Expiry_Days	Number	The duration remaining till the company's trade license expiry date.
Payment_Aging	Number	The sum of days for all of the previous outstanding transactions for that client.
Country	Category	The mode of the import countries in the client declarations history
Commodity	Category	The mode of the commodity appears in the client declarations history
Total_Value	Number	The sum of the invoices' values in the client declarations history in the last 36 month
No_Days	Number	The period in days to the most recent client transaction.
Total_Weight	Number	The total cargo weight in the client declarations history in the last 36 month
Frequency	Number	The count of declarations in the last 36 month
Total_Income	Number	The total value of duties paid in the client declarations history in the last 36 month
AEO_Flag	Boolean	A flag for the clients in the Authorized Economic Operator program
Gatepass	Number	Count of the gate pass approved at the time of transit out.
DESTINATION_CODE	Category	The declaration destination is Ship Stores or any other.
LICENSE_ISSUING_AUTH_CODE	Category	The license issuing authority for the trade license.

Table 1 Data Dictionary of the customs' dataset

2- Data Description

Dataset has the following structure, for each attribute it shows the minimum, maximum, mean, median, and number of null values.

This is the data description and structure after transforming the categorical attributes into number ones.

Variable <chr>	Min <dbl>	Median <dbl>	Mean <dbl>	Max <dbl>	NA.Count <dbl>
CALCULATED_AMOUNT	1	7417	1.010399e+05	8369556	0
DESTINATION_CODE	1	1	1.003192e+00	2	0
PAYMENT_AGING	0	0	1.298915e+03	106373	0
GATEEPASS	0	3	1.819919e+01	2903	0
AEO_FLAG	0	0	4.469036e-02	1	0
LICENSE_ISSUING_AUTH_CODE	1	9	8.390083e+00	9	0
TOTAL_INCOME	0	45910	1.656187e+05	13410321	0
FREQUENCY	0	490	1.888453e+03	66744	0
TOTAL_VALUE	0	152372597	1.415075e+09	47386565054	0
TOTAL_WEIGHT	0	4386521	2.818801e+08	14334284513	0
COUNTRY	0	15	1.970887e+01	48	0
COMMODITY_CODE	0	317	2.942466e+02	523	0
NO_DAYS	1	2	4.393190e+01	365	0
EXPIRY_DAYS	-5394	792	8.596461e+02	21371	0
FORFEITURE_STATUS_CODE	0	0	2.017451e-01	1	0

Table 2 Dataset Structure

3- Data Distribution Frequency

- Destination Code

Figure 5 shows that most of the sample has “Other” destination, than “Shop shore ship” destination.

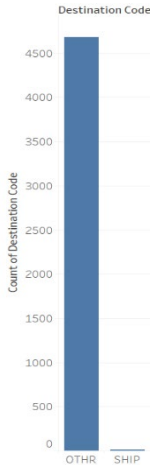

Figure 5 Destination Code, generated from custom dataset using Tableau.

- Calculated Amount

The histogram in Figure 6 shows the duty amounts are concentrated in the low value range.

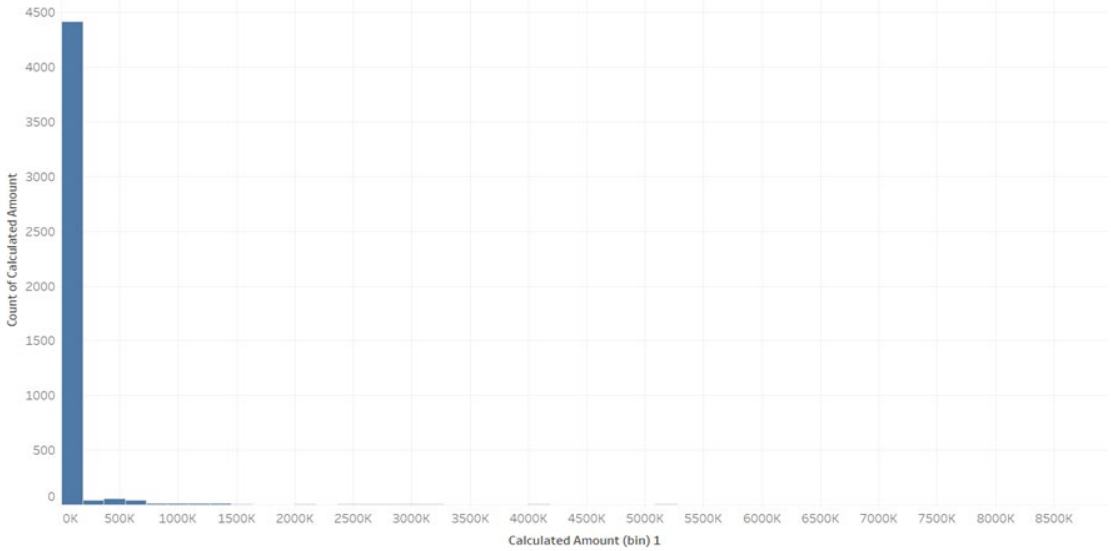

Figure 6 Calculated Amount, generated from custom dataset using Tableau.

- Payment Aging

The graphs in Figure 7 means that most the clients have few previous outstanding payments.

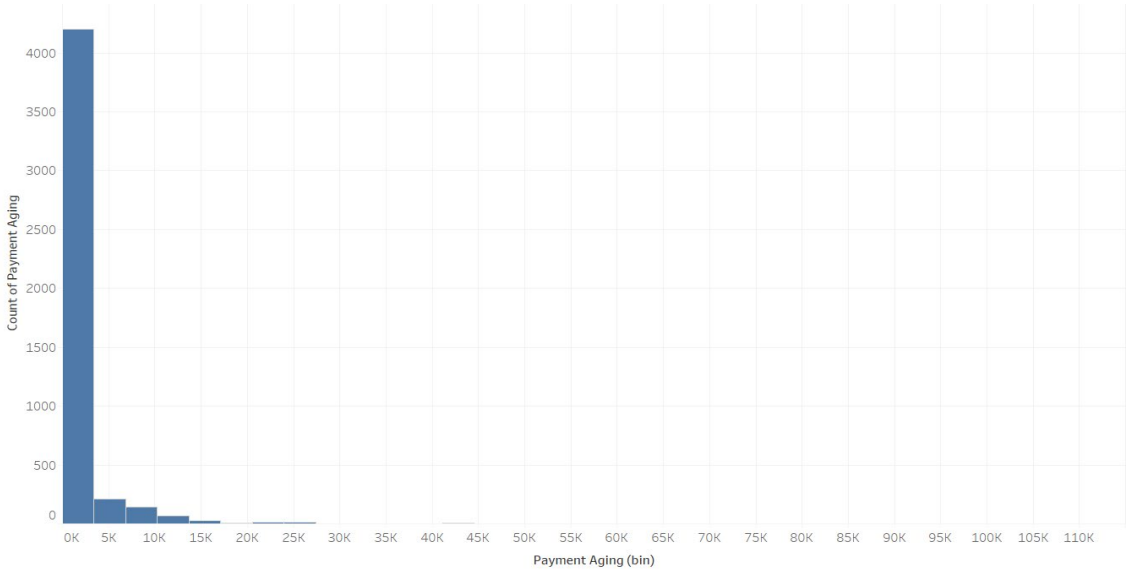

Figure 7 Payment Aging, generated from custom dataset using Tableau.

- GATEEPASS

Most of the companies have between zero and hundred gate passes in the week range of their export transaction as shown in Figure 8.

Figure 8 GATEEPASS, generated from custom dataset using Tableau.

- AEO_FLAG

Most of the companies are not part of Advanced Economic Operators as shown in Figure 9.

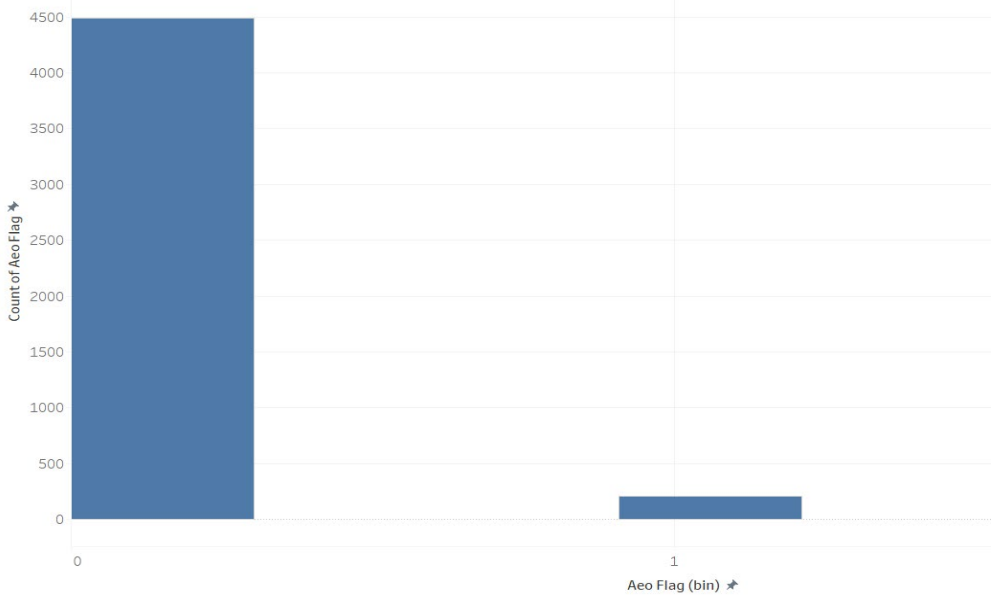

Figure 9 AEO_FLAG, generated from custom dataset using Tableau.

- LICENSE_ISSUING_AUTH_CODE

Figure 10 shows that most companies using the NR service belong to one freezone license authority.

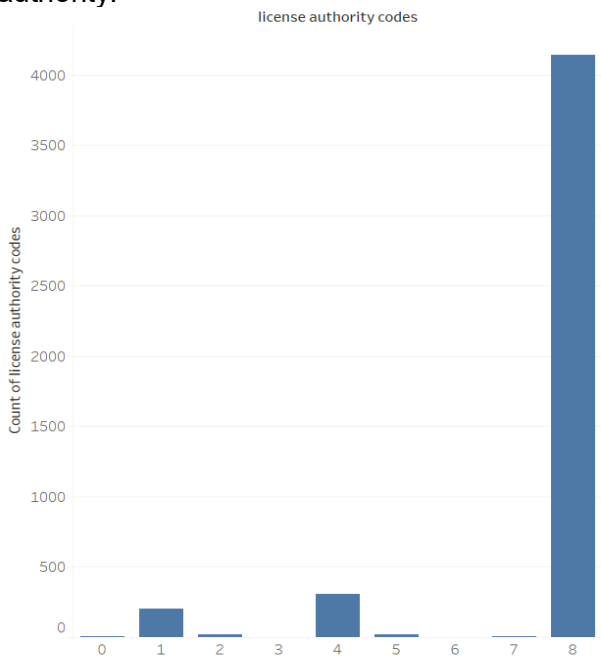

Figure 10 LICENSE_ISSUING_AUTH, generated from custom dataset using Tableau.

- TOTAL_INCOME

Figure 11 shows that low duties values are collected from the sample clients, mainly because the majority are freezone companies, some outliers contribute with high duty values.

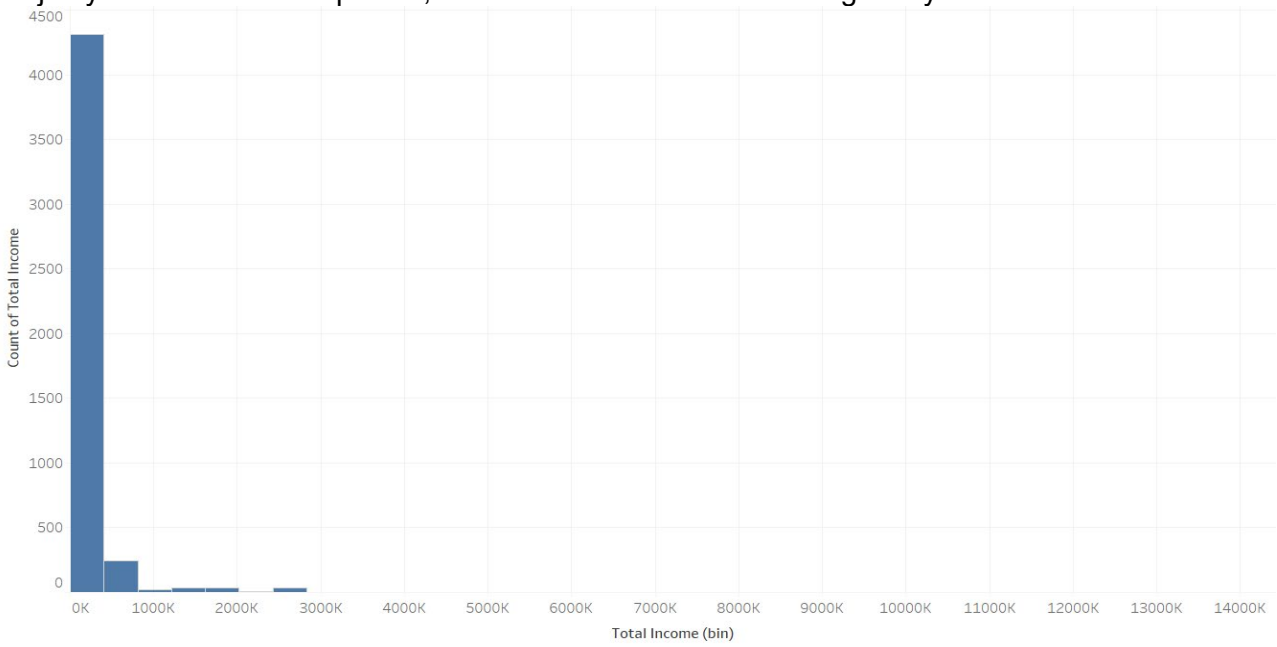

Figure 11 TOTAL_INCOME, generated from custom dataset using Tableau.

- FREQUENCY

Majority of companies have low transactions count during the last three years as depicted by Figure 12.

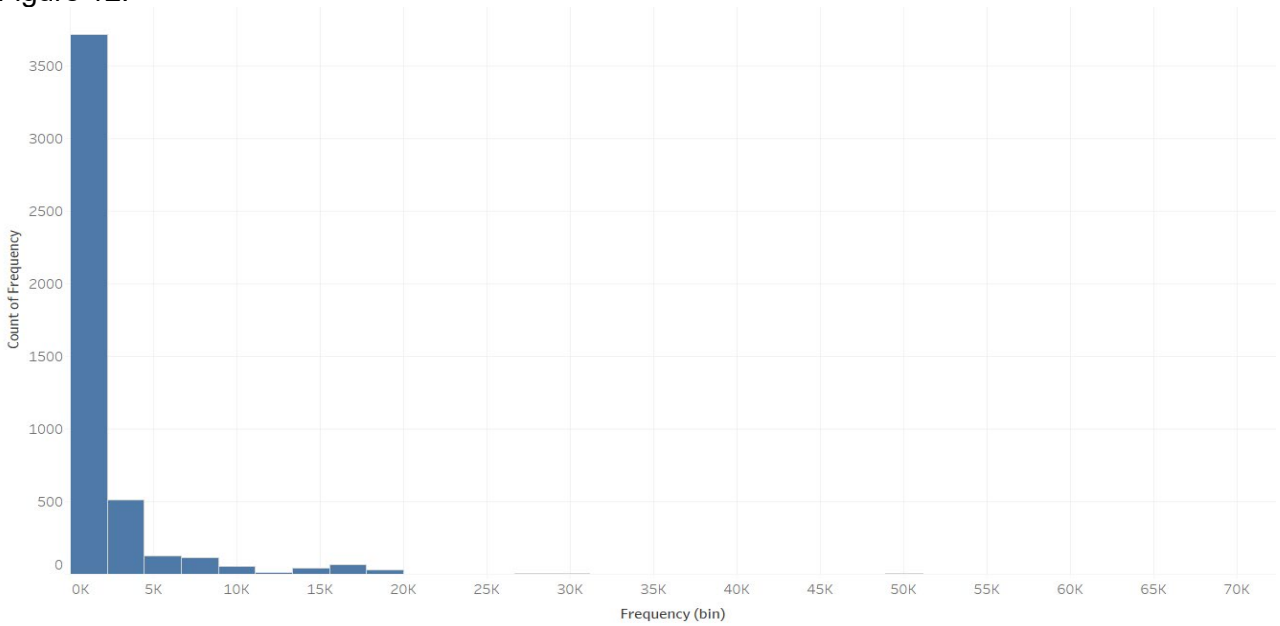

Figure 12 FREQUENCY, generated from custom dataset using Tableau.

- TOTAL_VALUE

Figure 13 shows that mostly the companies trade in low invoice values, with some outliers with high value goods.

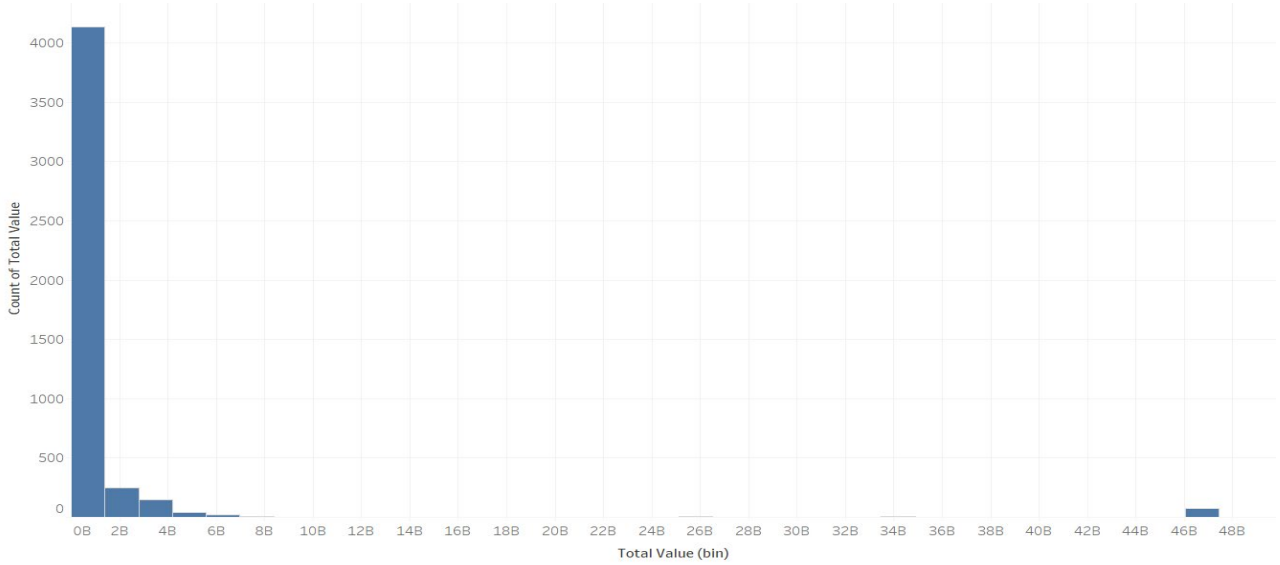

Figure 13 TOTAL_VALUE, generated from custom dataset using Tableau.

- TOTAL_WEIGHT

Figure 14 shows that the goods total weights are in the light weight range with some exceptions trading in the high weight shipments.

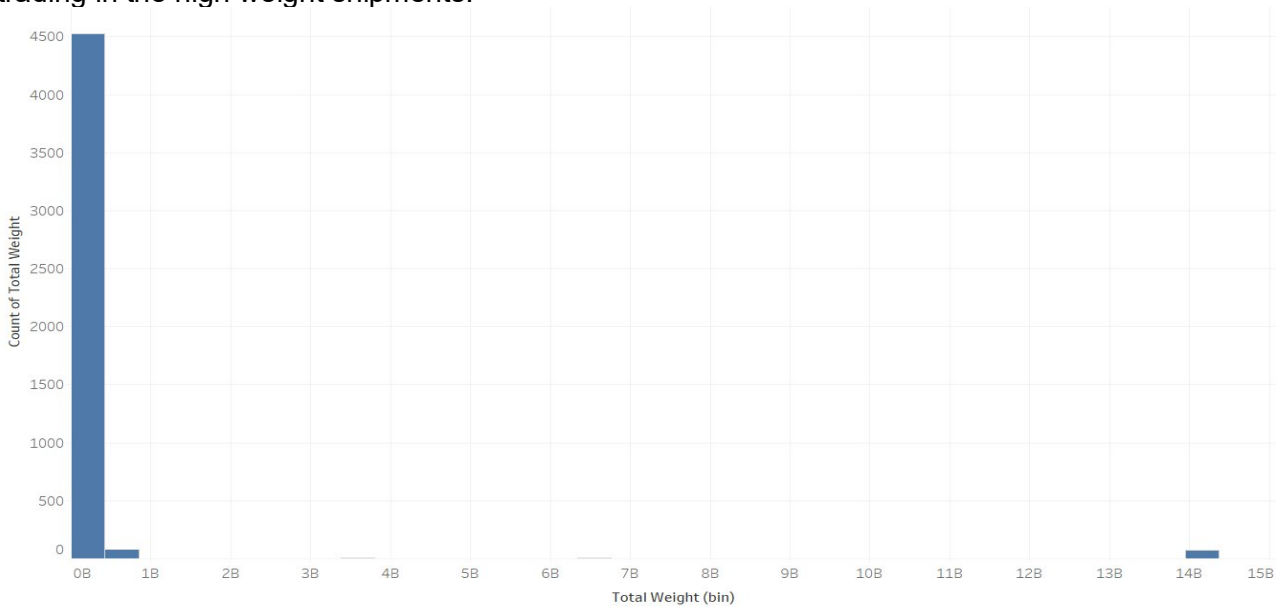

Figure 14 TOTAL_WEIGHT, generated from custom dataset using Tableau.

- COUNTRY

Figure 15 shows that trade partners are mainly China, US, Japan, and Korea.

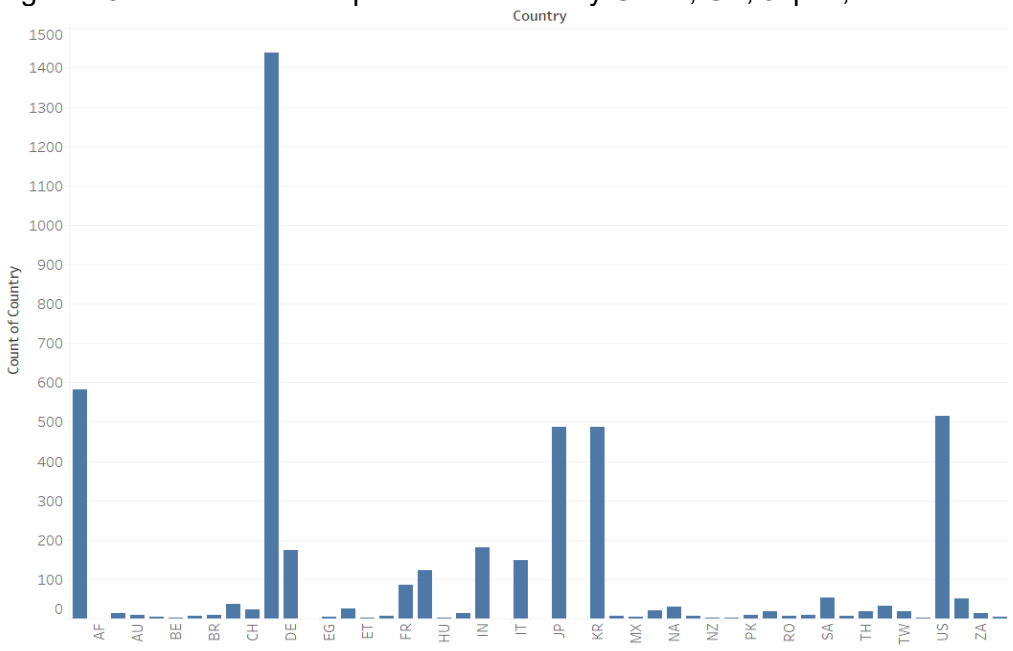

Figure 15 COUNTRY, generated from custom dataset using Tableau.

- COMMODITY_CODE

The top traded commodities are vehicles, parts, and motors as shown in Figure 16.

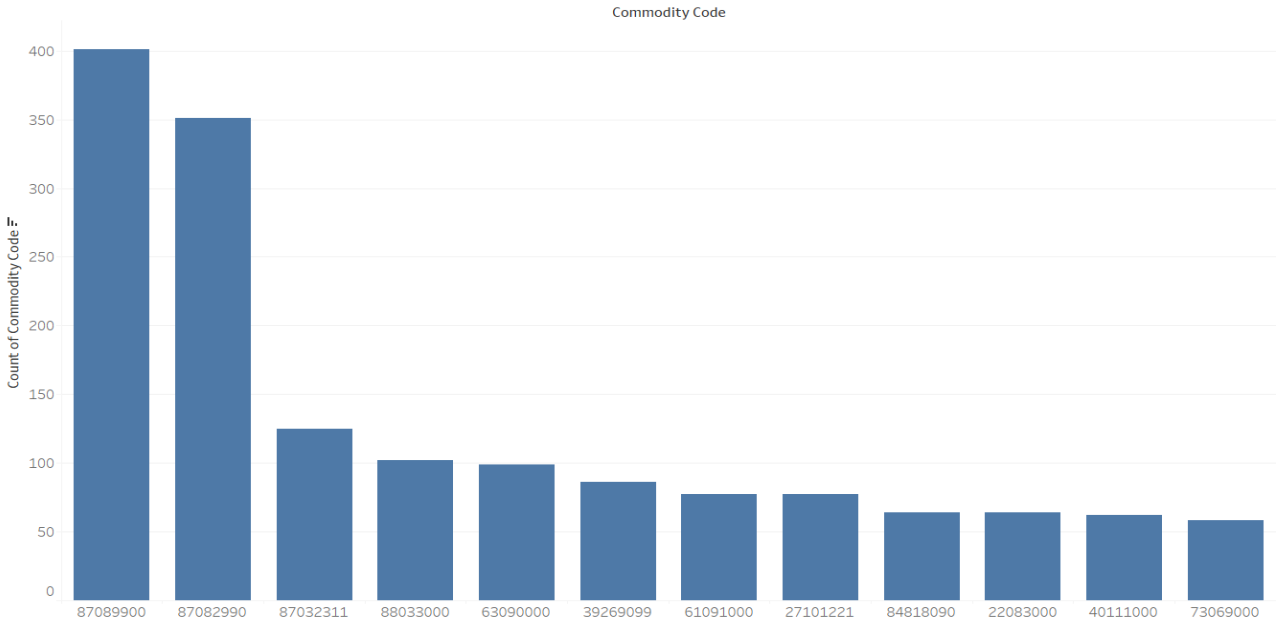

Figure 16 COMMODITY_CODE, generated from custom dataset using Tableau.

- NO_DAYS

A lot of companies are actively using the system within the last 20 days as depicted by Figure 17.

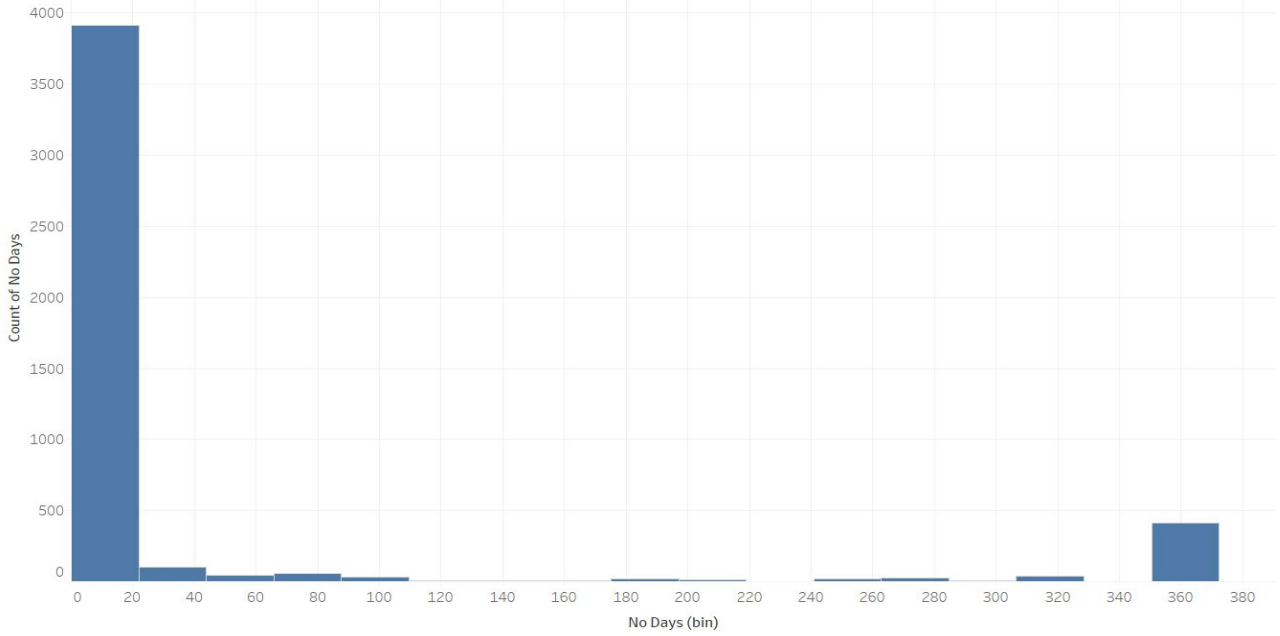

Figure 17 NO_DAYS, generated from custom dataset using Tableau.

- EXPIRY_DAYS

Figure 18 shows that many companies increase their transactions near expiry, and some of them still using the customs system even after the trade license expiry.

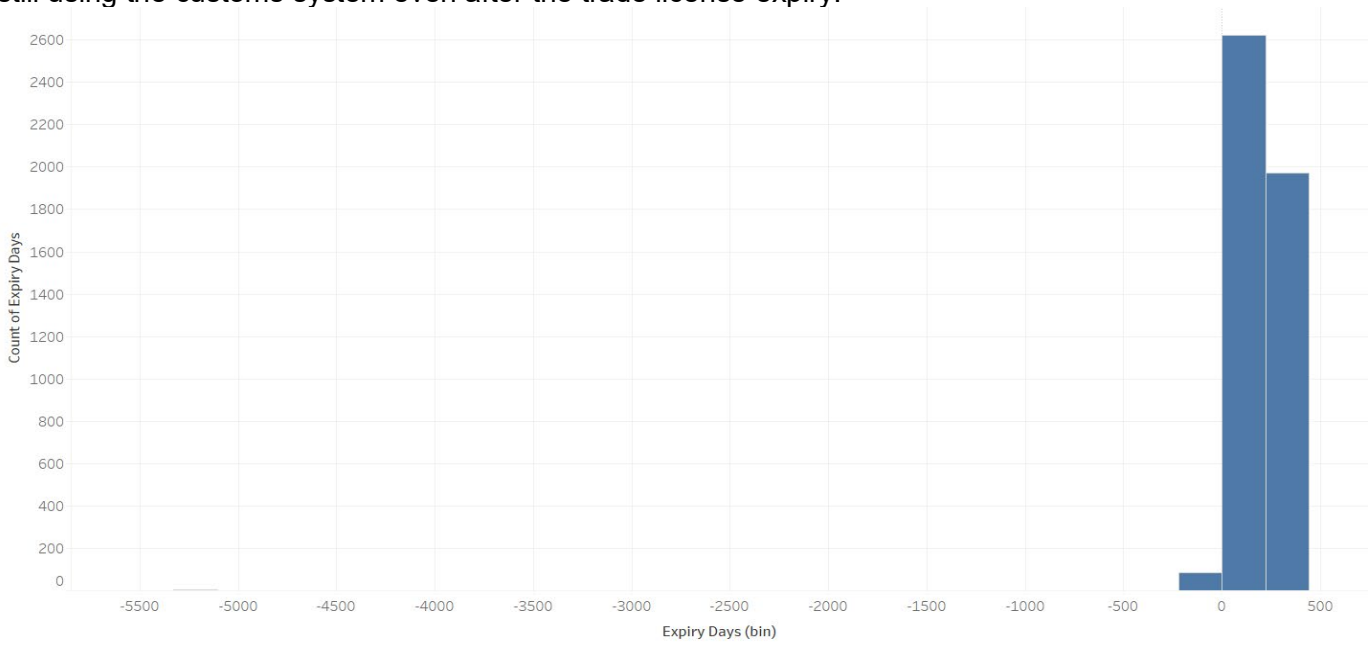

Figure 18 EXPIRY_DAYS, generated from custom dataset using Tableau.

- FORFEITURE_STATUS_CODE

There is a kind of class imbalance, as expected there are fewer defaulted transaction than the settled ones as shown in Figure 19.

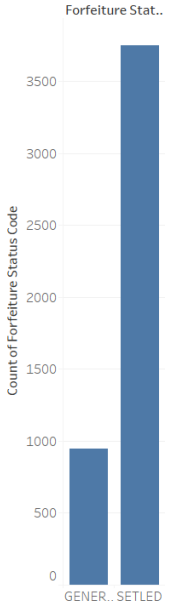

Figure 19 FORFEITURE_STATUS_CODE, generated from custom dataset using Tableau.

4- Data Correlations

The attributes of the dataset show a correlation between each other and correlation with the response attribute as shown in Figure 20.

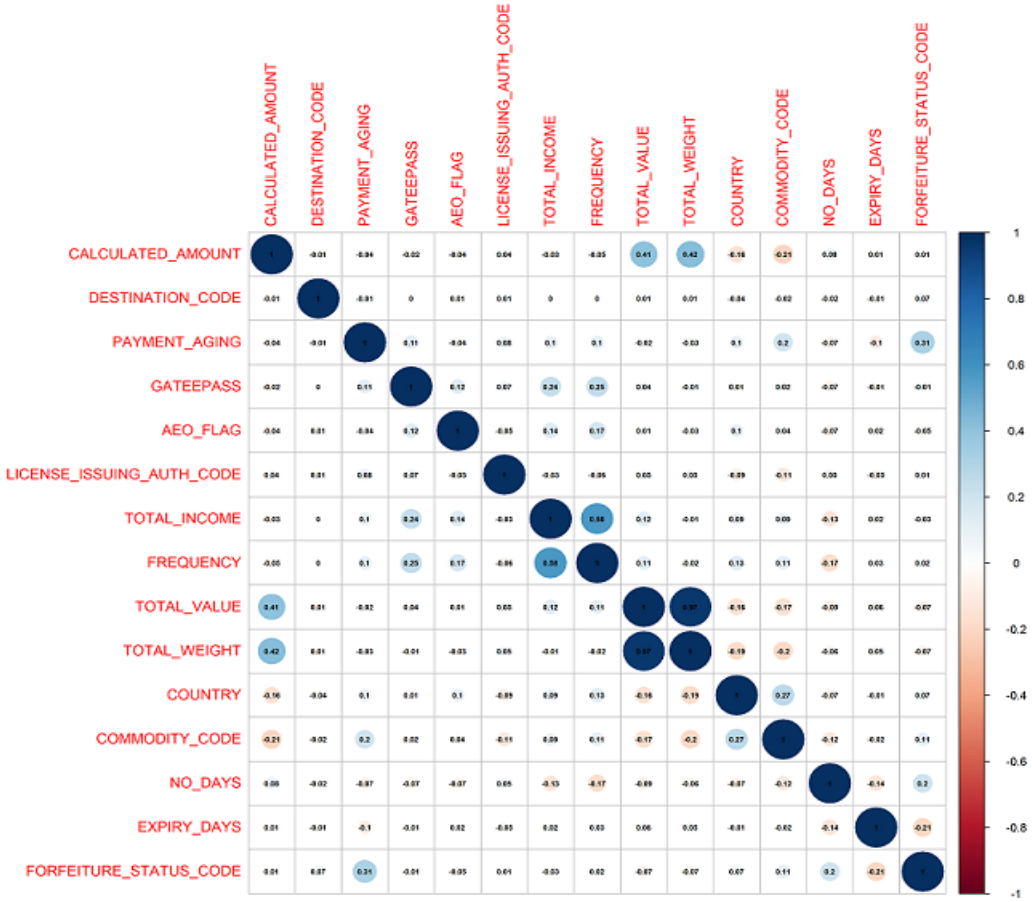

Figure 20 Correlation Diagram, generated from the customs dataset using R.

The graph highlights the following relations between the attributes and the response:

- The more the available aging payment for the clients the more the increase in the defaulting probability.
- The close the license expiry date the highly the defaulting chances.
- For the high value and weight goods shipments there are less chances for payment defaults.
- The clients in the Authorized Economic Operator program are less likely to miss on their payments.

A. Commodities to payment correlation

The diagram in Figure 21 shows some commodities tend to be related to the payment default status like the ones associated with vehicles, parts, and Whiskey.

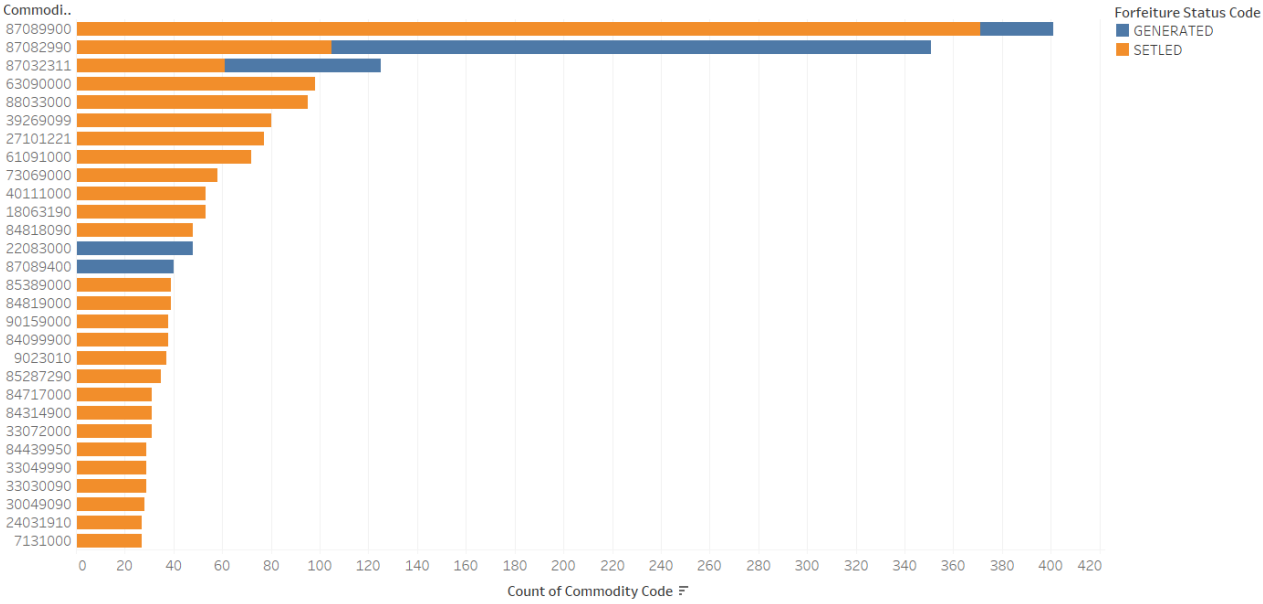

Figure 21 Commodities to payment correlation, generated from custom dataset using Tableau

B. Countries to payment status

Figure 22 contains a map shows that payment defaults most frequently taking place with the goods of far east origin.

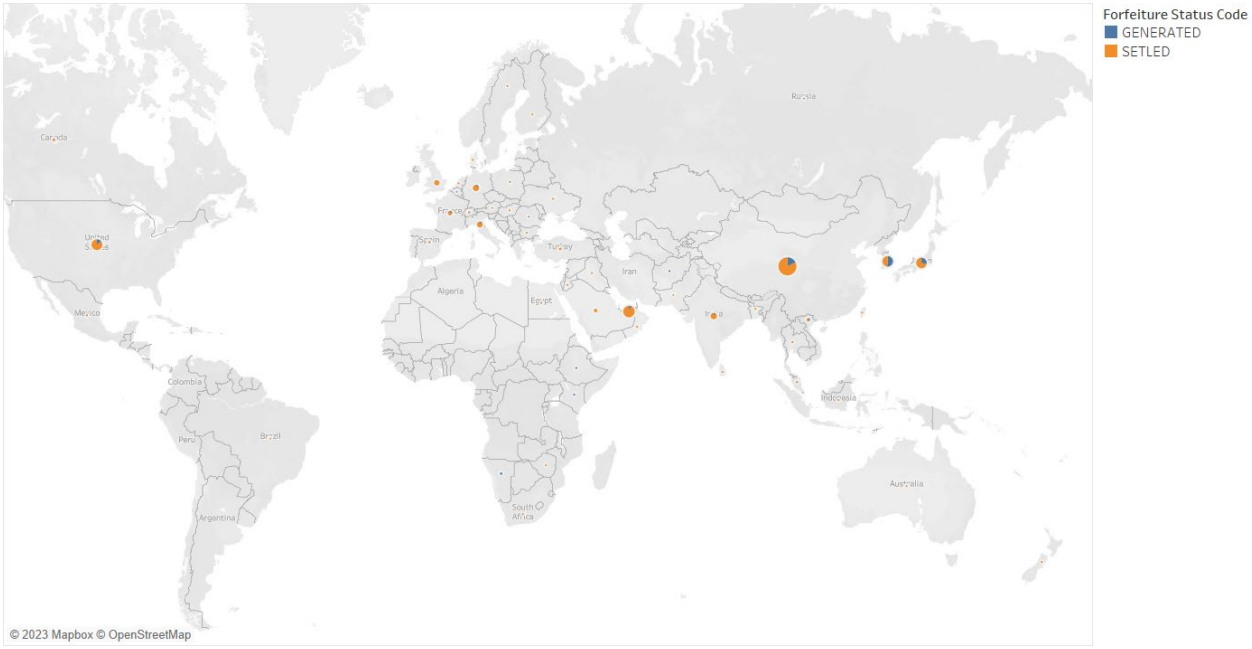

Figure 22 Countries to payment status, generated from custom dataset using Tableau.

C. License Issuing Authority to Payment

The graph in Figure 23 shows that some freezones are more connected to the defaulting status than others.

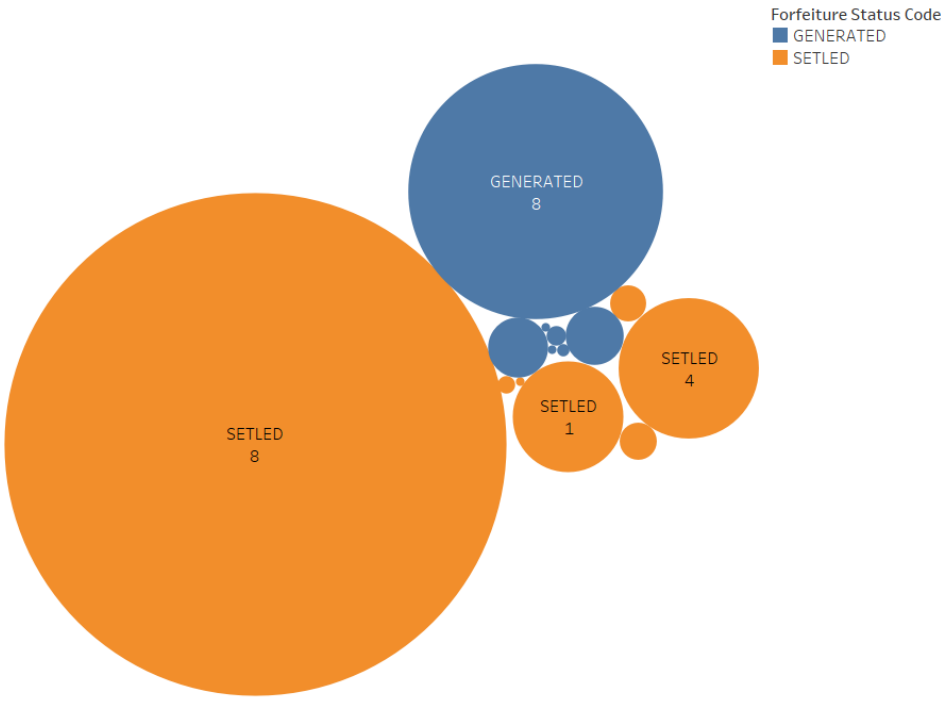

Figure 23 License Issuing Authority to Payment, generated from custom dataset using Tableau.

5- Feature Importance

The following diagram in Figure 24 shows the analysis results of feature importance using the R caret library.

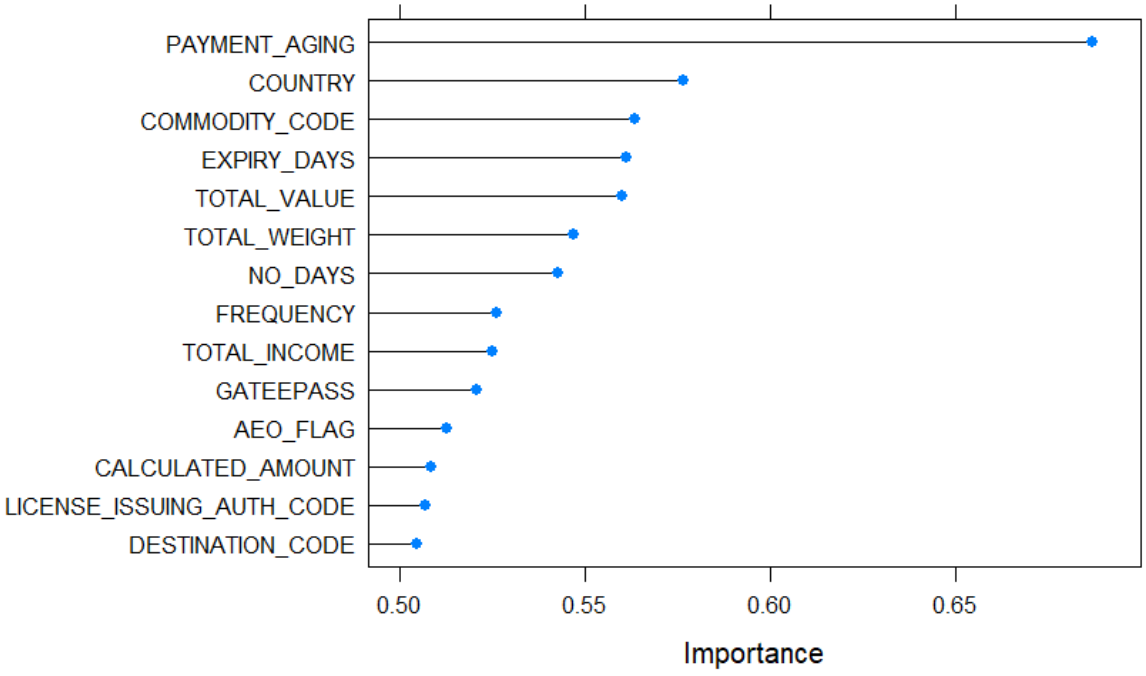

Figure 24: Features importance

The analysis reveals that the sum of payment aging contributes the most to the response attribute of payment default status, followed by the country of origin, type of commodity, trade license expiry remaining period, invoice total value, and total goods weight. However, we are going to try using all of the available dataset in the prediction models.

6- Data Scaling

To optimize the results of the statistical models, data scaling is required to bring all data attributes to the same range, the techniques used in data scaling are, Z-Score, Min-Max, and decimal scaling, the following code snippet shows how these methods were employed.

```
xy$CALCULATED_AMOUNT <- (xy$CALCULATED_AMOUNT-mean(xy$CALCULATED_AMOUNT))/sd(xy$CALCULATED_AMOUNT)
xy$TOTAL_INCOME <- (xy$TOTAL_INCOME-mean(xy$TOTAL_INCOME))/sd(xy$TOTAL_INCOME)
xy$FREQUENCY <- (xy$FREQUENCY-mean(xy$FREQUENCY))/sd(xy$FREQUENCY)
xy$TOTAL_VALUE <- (xy$TOTAL_VALUE-mean(xy$TOTAL_VALUE))/sd(xy$TOTAL_VALUE)
xy$TOTAL_WEIGHT <- (xy$TOTAL_WEIGHT-mean(xy$TOTAL_WEIGHT))/sd(xy$TOTAL_WEIGHT)
xy$NO_DAYS <- (xy$NO_DAYS-mean(xy$NO_DAYS))/sd(xy$TOTAL_WEIGHT)
xy$EXPIRY_DAYS <- (xy$EXPIRY_DAYS-mean(xy$EXPIRY_DAYS))/sd(xy$EXPIRY_DAYS)
xy$COMMODITY_CODE <- xy$COMMODITY_CODE/1000
xy$COUNTRY <- xy$COUNTRY/100
```

7- Hypothesis Testing

Here we are going to test the main research questions using logistic regression model to confirm or reject the hypothesis of whether the related attributes affect the payment default status or not.

- Whether the defaulting companies are avoiding the closing audit, the related feature is the remaining period to the license expiry date at the transaction date.
- Whether there is possible smuggling to local market activity, the related feature is the "Gatepass" flag which indicates the coexistence of a gate-pass to local market at the same time of the export declaration.

Examining the logistic regression results,

```
Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)               -5.619e+00  8.074e-01  -6.959 3.41e-12 ***
CALCULATED_AMOUNT          1.400e-07  9.650e-08   1.451   0.1469
DESTINATION_CODE           3.532e+00  7.478e-01   4.723 2.32e-06 ***
PAYMENT_AGING              3.131e-04  1.916e-05  16.337  < 2e-16 ***
GATEEPASS                 -5.472e-04  7.919e-04  -0.691   0.4896
AEO_FLAG                  -8.206e-02  2.825e-01  -0.290   0.7715
LICENSE_ISSUING_AUTH_CODE -2.183e-04  2.848e-02  -0.008   0.9939
TOTAL_INCOME              -1.940e-06  3.832e-07  -5.061 4.16e-07 ***
FREQUENCY                  8.931e-05  2.142e-05   4.170 3.04e-05 ***
TOTAL_VALUE               -1.706e-11  6.985e-11  -0.244   0.8070
TOTAL_WEIGHT              -1.959e-10  2.697e-10  -0.727   0.4675
COUNTRY                    9.013e-03  3.832e-03   2.352   0.0187 *
COMMODITY_CODE             8.055e-04  3.274e-04   2.461   0.0139 *
NO_DAYS                    4.616e-03  3.744e-04  12.332  < 2e-16 ***
EXPIRY_DAYS               -1.878e-03  4.370e-04  -4.298 1.73e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For the first hypothesis that companies are trying to avoid stock audit shortages, the remaining period to expiry represented by EXPIRY_DAYS attribute is significant with P-Value less than alpha, therefore this hypothesis cannot be rejected.

The second hypothesis regarding the possibility of smuggling to local market during the time of export declaration, and represented by the GATEPASS attribute; the related P-Value (0.49) is larger than the alpha value, which indicates that the attribute is not significant and the null hypothesis can be accepted.

8- Data Split

When one has to deal with classification tasks for which the prior probabilities of class membership are very different, it becomes important to choose training and test sets that faithfully represent the original label distribution. This is where stratification of the subsampling becomes essential. The study is using a function performs stratified hold out subsampling to split the training and test datasets.

```{R}

  stratified.holdout <- function(y, ptr)
  {
     n             <- length(y)
     labels        <- unique(y)      # Obtain classifiers
     id.tr <- id.te <- NULL

     # Loop once for each unique label value

     y <- sample(sample(sample(y)))

     for(j in 1:length(labels))
     {
       sj    <- which(y==labels[j])  # Grab all rows of label type j
       nj    <- length(sj)           # Count of label j rows to calc proportion below

       id.tr <- c(id.tr, (sample(sample(sample(sj))))[1:round(nj*ptr)])
     }                           # Concatenates each label type together 1 by 1

     id.te  <- (1:n) [-id.tr]         # Obtain and Shuffle test indices to randomize

  return(list(idx1=id.tr,idx2=id.te))
}
```

9- Models Evaluation

This study tests a variety of statistical and machine learning models for prediction of the payment default status, here are the list of tested models with its hyper parameters,

- LDA (Linear Discriminant Analysis)
- QDA (Quadratic Discriminant Analysis)
- Logistic Regression, with parameters family=binomial(link='logit')
- Nearest Neighbors Learning Machine (KNN) with k=12
- Support Vector Machines, with kernel='rbfdot', and type='C-svc'
- Random Forest with 100 estimators.
- Stochastic Adaptive Boosting with 100 estimators.

The following diagram in Figure 25 compares the different model's accuracy using the ROC curve,

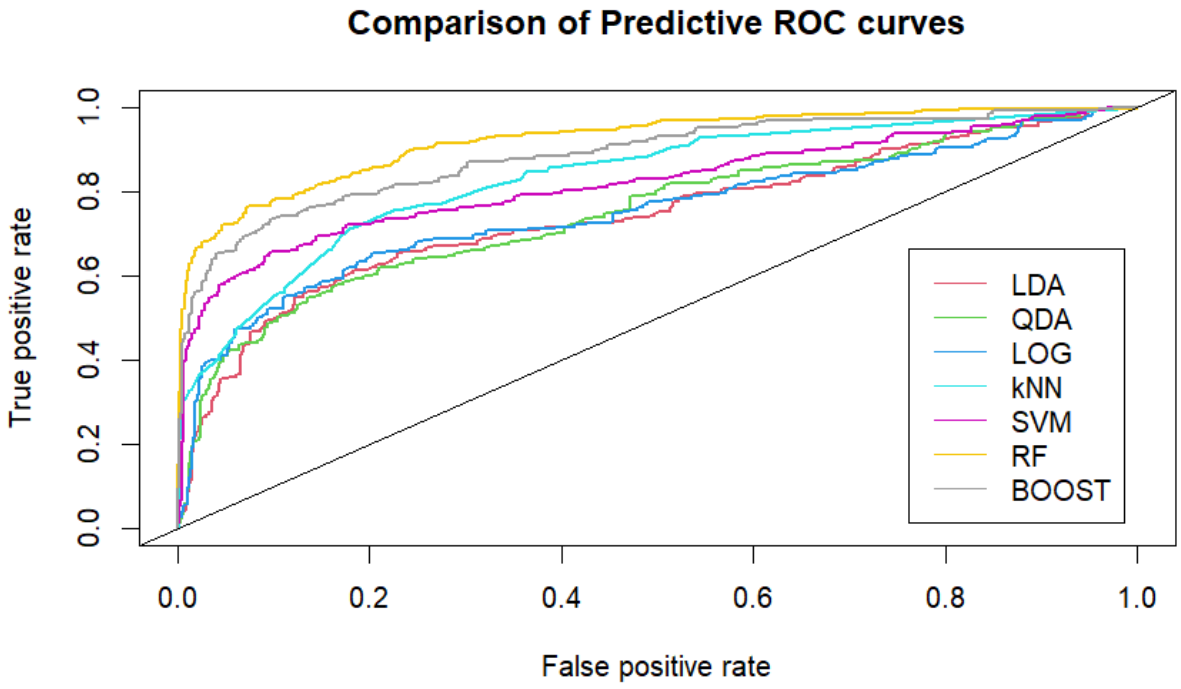

Figure 25 model comparison

According to the graph above the top performing models are Random Forest (RF) and Gradient Boost followed by the KNN model.

Calculating the RF model confusion matrix and accuracy results in,

- Accuracy: 0.9166
- Precision: 0.9727
- Recall (TPR): 0.9260
- Fi Score: 0.9488229

10- Class Imbalance

The collected sample has quite class imbalance between the settled and defaulted transactions. The parentage of defaulted transactions is around 20%. To rectify this imbalance SMOTE (Synthetic Minority Over-sampling Technique) is used, which is a popular technique used in machine learning to address class imbalance by generating synthetic samples for the minority class.

After using SMOT the class distribution is now 56% settled to 42% defaulted transactions. Running the same models again of the SMOT data resulted in the following diagram in Figure 26, which shows similar comparison to the models in the previous section without applying class balancing technique.

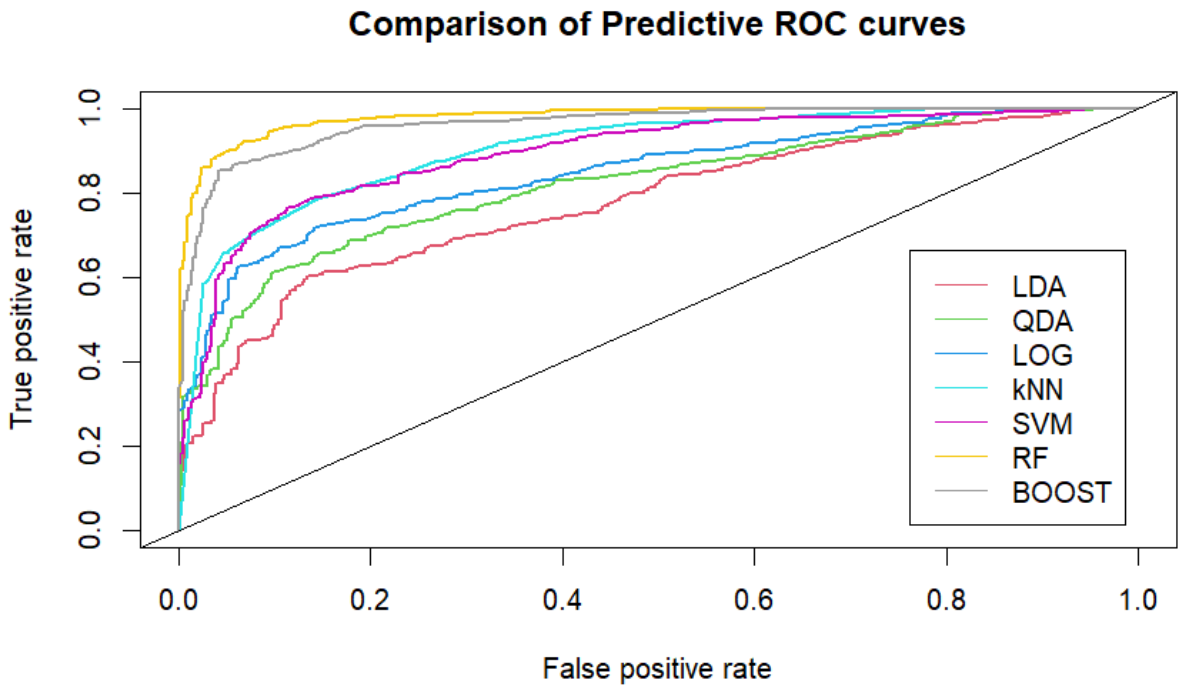

Figure 26 Models Run on SMOT results

The resulted confusion matrix of the Random Forest model is showing the following measures,

	Class Imbalance	**With SMOT**
Accuracy	91.82%	92.76%
Precision	92.72%	91.70%
Recall	97.58%	89.49%
Fi Score	95.09%	90.58%

Table 3 RF Accuracy comparison

The results in Table 3 show that implementing the model without balancing the classes by synthetic data gave a better result.

11- Cross Validation

The study uses the bootstrap resampling method to test the model for overfitting and check the consistency of the prediction across different samples. The results are shown in Figure 26 as a Boxplot diagram for comparison.

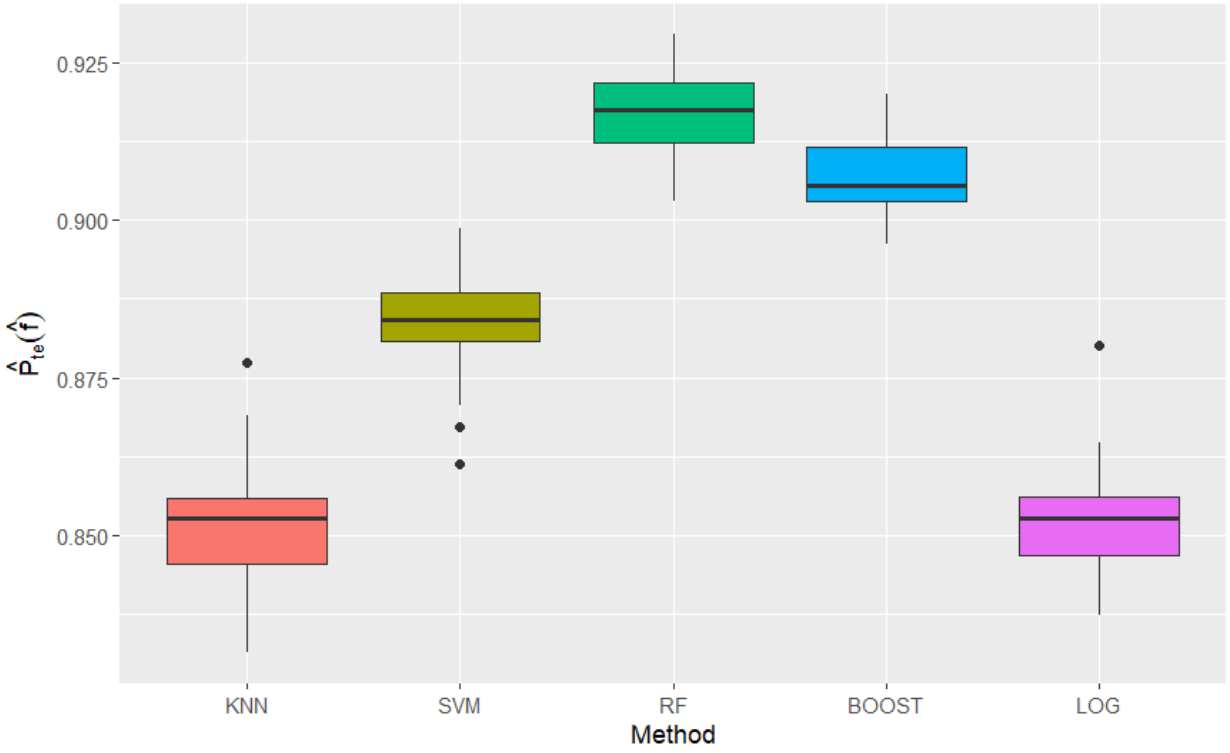

Figure 27 Cross Validation

The diagram shows the Random Forest model is consistently performing better than other model with average accuracy of 0.9175149. Followed by Boosting model with close difference.

The worst performing model is Logistic regression which indicate that the data relationship with the logit is not following a linear distribution.

Chapter 5 Discussion

The study addressed the answers to the main business questions and tested the hypothesis of the related relations. The first question was whether the payment defaults were linked to end of license audit evasion, the logistic regression results couldn't reject this hypothesis. Which is also evident in the following Figure 27 diagram, which shows that most of the defaulting happen near license expiry or during the grace period provided after expiry.

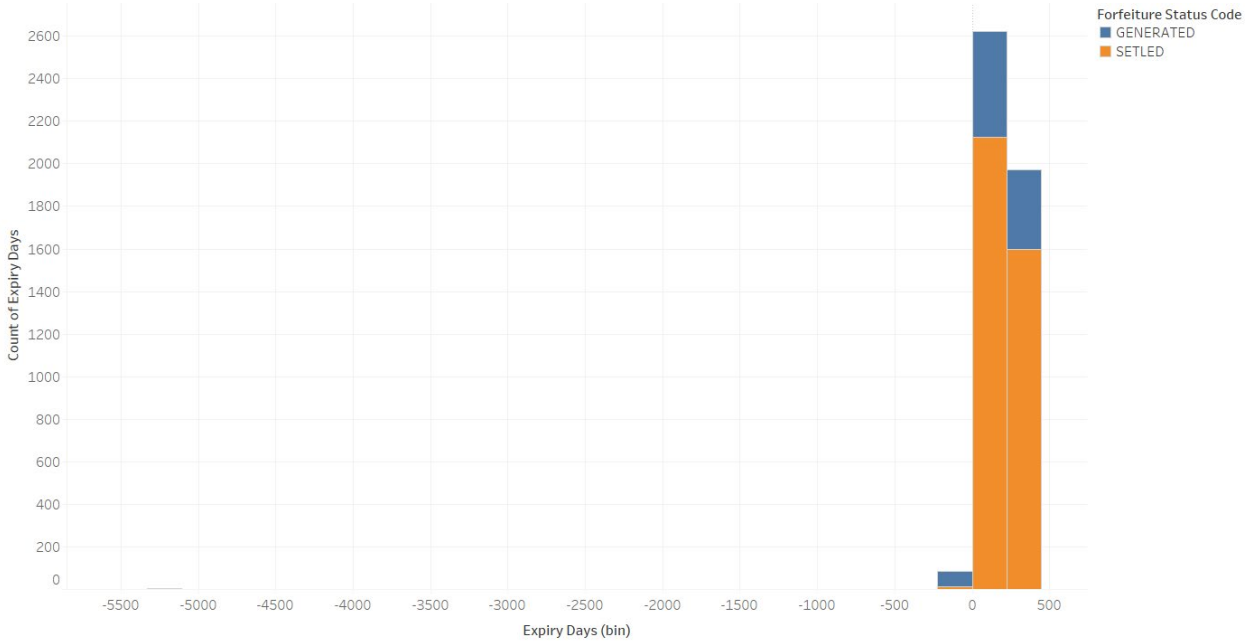

Figure 28 Audit Evasion relation, generated from custom dataset using Tableau.

The customs audit department needs to put new business rules and procedure to consider this declared as export goods in the audit scope. And educate the clients that this behaviour will not relief them from audit banalities and fines in case the mentioned goods are not available in the company's warehouse.

The second question was whether the payment defaults were because of a smuggling operation taking place at the same time of fake exports, the logistic regression p-value rejected that hypothesis which is also evident in the following diagram.

Figure 28 shows number of local gate-passes that took place during the declaration time does not increase the defaulting (Generated) status.

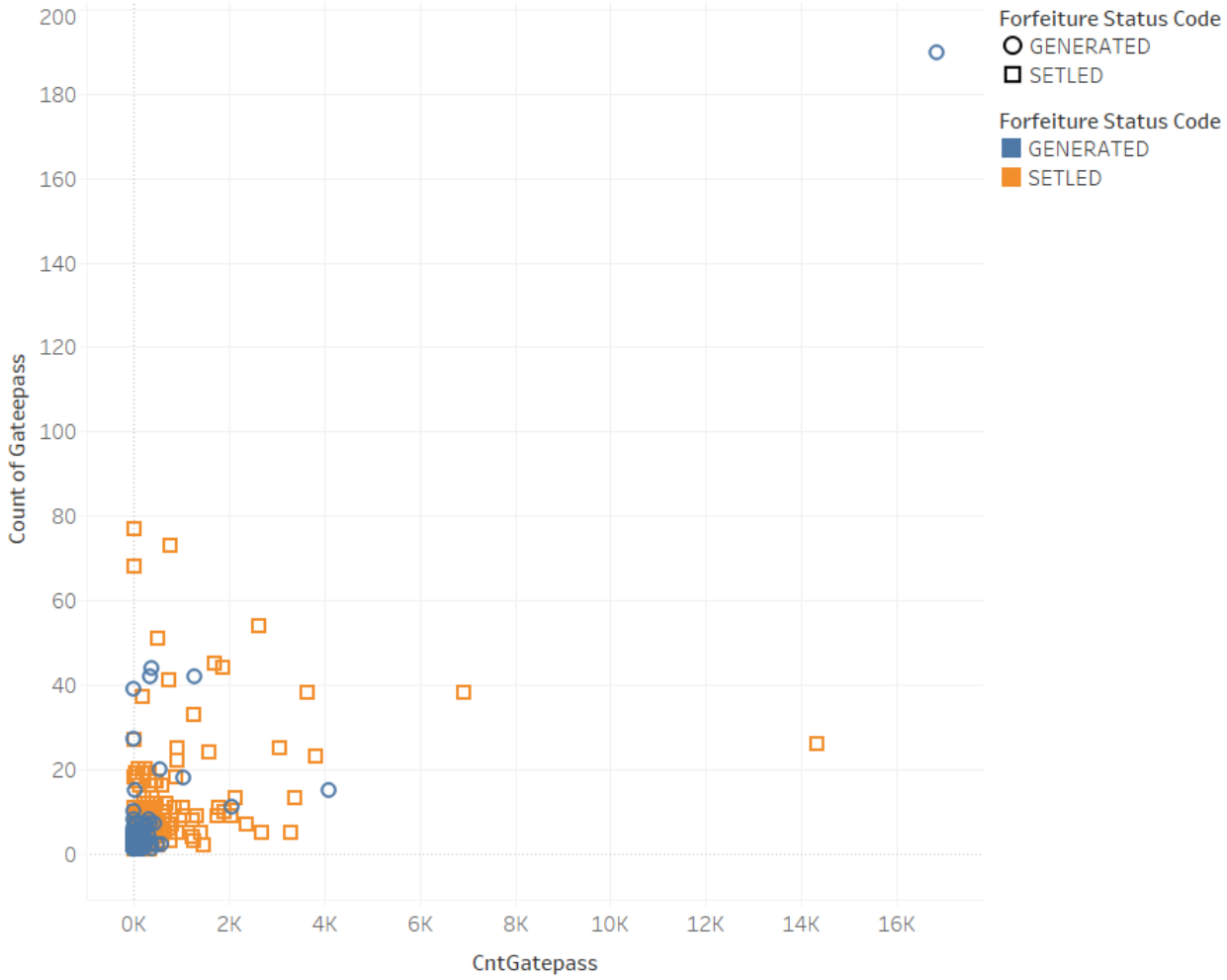

Figure 29 Defaulting status to gate-pass count, generated from custom dataset using Tableau.

1- Prediction Discussion

The last question and the core one where it's possible to predict the defaulting status using different machine learning algorithms. It was observed that there are distinguishable patterns in the data which is detectible by different machine learning techniques, the Random Forest which is an ensemble of various decision trees contributing together for the final decision, emerged as the top performing model compared to the others. With accuracy around 91% the model demonstrated that it could predict the companies with potential defaulting compared to those who were compliant. The customs could use such a model to stop the revenue leakage due to the duty evasion through the use of fraudulent export declarations.

The study is aligned with the literature reviews in the following points,

- The successful use of machine learning in customs authorities has enabled efficient prediction of delinquent payments and frauds, leading to increased compliance and reduced financial losses.
- One of the highly important features is the existence of a previous defaults for the clients, which is repeatedly used in many of the referenced researches. The study shows it's on the top in importance among the other features.
- The machine learning models can predict payment defaults both before and after transactions occur. A priori models can help prevent fraud, while a posteriori models can aid in audit planning. This study is one of the A Priori mechanisms to prevent frauds.
- Tree-based models such as Random Forest, Boosting, and Bagging have been found to be particularly effective in predicting duty and payment fraud cases, which is evident in this study having Random Forest and Boosting as the top performing models.

Chapter 6 Conclusions

1- Conclusion

The analysis indicates a strong likelihood that the fraudulent freezone exports is connected to the remaining period till the trade license expiry. However, the statistical models applied to the provided sample data did not confirm the other hypothesis regarding smuggling. It's important to note that the statistical models can provide insights about the relations in the data, however, it doesn't always confirm or deny hypothesis definitely as it works on a sample data. Although the data in this case did not support the hypothesis of smuggling, customs should conduct further investigations. However, the connection between the fraudulent freezone exports and the remaining license expiry period still appears to be a significant factor as suggested by the analysis.

By leveraging the insights provided by the Random Forest model, customs agencies can proactively target companies that are more likely to engage in fraudulent transactions, thereby reducing revenue losses associated with duty evasion. Implementing such a predictive model as part of the decision-making process can enable customs' authorities to allocate resources effectively and focus their efforts on high-risk entities, contributing to enhanced compliance and revenue protection.

Overall, the study highlights the utility of machine learning algorithms, particularly the Random Forest model, in identifying potential defaulting companies and in preventing revenue losses resulting from fraudulent practices.

2- Recommendation

It is recommended that the customs audit team revise its business rules and policies in order to ensure that fraudulent exports are not considered in inventory audits. It is important that the customs department educates its clients about the possible penalties and fines that may occur as a result of such fraudulent transactions, informing them that such transactions won't affect the audit results.

Employing machine learning models, specifically Random Forest to predict and avert the fraudulent transactions and prevent the resulting revenue loss is strongly recommended. To ensure optimal performance, it is advisable to integrate the predictive model within the MLOps (Machine Learning Operations) environment. Regular retraining of the model should be conducted at suitable intervals to enhance its accuracy and effectiveness. Additionally, continuous monitoring of the model for potential concept drift is crucial to maintain its reliability over time.

3- Future Work

Among the limitations of the study was the absence of financial data, such as importer companies' income statements, in the customs department. Some of the reviewed literatures particularly in Tax evasion domain found that the companies' financial data is useful. As a future work it's recommended to collect clients' financial statements related datasets.

The study did not cover neural networks and deep learning techniques. In spite of the acceptable accuracy of the machine learning algorithm, it is always a good idea to evaluate deep learning methods with this topic for future research.

References

1. Abedin, M.Z. et al. (2021) 'Tax Default Prediction Using Feature Transformation-Based Machine Learning', IEEE Access, 9, pp. 19864–19881. Available at: https://doi.org/10.1109/ACCESS.2020.3048018.
2. Bonchi, F. et al. (1999) 'Using Data Mining Techniques in Fiscal Fraud Detection', in M. Mohania and A.M. Tjoa (eds) DataWarehousing and Knowledge Discovery. Berlin, Heidelberg: Springer Berlin Heidelberg (Lecture Notes in Computer Science), pp. 369–376. Available at: https://doi.org/10.1007/3-540-48298-9_39.
3. Fan Yu, Zheng Qin, and Xiao-Ling Jia (2003) 'Data mining application issues in fraudulent tax declaration detection', in Proceedings of the 2003 International Conference on Machine Learning and Cybernetics (IEEE Cat. No.03EX693). 2003 International Conference on Machine Learning and Cybernetics, Xi'an, China: IEEE, pp. 2202–2206. Available at: https://doi.org/10.1109/ICMLC.2003.1259872.
4. González García, I. and Mateos Caballero, A. (2021) 'A Multi-Objective Bayesian Approach with Dynamic Optimization (MOBADO). A Hybrid of Decision Theory and Machine Learning Applied to Customs Fraud Control in Spain', Mathematics, 9(13), p. 1529. Available at: https://doi.org/10.3390/math9131529.
5. Granados, O.M. and Nicolás-Carlock, J.R. (eds) (2021) Corruption Networks: Concepts and Applications. Cham: Springer International Publishing (Understanding Complex Systems). Available at: https://doi.org/10.1007/978-3-030-81484-7.
6. Höglund, H. (2017) 'Tax payment default prediction using genetic algorithm-based variable selection', Expert Systems with Applications, 88, pp. 368–375. Available at: https://doi.org/10.1016/j.eswa.2017.07.027.
7. Hua Shao, Hong Zhao, and Gui-Ran Chang (2002) 'Applying data mining to detect fraud behavior in customs declaration', in Proceedings. International Conference on Machine Learning and Cybernetics. 2002 International Conference on Machine Learning and Cybernetics, Beijing, China: IEEE, pp. 1241–1244. Available at: https://doi.org/10.1109/ICMLC.2002.1167400.
8. Mojahedi, H., Babazadeh Sangar, A. and Masdari, M. (2022) 'Towards Tax Evasion Detection Using Improved Particle Swarm Optimization Algorithm', Mathematical Problems in Engineering. Edited by K. Sun, 2022, pp. 1–17. Available at: https://doi.org/10.1155/2022/1027518.
9. Seify, M. et al. (2022) 'Fraud Detection in Supply Chain with Machine Learning', IFAC-PapersOnLine, 55(10), pp. 406–411. Available at: https://doi.org/10.1016/j.ifacol.2022.09.427.
10. Triepels, R., Daniels, H. and Feelders, A. (2018) 'Data-driven fraud detection in international shipping', Expert Systems with Applications, 99, pp. 193–202. Available at: https://doi.org/10.1016/j.eswa.2018.01.007.
11. Vanhoeyveld, J., Martens, D. and Peeters, B. (2020) 'Customs fraud detection: Assessing the value of behavioural and high-cardinality data under the imbalanced learning issue', Pattern Analysis and Applications, 23(3), pp. 1457–1477. Available at: https://doi.org/10.1007/s10044-019-00852-w.
12. Wu, R.-S. et al. (2012) 'Using data mining technique to enhance tax evasion detection performance', Expert Systems with Applications, 39(10), pp. 8769–8777. Available at: https://doi.org/10.1016/j.eswa.2012.01.204.
13. Wu, Y. et al. (2019) 'TEDM-PU: A Tax Evasion Detection Method Based on Positive and Unlabeled Learning', in 2019 IEEE International Conference on Big Data (Big Data). 2019

IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA: IEEE, pp. 1681–1686. Available at: https://doi.org/10.1109/BigData47090.2019.9006325.
14. Zhu, X. et al. (2018) 'IRTED-TL: An Inter-Region Tax Evasion Detection Method Based on Transfer Learning', in 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE). 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE), New York, NY, USA: IEEE, pp. 1224–1235. Available at: https://doi.org/10.1109/TrustCom/BigDataSE.2018.00169.
15. Zumaya, M. et al. (2021) 'Identifying Tax Evasion in Mexico with Tools from Network Science and Machine Learning', in O.M. Granados and J.R. Nicolás-Carlock (eds) Corruption Networks. Cham: Springer International Publishing (Understanding Complex Systems), pp. 89–113. Available at: https://doi.org/10.1007/978-3-030-81484-7_6.
16. Alsadhan, N. (2023) 'A Multi-Module Machine Learning Approach to Detect Tax Fraud', Computer Systems Science and Engineering, 46(1), pp. 241–253. Available at: https://doi.org/10.32604/csse.2023.033375.
17. Murorunkwere, B.F. et al. (2023) 'Predicting tax fraud using supervised machine learning approach', African Journal of Science, Technology, Innovation and Development, pp. 1–12. Available at: https://doi.org/10.1080/20421338.2023.2187930.
18. Vanhoeyveld, J., Martens, D. and Peeters, B. (2020) 'Value-added tax fraud detection with scalable anomaly detection techniques', Applied Soft Computing, 86, p. 105895. Available at: https://doi.org/10.1016/j.asoc.2019.105895.
19. Murorunkwere, B.F. *et al.* (2022) 'Fraud Detection Using Neural Networks: A Case Study of Income Tax', *Future Internet*, 14(6), p. 168. Available at: https://doi.org/10.3390/fi14060168.
20. Baghdasaryan, V. et al. (2022) 'Improving Tax Audit Efficiency Using Machine Learning: The Role of Taxpayer's Network Data in Fraud Detection', Applied Artificial Intelligence, 36(1), p. 2012002. Available at: https://doi.org/10.1080/08839514.2021.2012002.
21. Hooda, N., Bawa, S. and Rana, P.S. (2020) 'Optimizing Fraudulent Firm Prediction Using Ensemble Machine Learning: A Case Study of an External Audit', Applied Artificial Intelligence, 34(1), pp. 20–30. Available at: https://doi.org/10.1080/08839514.2019.1680182.
22. Ben Ismail, M.M. and AlSadhan, N. (2023) 'Simultaneous Classification and Regression for Zakat Under-Reporting Detection', Applied Sciences, 13(9), p. 5244. Available at: https://doi.org/10.3390/app13095244.
23. Castellón González, P. and Velásquez, J.D. (2013) 'Characterization and detection of taxpayers with false invoices using data mining techniques', Expert Systems with Applications, 40(5), pp. 1427–1436. Available at: https://doi.org/10.1016/j.eswa.2012.08.051.
24. Ngah, Z.A., Ismail, N. and Abd Hamid, N. (2022) 'A cohesive model of predicting tax evasion from the perspective of fraudulent financial reporting amongst small and medium sized enterprises', Accounting Research Journal, 35(3), pp. 349–363. Available at: https://doi.org/10.1108/ARJ-09-2020-0315.
25. González-Martel, C., Hernández, J.M. and Manrique-de-Lara-Peñate, C. (2021) 'Identifying business misreporting in VAT using network analysis', Decision Support Systems, 141, p. 113464. Available at: https://doi.org/10.1016/j.dss.2020.113464.
26. Alejandrino, J.C., P. Bolacoy, J.Jr. and Murcia, J.V.B. (2023) 'Supervised and unsupervised data mining approaches in loan default prediction', *International Journal of Electrical and Computer Engineering (IJECE)*, 13(2), p. 1837. Available at: https://doi.org/10.11591/ijece.v13i2.pp1837-1847.

27. Madaan, M. et al. (2021) 'Loan default prediction using decision trees and random forest: A comparative study', IOP Conference Series: Materials Science and Engineering, 1022(1), p. 012042. Available at: https://doi.org/10.1088/1757-899X/1022/1/012042.
28. Aslam, U. et al. (2019) 'An Empirical Study on Loan Default Prediction Models', Journal of Computational and Theoretical Nanoscience, 16(8), pp. 3483–3488. Available at: https://doi.org/10.1166/jctn.2019.8312.
29. Zhu, Q. et al. (2022) 'Loan Default Prediction Based on Convolutional Neural Network and LightGBM':, International Journal of Data Warehousing and Mining, 19(1), pp. 1–16. Available at: https://doi.org/10.4018/IJDWM.315823.
30. Owusu, E. et al. (2023) 'A Deep Learning Approach for Loan Default Prediction Using Imbalanced Dataset':, International Journal of Intelligent Information Technologies, 19(1), pp. 1–16. Available at: https://doi.org/10.4018/IJIIT.318672.