ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Master Thesis Data Science and Marketing Analytics

Classification on Imbalanced Data – The Case for Credit Card Fraud Detection

Student name: Jivago de França Aires Galvão

Student number: 579943

Supervisor: Dennis Fok

Second assessor: Bas Donkers

Date Final Version: 06/11/2022

Abstract

The application of machine learning algorithms to classification problems in imbalanced data has a range of different considerations with regards to data preparation, performance metrics and the reliability of its generated probability scores, when compared to cases in which datasets are balanced. This research aims to assess suitable combinations of machine learning algorithms with resampling and probability calibration techniques used to tackle the issues that arise from imbalanced data. Specifically, Random Forest and Support Vector Machine were combined in different ways with SMOTE, Platt Scaling and Isotonic Regression, using one real-world and four generated datasets, all with different proportion of class imbalance. The performance of these combinations of techniques were measured in terms of ROC AUC, PR AUC, and Brier Score. It was found that, overall, Random Forest achieved superior results when compared to SVM. Furthermore, the combined use of Random Forest and Isotonic Regression provided the best results of PR AUC and Brier Score in all the five datasets analysed.

Table of Contents

Introduction

Binary classification tasks have a wide range of applications in different business sectors and have been continuously object of academic research. The cases of rare events, in which binary classification is performed using imbalanced data, are often examined. On these cases, conventional metrics for measuring classification performances might not be suitable to use. Classification of rare events using imbalanced data is applied in different sectors such as finance, medicine, and marketing, for instance. In marketing, one of the cases in which datasets are imbalanced usually involves the so-called database marketing models, which focus marketing efforts to existing clients aiming at growing customer loyalty. These efforts come under the assumption that enhancing the relationship with current customers brings more profit than acquiring new ones (Duman, Ekinci & Tanrıverdi, 2012). The classification models in database marketing focus on classifying customers as buyers and non-buyers in the context of cross selling or up selling. The former refers to finding a given existing customer who would be more likely to buy a new product he does not have, whereas the up selling refers to finding customers that may grow the volume of purchases of a given product they already buy. In medicine, one of the applications of classification tasks of rare events is on cancer detection where the patient is classified as ill or not ill, a problem which relies on the minority class samples (Fotouhi, Asadi & Kattan, 2019).

This paper focuses on one of the applications in finance, namely credit card fraud detection, and aims at finding a suitable combination of techniques to address some of the matters that usually arise when classification of rare events is performed. Specifically, it will deploy Random Forest and Support Vector Machine, two machine learning models commonly used in classification of rare events, as will be shown in the literature review section below. Following that, it will assess combinations of a data resampling technique and two probability calibration techniques in order to enhance the performance of the classification model.

In this section, the background of imbalanced classification applied in fraud detection for credit card transactions will be discussed, with emphasis on the overall effects it has on the credit card industry. Furthermore, the research objectives of this paper will be outlined, as well as the organizational structure it will follow.

Background

The issuance of credit to economic agents is a vital part of the dynamism in any economy, being an essential driver of household consumption and entrepreneurship activities. The organic evolution of banks' financial instruments led to the creation of credit cards as a portable way to purchase on credit. As it happens in other sectors, transactions involving credit cards are also subject to fraud. A report from the European Central Bank shows that in 2018, the value of credit card transactions that were fraudulent reached 1.8bn Euros, out of a total 4.84tr Euros in transactions that happened in the same period (European Central Bank, 2018). This encompasses those credit cards that were issued within the Single Euro Payments Area (SEPA). The same report shows that the value of fraudulent transactions grew at a higher rate than the value of overall credit card transactions between 2017 and 2018. The value in frauds for credit cards transactions increased by 13% year on year, whereas the overall value in transactions increased by 6.5% year on year.

Currently, fraud in credit card transactions can be separated into two types: card-present (CP) and card-not-present (CNP). The former refers to frauds in which the physical card is necessary such as in the case of transactions in ATM machines, whereas the latter refers to situations in which no physical card is required such as transactions on the internet. The aforementioned report by the European Central Bank points that in 2018, the frauds of type CNP represented 79.5% of the total value of frauds. Hence showing that nowadays frauds conducted remotely on the internet, mail or phone are more usual than the ones conducted physically.

Financial institutions are increasingly using machine learning techniques to be able to cope with the large amount of data that is available nowadays and promote the drive towards automated credit card fraud detection. In the past few years, different studies have focused on the application of machine learning to credit card fraud detection (Adewumi & Akinyelu, 2017; Popat & Chaudhary, 2018; Priscilla & Prabha, 2019). Those studies used different machine learning algorithms to perform the classification task, such as Random Forest, Naïve Bayes, Neural Networks, Support Vector Machines and Logistic Regression, to name a few.

Rare events are those in which an outcome occurs with remarkedly less frequency than other more common outcomes (Maalouf & Trafalis, 2011). Credit card fraud is one of the examples of rare events, given the rare nature of fraudulent transactions when compared to those that are legit. Rare events can have a substantial impact on their related fields if not detected properly. For instance, consider the impact of a non-diagnosed cancer on a patient or the effect of large frauds to a credit card company. Hence, the size of the effects of rare events makes their detection an important task, which can contribute to enhance their understanding and prevention. Decisions involving rare events usually present asymmetric costs (Byron, Wallace & Dahabreh, 2012), and credit card detection is not an exception to that. The costs involving correct and incorrect classifications are different. When it comes to fraud detection, the main desired outcome is to accurately detect the frauds so that the costs associated with fraudulent activities are avoided. However, once a fraud is detected, there are also costs associated with the investigations or other measures that are taken to tackle the fraud. In addition to those, there are also costs involved for the cases of false positive and false negative outcomes, and they are also asymmetric. When false negative outcomes happen, the company incurs on the losses coming from the frauds, whereas in case of false positives it incurs on the costs involved with the measures it takes when frauds are detected, as well as the reputation cost arising from bothering clients due to misclassification of fraud.

In imbalanced scenarios such as the case of credit card fraud, the errors tend to occur more often towards the minority class, if the classification methods are applied naively. That is, the model tends to classify an observation as belonging to the majority class, when in fact it belongs to the minority. In the case of credit card fraud, the model misclassifies a transaction as legitimate, when in fact it is fraudulent, which consists in a false negative. This situation happens because the classification models are trained in an imbalanced scenario in which most of the observations consist of legitimate transactions. The fact that in fraud detection for credit card transactions the costs associated with false negatives are usually larger than false positives highlights the importance of addressing the issue of classification under imbalanced data.

There is a wide range of techniques that can be used to tackle the issues of classification tasks on imbalanced data. Some of them focus on manipulating the data

itself, while others focus on tweaking the machine learning algorithms. The former type usually involves changing the training dataset through resampling so that the data gets a more balanced proportion of observations between the two classes. That can be applied through undersampling the majority class (e.g., random undersampling) or oversampling the minority class (e.g., SMOTE). The other sort of techniques encompasses cost-sensitive analysis and involve tweaking the machine learning algorithms so that it considers the different costs associated with each outcome. Another group of techniques used when classification is applied on imbalanced data is probability calibration techniques, which tackle the issue of predicted probabilities not matching the true probabilities. There are different methods used for calibrating probabilities, such as Platt Scaling and Isotonic Regression. These techniques will be further explained on the literature review and methodology sections.

Research Objective

In view of the background provided above and the current challenges involving the detection of credit card fraud, this paper will aim at analysing possible ways to improve the performance of machine learning models trained for fraud detection by combining resampling and probability calibration methods commonly used to handle issues arising from imbalanced data. Given the difficulty in finding public real-world data related to credit card frauds, this paper will simulate four datasets with different levels of imbalance, in addition to the use of one real-world dataset. Random Forest (RF) and Support Vector Machine (SVM) models will be trained, and different combinations of Platt Scaling, Isotonic Regression and the SMOTE technique will be subsequently applied on the different datasets used in this study. The steps taken aim at assessing whether SMOTE, Isotonic Regression and Platt Scaling can be used combined to improve the performances of RF and SVM models for the simulated datasets, and which combination performs the best for each proportion of class imbalance considered.

Furthermore, by applying the methodology on four different imbalance scenarios, the research aims at giving results applicable under the normal highly imbalanced fraud scenario, as well as under exceptional ones, in which the rate of fraudulent detections would increase drastically compared to what has been usually observed by real-world

data. In that way, this paper also aims at detecting which combination of techniques are more suitable for each type of dataset, according to the level of imbalance it presents. Specifically, the level of imbalance on the first dataset was 10%. The other simulated datasets have a gradually decreased proportion of minority class, with 5%, 1% and 0.2%, respectively. The latter proportion was specifically chosen to be close the imbalance of the real-world dataset used, which has around 0.17% of observations classified as fraud.

Paper Structure

Following this introduction section, this paper will present a literature review involving credit card fraud, as well as classification algorithms, resampling methods and probability calibration techniques previously implemented for that problem. After the literature review, a section about the data simulation will be exposed, focusing on the methodology used to simulate the dataset. Then, there will be a section explaining the methodology used for conducting the investigations of this paper. In particular, the machine learning models, resampling technique and calibration methods, as well as the performance measures will be explained. Then, the results obtained will be presented and the insights derived from them will be exposed. Finally, a general discussion section will explore the implications of the insights, as well as the limitations of this paper and the fields for future research.

Literature Review

This section will go through the previous studies conducted about fraud detection, as well as past research about methods that are applied throughout the paper to reach the prediction of credit card fraud.

Fraud Detection

The problem of fraud detection has been widely investigated in previous research. Phua et al., (2010) presented an extensive review of studies addressing the issue of automated fraud detection in many domains. They identify the most common occasions in which fraud occurs, for instance, on medical and housing insurance, credit application, telecom subscription, credit card transactions, among others.

In the specific field of fraud in credit card transactions, different investigations were conducted. Chen et al., (2004) proposed the so-called questionnaire-responded transaction (QRT) data, collected through online questionnaires, and then use SVMs to train the data and develop a model to predict new transaction. Maes et al., (2002) used Artificial Neural Networks (ANN) and Bayesian Belief Networks (BNN) to credit card fraud detection. They found that the use of BNN produces better results and has a faster training process, however, the process of fraud detection was found to be quicker using ANN.

Bhattacharyya et al., (2011) further investigated the use of data mining techniques in credit card fraud by comparing the use of Random Forest, Support Vector Machine and Logistic Regression. They found that, overall, Random Forest models presented a better performance than SVM. They also emphasized that Logistic Regression had good performance, often better than the ones achieved using SVM.

Chan et al., (1999) examines credit card fraud detection in the context of e-commerce, with a focus on the imbalance of data, the non-uniform nature of costs when classification errors occur, as well as the issue of distributed databases. They proposed combining multiple machine learning algorithms, training them on subsets distributed data in a distributed environment designed by the researchers. They also considered the asymmetry of costs by implementing a cost-sensitive version of the AdaBoost algorithm, the so-called AdaCost. They found their proposed methods were

effective in building fraud detectors, as well as scalable in a distributed environment. One limitation, however, is the necessity of determining the desired distribution of the training set in accordance with a cost model.

Imbalanced data

Real-world data sets used in binary classification problems are usually imbalanced, in the sense that the majority observations belong to one of the classes only. That happens in different applications, from medicine (e.g., cancer detection) to finance (e.g., fraud detection in credit card transactions). On imbalanced data sets, the target binary variable Y has unequal distribution of the classes, that is, one of the classes is over-represented whereas the other is under-represented. For instance, in the case of credit card transactions, most of the observations of transactions are classified as non-fraudulent (Y=0) and represent the majority class, whereas the minority class encompasses the fraudulent transactions (Y=1), which are a low proportion of the overall data.

The presence imbalance in the data classes make the accuracy of the model an unreliable metric of performance measure. Hence, if the algorithm is built on the assumption of maximising the accuracy, the machine learning model will generate unsatisfactory classifiers (Provost, 2000). For instance, in hypothetical data set which has 99% of the observations belonging to Y=0 and 1% belonging to Y=1, a model that predicts that all the observations in the test set would belong to Y=0 would have a 99% of accuracy. Hence, accuracy is not the most suitable measure for evaluate model performance in the case of imbalanced data.

Sun, Wong & Kamel, (2009) addressed the use of evaluation metrics for classification under imbalanced data, emphasizing that different metrics can be used according to the learning objective of the classification task. They also point to the fact that accuracy is not an appropriate measure because the minority class has little impact on the accuracy, when compared to the majority class. The study mentions the fact that, when the performance of the positive class is important, the precision and recall are measures two important measures, also highlighting the link between the two measures and the F-measure, which would be the harmonic mean of these two metrics. They also mention the ROC AUC as a threshold-free metric that can be used

on these scenarios of classification under imbalance. In the research examining the relationships between ROC curves and PR curves, Davis & Goadrich, (2006) argue that in case of highly skewed dataset, the PR curve provides a more accurate portrait of the model's performance.

In view of the issues that imbalance data can bring to the model performance, different techniques can be applied to handle this situation. Examples of techniques are data resampling methods, cost-sensitive learning, and probability calibration.

Resampling Techniques

The resampling techniques can be divided into undersampling, oversampling and hybrid. In general, they focus on resampling the training data to make the data set more balanced.

Oversampling

In oversampling, the minority class has its proportion of observations increased in the dataset. One of its forms is the so-called random oversampling (ROS), by which the observations are duplicated randomly. There are also more sophisticated approaches to oversampling, such as Synthetic Minority Oversampling Technique (SMOTE) and Adaptative Synthetic Sampling (ASASYN). In SMOTE, the oversampling of the minority class occurs by the generation of synthetic examples in the feature space (Chawla et al., 2002). The ASASYN, in turn, considers the data distribution and adaptatively generates synthetic samples belonging to the minority class. It considers the level of difficulty in learning the observations belonging to the minority class, hence generating more synthetic observations for the minority class examples that are relatively more difficult to learn (He et al., 2008).

Undersampling

The methods belonging to undersampling remove observations from the majority class to make it more balanced. Analogously to oversampling, the simplest method to conduct undersampling is to randomly remove these samples from the majority class, known as random undersampling (RUS). Other approaches for undersampling involve the use of the KNN method (Mani & Zhang, 2003), Tomek Links (Tomek, 1976) and the so-called Edited Nearest Neighbours (DL Wilson, 1972), all focusing on the

removal of observations from regions that overlap. Another technique involves undersampling with a cluster-based approach (SJ Yen, YS Lee 2009).

Hybrid

Hybrid techniques involving the combination of oversampling and undersampling techniques used together have also been proposed by previous research. Batista, Prati & Monard (2004), for instance, propose the combination of SMOTE with Tomek and ENN.

Resampling Methods Applied in Fraud Detection

Recent studies have applied resampling techniques in the context of credit card fraud transactions. Mrozek, Panneerselvam, Bagdasar (2020) applied random undersampling and the SMOTE methods into a real-world dataset, using Random Forests, Logistic Regression, K-Nearest Neighbours, as well as Stochastic Gradient Descent. They found that Random Forest in combination with Random Undersampling got the best recall score, when compared to just using the Random Forest without addressing the imbalance of the data. Sisodia, Reddy & Bhandari, (2017) applied different oversampling techniques (e.g., SMOTE, SMOTE ENN, SAFE SMOTE, SMOTE TL, and Random Oversampling), and then used cost-sensitive analysis on Adaboost and Bagging. They concluded that SMOTE ENN performs better in detecting frauds.

Calibration of Probabilities

In binary classification problems, it is not unusual that the probabilities obtained by the machine learning models applied do not match the true probabilities. That issue is particularly relevant when the dataset used is imbalanced, in which the costs of misclassification are commonly asymmetric (Wallace & Dahabreh, 2012). In particular, Niculescu-Mizil & Caruana (2005) demonstrate that models such as SVMs, boosted trees and boosted stumps usually push the predicted probabilities away from the 0 and 1 threshold.

The probability calibration of classification models has been examined in previous studies. In particular, methods which are applied after a given model is fit have received attention, the so-called post-processing calibration methods. Different

combinations of machine learning models and calibration methods have been applied by previous research. Zadrozny & Elkan, (2002) propose the use of a common algorithm used in Isotonic Regression, the so-called pair-adjacent violators (PAV) algorithm, to enable the learning of mapping from the ranking scores to the estimates of calibrated probabilities. They also use the PAV algorithm for the case of multiclass probabilities.

Combining different calibration metrics with learning models, Niculescu-Mizil & Caruana, (2005) demonstrate that AdaBoost predicts probabilities that are distorted and apply three different calibration methods to handle this issue, namely, Platt Scaling, Isotonic Regression and Logistic Correction. They found that Platt Scaling and Isotonic Regression enhance the probabilities predicted by both Boosted Trees and Boosted Stumps, while Logistic Correction worked well only on Boosted Stumps. They also found that boosted full decision trees after calibrations perfom better probability predictions than SVMs, KNNs and Neural Nets. Wallace & Dahabreh, (2012) also address the problem of probability calibration in imbalanced datasets. They argue that in imbalanced scenarios, the estimator provides unreliable probability estimates when it comes to the minority class and propose the use of the so-called stratified Brier score metric to measure this issue. The solution they propose is to use balanced bootstrap samples of the training data to induce the probability calibration.

In a recent research, Huang, et al., (2020) conducted a large experimental investigation on calibration under imbalanced datasets. They used Logistic Regression, Random Forests, Support Vector Machines and Gradient Boosting Decision Tree as classification models, and tested the calibration using different methods, namely, Platt Scaling, Histogram Binning, Isotonic Regression and Bayesian Binning into Quantiles. They concluded that Isotonic Regression has the best performance overall.

In the context of the literature review presented in this section and the contributions made by the wide range of studies conducted in the field, this paper aims at addressing different issues that arise from classification under imbalanced data, within the same framework. Although the use of probability calibration methods was previously investigated to improve the match between predicted and true probabilities, and although methods of resampling were previously applied on imbalanced classification

tasks to improve their performance metrics, this research will expand their scope to the context of credit card fraud under different imbalance scenarios that were generated by simulation. In particular, it aims to combine the use of these two types of methods and provide a suitable choice combination so that both probability calibration metrics (i.e., Brier Scores) and measures related to multiple thresholds (i.e., the areas under the curve of ROC and PR curves) are optimised.

Data

One of the main issues on the research about data mining applied to fraud detection is the scarce amount of publicly available real-world data that can be used to conduct the studies (Phua et al., 2020). In the case of datasets regarding credit card fraud, this is specially the case due to confidentiality issues. Considering this matter, part of the datasets used in this paper are the results of data generation process. Four datasets were generated using different parameters to control for imbalance and how easily the classification could be done. Furthermore, one real-world dataset is also used with the aim to compare the results and assess whether the most suitable combination of methods obtained on the simulated datasets would also be applicable to this particular real-world data. The details of the datasets are explained below.

Real-World Data

The dataset used to represent a real-world scenario encompasses two days of transactions that occurred in September 2013 by cardholders in the European continent. In total, there were 284,807 transactions, out of which 492 were labelled as fraudulent, representing roughly 0.172% of the whole sample. The dataset has 31 variables, with 28 of them being the result of Principal Component Analysis (PCA). The variables which are transformed using PCA usually display a wide range of behavioral and demographic information about the clients. The others are “Time”, “Amount” and “Class”. The latter is binary and labels the transaction as Fraud (Class = 1) or Non-Fraud (Class = 0). The variable “Amount” refers to the transaction amount, whereas “Time” refers to the number of seconds that have passed from the very first transaction computed and any other given one. This dataset was produced and collected as part of a partnership between the Machine Learning Group of the Université Libre de Bruxelles and the payment solutions company Worldline.

Data Simulation

The data generation process was conducted such that four datasets with different class imbalance were generated. The simulation was done using the *make_classification* function present in Python’s Scikit-learn library, which enables the creation of a random n-class classification problem, generating clusters of points normally distributed in a hypercube. The number of features generated was set to 30,

which is the same number of features used in the training of the real-world dataset, after the feature engineering performed. The features were generated to resemble the ones present in the real-world data described above, including the binary variable “Class”. The process generated 200,000 observations to resemble the number found in the real-world dataset. To control for ease with which the classification can be performed, the parameter *class_sep* was adjusted. This parameter multiplies the hypercube size, with larger values implying that the classes are more spread out in the hypercube and making the classification easier. The default for this parameter is 1, but the value used was 0.5, so that the classification task is harder, and the models applied do not predict fraud perfectly, which wouldn’t be compatible with what happens in real-life.

The control of the class imbalance was done by adjusting the parameter *weights*, which basically assigns the proportion of weights on each of the classes for the simulation. For the purposes of this research, four different proportions were considered in the data generation process, namely 10%, 5%, 1% and 0.2%. The choice of proportion in the highly imbalanced scenarios of 1% and 0.2% was done to resemble to the scenarios displayed in the reports by the European Central Bank, mentioned in the section above, in addition to be near the imbalance proportion of the real-world dataset used in this study. Hence, these two scenarios would be easier to observe, according to the reports on fraud publicly available currently. The other two less imbalanced scenarios, in which 10% and 5% of observations are frauds, aims at assessing how different the combination of methods would be in case of an improbable scenario in which the proportion of frauds increased drastically, compared to what is currently documented publicly.

Methodology

In this methodology section the machine learning models, resampling technique, probability calibration methods, and performance metrics used in this paper will be explained one by one. Following that, the exact steps taken for conducting the investigation will be exposed, in order. Each step taken will be justified considering the methods explained in this section, as well as the previous approaches used to investigate credit card fraud, explained in the literature review section above.

Machine Learning Models

Support Vector Machine

Support Vector Machine (SVM) will be one of the machine learning methods used to perform the detection of fraud in credit card transactions. It is a type of supervised learning that can be used for both classification and regression tasks. The main idea of the method is to discover a hyperplane in a N-dimensional hyperspace that can classify data points. The hyperplane (the classification function) is determined such that it has the maximal margin separating the classes (i.e., the maximum distance between the nearest data points and the hyperplane). This characteristic minimizes the risk of overfitting the training data. The hyperplane separating the different classes can be expressed as shown below.

$$< w, x > + b = 0$$

Where $< w, x >$ is the dot product of the coefficient vector w and vector variable x.

The SVM can use a Kernel function, which represents the dot product of two data points in a high-dimensional feature space. Hence, the SVM classification function can be expressed in terms of dot products of input data points in a high-dimensional feature space. SVM has some characteristics that make it appropriate to apply in classification problems with imbalanced data such as credit card fraud. Particularly, the fact that it is a linear classifier but can work properly in a high-dimensional feature space without the need of implementing further computational complexity (Bhattacharyya et al., 2011).

Random Forest

The other method used in this paper for fraud detection will be Random Forest (RF), which belongs to the class of the so-called ensemble models. The latter refer to the sort of methods that develop a set of models and aggregate the predictions made to determining the output class label for a given data point. In the case of Random Forests, the aggregation involves several decision trees. After the trees are generated, they vote for the most popular class for the input x (Breiman, 2001). Specifically, it addresses one of the drawbacks of decision trees, namely their sensitivity to specific training set, which can lead to overfit. The trees built in a Random Forest model are done on bootstrapped samples of the chosen training data. Another aspect of this method is that each time a node is built, only some previously selected subsamples of attributes are selected, randomly. That differs from the Bagging method, in which all the attributes of a model are considered when the node is built.

Performance Measures

The performance measures for classification problems, either binary or multi-class, take into consideration the so-called confusion matrix (aka contingency table). The confusion matrix shows the possible outputs of a classification model. In the case of fraud detection, which is a binary classification problem, the confusion matrix will be 2x2 showing four possible outcomes. An illustration of confusion matrix for binary classification, as defined by Tharwat, A. (2020), is shown below.

	Y_0	Y_1
$\hat{Y}_0$	TN	FN
$\hat{Y}_1$	FP	TP

In the matrix, TN and TP represent true negative and true positive, respectively, whereas FP and FN refer to false positive and false negative. In the context of credit card fraud detection, a true positive instance would indicate that a transaction was predicted to be fraudulent, and it actually was. A false positive, in turn, means a given transaction was predicted as fraudulent but in fact it was legit. True negative refers to those that were predicted as genuine and were genuine, and finally false negative are those that are predicted as genuine but were actually fraudulent.

The imbalance present makes accuracy an unsuitable metric for measuring the performance of the machine learning models used. Accuracy is defined by the equation below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

To illustrate this in the case of fraud detection in credit cards, if a model naïvely predicts all the observations in the test set to belong to the majority class of genuine transactions (Y=0), this model will have a high accuracy. However, a model like that would not be useful in practice, since the main interest of a fraud detection system is to identify the frauds, which would not happen in case all the observations were classified as legit.

Other set of metrics are also used for measuring performance involves Sensitivity (True Positive Rate), Specificity (True Negative Rate), False Negative Rate, and False Positive Rate. All of them are based in the given probability threshold that is chosen in the problem, that is, if the threshold changes, so does the value computed by the metric. Each of them is briefly explained below.

True Positive Rate

The True Positive Rate (aka sensitivity, recall or hit rate) is computed as the number of true positive instances divided by the total number of positive samples, as shown in the equation below.

$$Recall = \frac{TP}{TP + FN}$$

Analogously, the False Negative Rate (FNR) is defined as $FNR = 1 - Recall$

True Negative Rate

The True Negative Rate (aka specificity or inverse recall): is the ratio of correctly classified negative samples and the total number of negative instances, shown below.

$$Specificity = \frac{TN}{TN + FP}$$

The False Positive Rate (FPR) is defined as $FPR = 1 - Specificity$

The metric Precision is also commonly used for evaluating the performance of models. It is given by the following formula.

$$Precision = \frac{TP}{TP + FP}$$

In the context of fraud detection, precision measures the proportion of transactions that are actually fraudulent, among those that were classified as frauds.

There are applications of more metrics to address the model performance. However, this paper will keep the ones mentioned above as the main ones, as they are important for the understanding of the other two metrics that will be the main ones used in the methodology of this paper, as explained in the next sub-section.

The methods explained above can only be calculated once a confusion matrix is designed and they depend on the probability threshold chosen for the problem. It is often valuable to also assess the performance of a model considering different thresholds. Fort that, two metrics will be explained, the Receiving Operating Characteristic (ROC) and the Precision-Recall (PR) curve. These two metrics can also be used for the selection of an optimal probability threshold for the classification problem, which will be explained below.

Receiving Operating Characteristic

The Receiving Operating Characteristic is a graph used for visualizing and selecting classifiers according to their performance Fawcett, (2006). It is obtained by plotting the TPR (y-axis) against the FPR (x-axis), considering different thresholds. The comparison between different classifiers in this case is done by using the area under the curve, the AUC ROC. The model with a higher ROC AUC is said to have a better average performance. Theoretically, the value of ROC AUC can be in the interval [0,1], but since the random classifier produced in the graph connects the coordinates (0,0) to (1,1), its AUC is 0.5 and no realistic model will have an area below that. Hence, ROC AUC value usually ranges in the interval [0.5,1]. The ROC AUC can be seen as the probability that a given classifier will rank a positive instance higher than a negative

one, both chosen randomly. Furthermore, ROC are monotonic functions, i.e., the TPR only increases if the FPR also increases.

The ROC curve can be used for the selection of a probability threshold for the problem, which would visually be located the closest to the top-left part of the plot. The threshold corresponding to this point is the one with the largest value of the so-called Youden's J statistic, defined by the equation below.

$$J = Sensitivity + Specificity - 1$$

Precision-Recall Curve

The Precision-Recall (PR) Curve is the plot of Precision (y-axis) and Recall (x-axis), considering different thresholds. One of the interesting points of PR curves in the context of credit card fraud detection is that it enables to highlight those classifiers that have both high Recall and high Precision (i.e., high TPR and low FPR), which is something desirable when obtaining a model to detect fraud. A detection system with high Recall and high Precision would generate many fraud alerts, and most of them would be correct. The performance is also assessed using the area under the curve. A model with a higher PR AUC is said to perform better, comparatively. For computing the PR AUC, the Average Precision (AP) is used as a metric. It is calculated by the formula below.

$$AP = \sum_{n} (R_n - R_{n-1}) * P_n$$

Where R_n and P_n are the Recall and Precision, respectively, and the subscript n refers to the nth threshold.

When using both ROC and PR curves for measurement of model performance, it makes easier to compare both by noting that in the PR curve the recall is in the x-axis, whereas in the ROC curve recall is in the y-axis. Although PR AUC does not have a statistical interpretation such as the ROC AUC, Davis & Goadrich, (2006) point that for a curve to dominate in the ROC space, it has necessarily to dominate in the PR space as well. However, the two plots also have differences. Unlike the ROC, the PR curve is not monotonic. Furthermore, while the ROC AUC value for the random

classifier is always 0.5, that is not the case for the PR AUC, as the value of the latter depends on how imbalanced the data is.

In the case of imbalanced datasets, the PR curve is normally used as alternative to the ROC, as it can enables one to grasp some differences in classifiers' performances that are not grasped by the ROC (Boyd, Eng, Page, 2013). Pointing to the same issue, Saito & Rehmsmeier, (2015) argue that PR curve plots provide a more intuitive and accurate interpretation of the performance of the classifier, as well as show the susceptibility of the models to the imbalanced dataset. Given that, the PR AUC will be the main method used for analysing the performance of the models in our problem. The PR AUC will be used to assess how the model changes its performance when probability calibration and/or SMOTE is done, that is, performance will be assessed before and after the proposed methods are applied.

The PR curve can also be used to choose the best probability threshold in the classification problem. The best threshold is the one that provides the best balance of Precision and Recall, located the closest to the top-right part of the plot. This can be done by maximising the F-Measure, which is the harmonic mean of Precision and Recall, given by the equation below.

$$F - Measure = \frac{(2 * Precision * Recall)}{(Precision + Recall)}$$

Brier Score

The Brier Score computes the fit of the probability estimates obtained by a given model to the true label (i.e., observed data). The metric is given by the equation below.

$$\frac{\sum_{i=0}^{N}(y_i - \hat{P}\{y_i|x_i\})^2}{N}$$

Where N is the sample size and $y \in \{0,1\}$.

If the probability estimates diverge from the true label, the score is high. Analogously when the estimates are near the observed data, the score is low. Therefore, in case the calibration method generates enhanced probability estimators, the Brier Score should be smaller, when compared to the one obtained by uncalibrated models.

Resampling Technique - SMOTE

One of the commonly used approaches to deal with imbalanced datasets involve the application of resampling techniques, aiming at making the target variables more equally distributed. In the context of credit card fraud, it implies that there is no such a big difference in the number of frauds and non-frauds in the dataset. As pointed out in the literature review, there are different ways to conduct resampling. To avoid unnecessary loss of information that might come from undersampling the majority class, this paper will focus on the application of oversampling. Given its extensive use in previous studies related to fraud detection, the SMOTE technique will be the one used in this research. This technique performs the synthetic oversampling of the minority class, which contrasts the other commonly used technique of random oversampling with replacement. To create the synthetic observations, the difference between the feature vector being considered and its nearest neighbour is taken, then this difference is multiplied by a random number in the [0:1] interval and following that the result is added to the feature vector that is being considered (Chawla et al., 2002). Hence, this technique introduces bias towards the minority class, which is something desirable in the context of credit card fraud.

Probability Calibration

In the context of fraud detection, and in many other real-world applications, it is also useful to obtain the class probabilities. In addition to the previously mentioned bias in the outputs towards the majority class, the use of imbalanced datasets for classification tasks also results in uncertainty regarding the probability calibration. Uncalibrated probabilities can also make the default decision thresholds not to be optimal. To address this issue, this paper will use two post-processing methods for calibrating the probabilities of the models, namely Platt Scaling and Isotonic Regression, explained below.

Platt Scaling

The Platt Scaling (Platt, 1999) is a probability calibration method, originally proposed to convert to posterior probabilities the outputs of SVMs by passing them through a sigmoid function, hence transforming the outputs from $[-\infty, +\infty]$ to probabilities of

outcome. The transformation is expressed by the equation below, where $f(x)$ is the output of a given machine learning method, such as SVM.

$$P(y = 1 \mid f) = \frac{1}{1 + \exp(Af + B)}$$

Where A and B are parameters fitted with maximum likelihood estimation using a given training set (f_i, y_i). The parameters A and B are found by using the gradient descend, coming from the solution of a particular loss function.

To avoid introducing bias and get good probability estimates, a different set should be used to perform the calibration. Hence, a validation set will be used to calibrate the probabilities, which is a different set from the training and test sets.

Isotonic Regression

The Isotonic Regression (Robertson et al., 1988) is a general method used for calibration of probabilities. This method requires that the mapping function should be monotonically increasing. The Isotonic Regression assumes the following:

$$y_i = m(f_i) + \epsilon_i.$$

Where y_i are the true values, f_i are the predictions and m is a given isotonic function. The regression itself is expressed as the optimization problem shown below, which finds the isotonic function $\hat{m}$ for a train set (y_i, f_i):

$$\hat{m} = argmin \sum (y_i - z(f_i))^2$$

Like in the case of Platt Scaling, an independent validation set will be used to perform the Isotonic Regression, aiming to avoid bias.

Research steps

In view of the methodology presented above, as well as the approaches taken by previous studies in the field, this paper will conduct the investigations on five datasets using the framework presented in the data section – four simulated datasets with four different proportion of imbalance in the data, and a real-world dataset. Each of them

will be split into train, validation, and test data. The train data will be used to fit the machine learning methods deployed (i.e., RF and SVM), while the validation set will be used for implementing the probability calibration methods. Finally, the test set will be used for assessing the fit of the model and whether it has improved its performance after the resampling and calibration methods. The probability scores for Random Forest will be obtained by applying the *predict_proba* method in Python's Scikit-learn library. In this case, the value of the predicted probabilities is obtained by calculating the mean predicted class probabilities of all the trees used in the Random Forest classifier. In each of the trees, the class probability is measured as the fraction of samples that belong to the same class in a given leaf. In the case of the SVM model, the probability scores will be obtained after the Platt Scaling and Isotonic Regression are applied.

The train set will then be fit on RF and SVM models. This will be done in the original datasets, as well as the train set that is generated after SMOTE is applied. Prior to the implementation of calibration methods on the models, their performances will be measured by using both the ROC AUC and PR AUC. Then, probability calibration with Platt Scaling will be applied on the models trained with and without SMOTE. Following that, the same step will be repeated, but this time applying Isotonic Regression as the calibration method. The Brier Score will also be computed to check the quality of the probability estimates after calibration. Furthermore, ROC AUC and PR AUC will also be computed to assess whether the models have improved their classification performances. The steps mentioned will be done on each dataset separately.

Results

This section will show the results obtained following the methods explained in the methodology section. As previously mentioned, the analysis was conducted on four different datasets that were generated using the same methodology. In addition, the same analysis was conducted in the real-world dataset about credit card fraud, whose details are also presented in the data section above. Apart from the different proportion of class imbalance, the generated datasets were simulated using the same parameters so that the comparison of performances could be more insightful.

On each dataset, the performance measures will be presented for different combination of techniques. The first case considered is the one in which no resampling (SMOTE) and no probability calibration technique is applied, that is, a naïve application of Random Forest and Support Vector Machine is used on the imbalanced datasets.

Another case considered is when SMOTE is applied but no probability calibration technique is conducted. This scenario enables the assessment of how well the model performs by only focusing on the application of SMOTE to the dataset to tackle the issue of imbalance.

Similarly, it is also considered the cases in which probability calibration is applied but SMOTE is not. Specifically, Platt Scaling and Isotonic Regression are the two calibration techniques used.

The last two cases will consist of the combination of all the techniques deployed in the scenarios mentioned above. One of them will be the application of SMOTE, followed by Platt Scaling, while in the other SMOTE will also be applied, but followed by Isotonic Regression instead.

The performance measures used for assessing the best combination of techniques will be the ROC AUC, PR AUC, and the Brier Score. Furthermore, the best probability threshold for the classification problem on each case will be presented, alongside the F1-score that is achieved if the threshold is used. The choice of best threshold for each case was done using the PR curve so that the model presented the best precision-recall balance. That was achieved by applying maximisation on the F-Score,

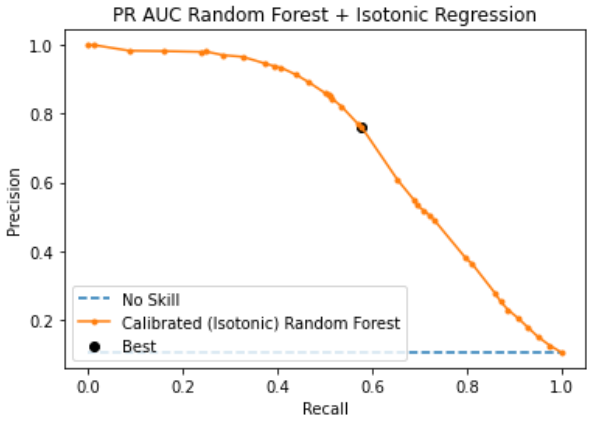
as explained previously in the methodology section. As the performance measure of each dataset is presented, a table summarizing them will be shown, as well as the PR and ROC curves resulted from the best combination of techniques on each case. In the PR AUC graph, the point in which the best threshold is located is also highlighted.

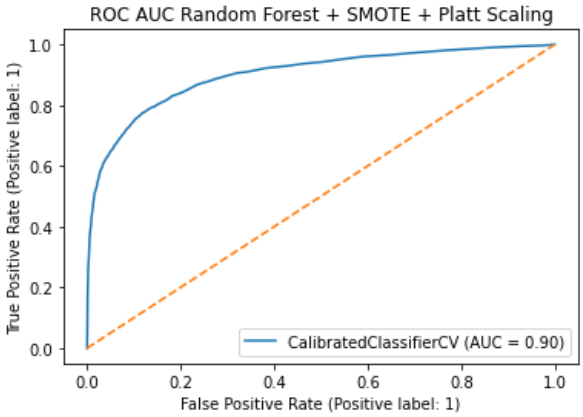
Dataset I (imbalance proportion: 10%)

Random Forest

The table below display the performance measures for each combination of techniques considered, applied on a Random Forest classifier. As can be noted, the case presenting the best results in terms of PR AUC and Brier Score is when Isotonic Regression is used solely. When it comes to ROC AUC, the case with highest AUC happens when SMOTE is applied, either in combination with Platt Scaling or used alone. The former case, however, comes with a better Brier Score since probability calibration is applied.

	ROC AUC	PR AUC	Brier Score	F-Score	Best Threshold	F-Score Best threshold
RF	0.8920	0.7129	0.0501	0.642	0.41	0.657
RF + SMOTE	**0.9012**	0.6811	0.0729	0.606	0.68	0.639
RF + Platt	0.8920	0.7129	0.0502	0.645	0.3498	0.657
RF + Isotonic	0.8915	**0.7131**	**0.0498**	0.636	0.3437	0.657
RF + SMOTE + Platt	**0.9012**	0.6811	0.0537	0.630	0.4402	0.639
RF + SMOTE + Isotonic	0.9002	0.6793	0.0530	0.636	0.3648	0.639

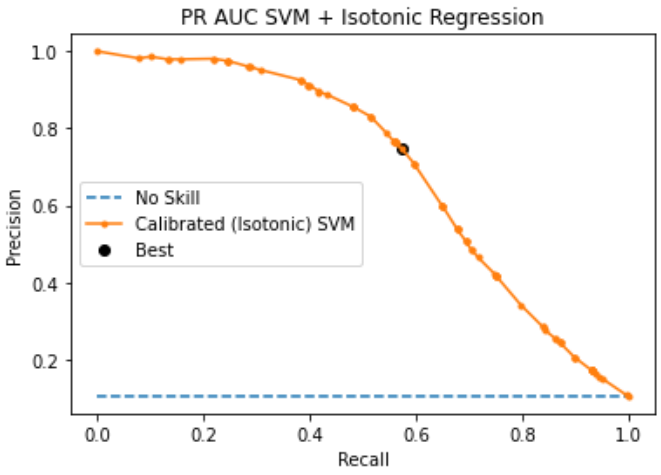


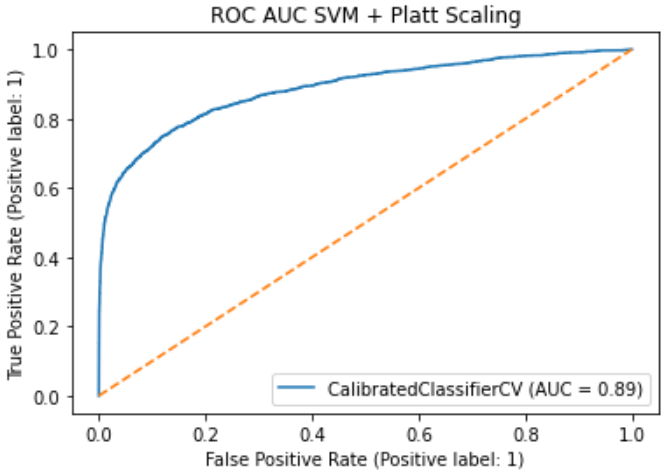


Support Vector Machine

In the case of Support Vector Machine, the use of Platt Scaling alone produces the highest ROC AUC. Similar to the case shown above, the best PR AUC and Brier Score are produced when the SVM classifier is calibrated with Isotonic Regression, without the application of SMOTE.

	ROC AUC	PR AUC	Brier Score	F-Score	Best Threshold	F-Score Best Threshold
SVM	0.8873	0.6976	-	0.577	0.3166	0.651
SVM + SMOTE	0.8725	0.6141	-	0.555	0.6757	0.576
SVM + Platt	**0.8873**	0.6976	0.0517	0.614	0.3168	0.651
SVM + Isotonic	0.8854	**0.6968**	**0.0515**	0.616	0.3548	0.649
SVM + SMOTE + Platt	0.8725	0.6141	0.0603	0.514	0.2938	0.576
SVM + SMOTE + Isotonic	0.8721	0.6144	0.0603	0.485	0.2812	0.576



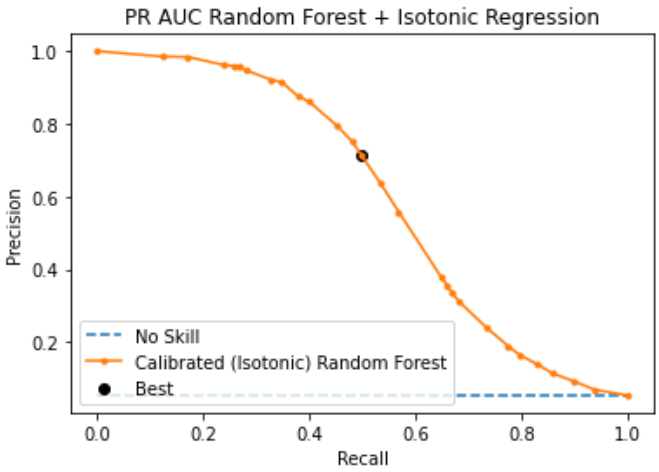


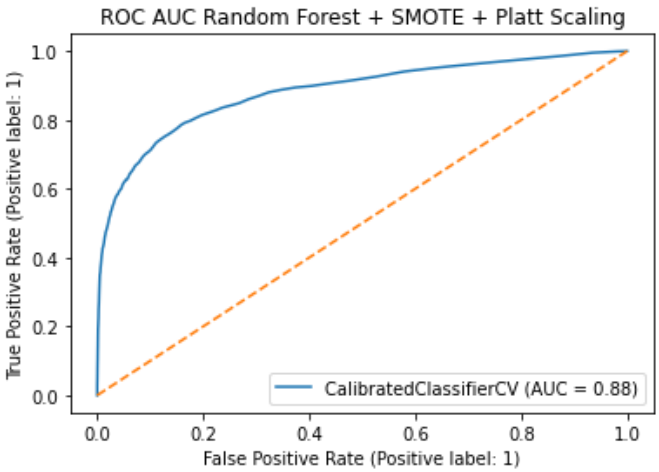
Dataset II (imbalance proportion: 5%)

Random Forest

The Random Forest applied in the dataset with class imbalance of 95% has the best ROC AUC when either SMOTE is used alone or in combination and Platt Scaling. The sole application of Isotonic Regression without SMOTE, on the other hand, produces the best measures for PR AUC and Brier Score.

	ROC AUC	PR AUC	Brier Score	F-Score	Best Threshold	F-Score Best threshold
RF	0.8626	0.6028	0.0311	0.556	0.39	0.590
RF + SMOTE	**0.8822**	0.5464	0.0569	0.5	0.71	0.54
RF + Platt	0.8626	0.6028	0.0308	0.564	0.3253	0.59
RF + Isotonic	0.8623	**0.6031**	**0.0307**	0.546	0.3207	0.587
RF + SMOTE + Platt	**0.8822**	0.5464	0.0340	0.532	0.3964	0.540
RF + SMOTE + Isotonic	0.8816	0.5488	0.0336	0.546	0.3842	0.538

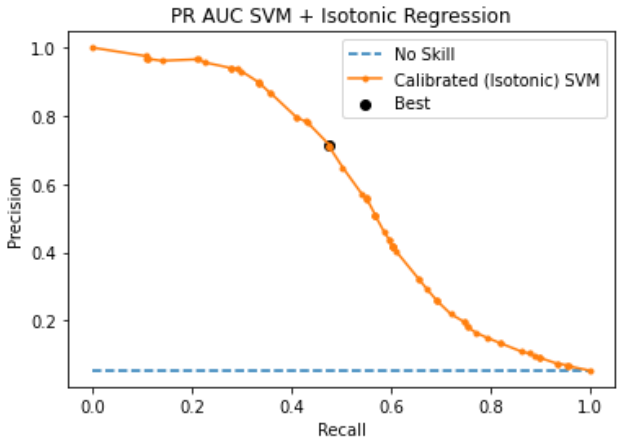


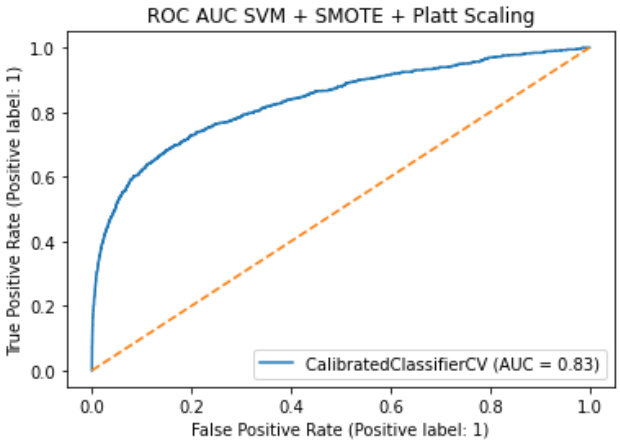


Support Vector Machine

The highest ROC AUC for the SVM classifier is obtained when Platt Scaling is used alone. The largest PR AUC in this case can be obtained by using two different combinations, both the Platt Scaling alone and the use of Isotonic Regression alone. Furthermore, the best Brier Score is also associated with the latter case of Isotonic Regression.

	ROC AUC	PR AUC	Brier Score	F-Score	Best Threshold	F-Score Best Threshold
SVM	0.8595	0.5785	-	0.374	0.2502	0.573
SVM + SMOTE	0.8343	0.4385	-	0.429	0.7001	0.452
SVM + Platt	**0.8595**	**0.5786**	0.0321	0.505	0.2530	0.573
SVM + Isotonic	0.8579	**0.5786**	**0.0319**	0.507	0.3687	0.572
SVM + SMOTE + Platt	0.8343	0.4385	0.0388	0.323	0.2285	0.452
SVM + SMOTE + Isotonic	0.8332	0.4383	0.0387	0.374	0.2343	0.450



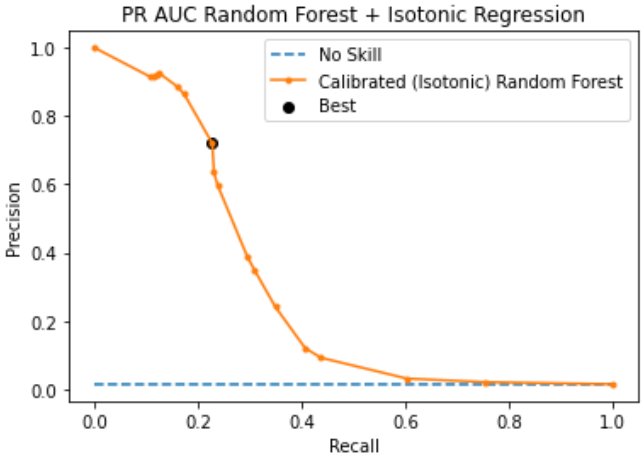


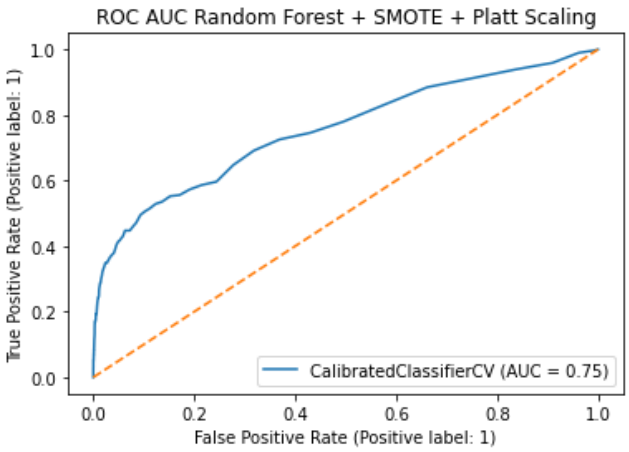
Dataset III (imbalance proportion: 1%)

Random Forest

The use of the Random Forest classifier in this case produces the best ROC AUC result when the SMOTE technique is applied, both alone and in combination with Platt Scaling. In terms of PR AUC, the use of both probability calibration techniques alone produces equally the best results, although the use of Isotonic Regression alone results in a slightly better Brier Score.

	ROC AUC	PR AUC	Brier Score	F-Score	Best Threshold	F-Score Best threshold
RF	0.7214	0.2891	0.0120	0.288	0.21	0.356
RF + SMOTE	**0.7539**	0.1779	0.0297	0.257	0.55	0.264
RF + Platt	0.7214	**0.2891**	0.0119	0.316	0.1084	0.356
RF + Isotonic	0.7180	**0.2891**	**0.0118**	0.288	0.4864	0.345
RF + SMOTE + Platt	**0.7539**	0.1779	0.01323	0.115	0.1219	0.264
RF + SMOTE + Isotonic	0.7476	0.1784	0.01329	0.288	0.1265	0.263

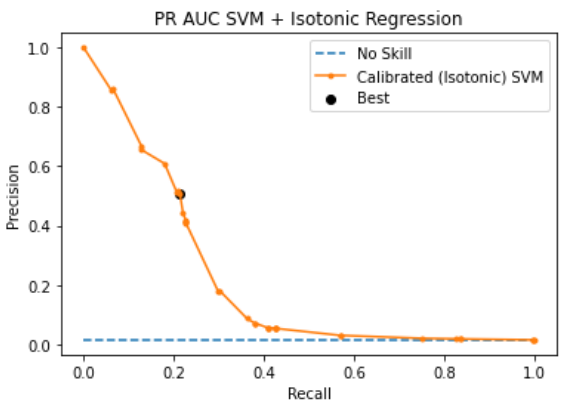


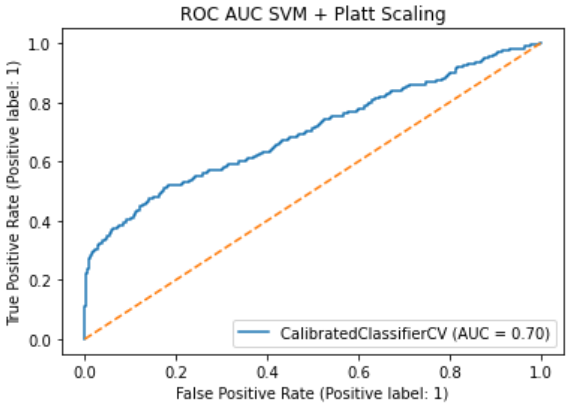


Support Vector Machine

The use of SVM in this dataset produces the best ROC AUC when Platt Scaling is used. The best PR AUC and Brier Score were obtained by using Isotonic Regression alone. In this dataset, the use of SMOTE with SVM did not produce superior results in any combination.

	ROC AUC	PR AUC	Brier Score	F-Score	Best Threshold	F-Score Best Threshold
SVM	0.6985	0.2140	-	≈ 0	0.1245	0.301
SVM + SMOTE	0.6335	0.1245	-	0.151	0.0100	0.167
SVM + Platt	**0.6985**	0.2141	0.0129	0.120	0.1468	0.301
SVM + Isotonic	0.6984	**0.2145**	**0.0127**	0.215	0.25	0.301
SVM + SMOTE + Platt	0.6883	0.1207	0.0139	0.007	0.0853	0.167
SVM + SMOTE + Isotonic	0.6810	0.1155	0.0138	0.088	0.1428	0.165



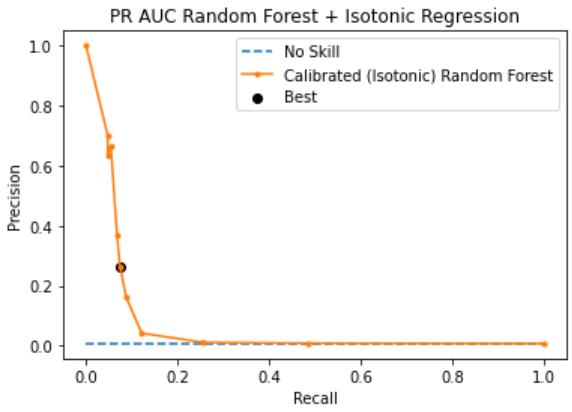


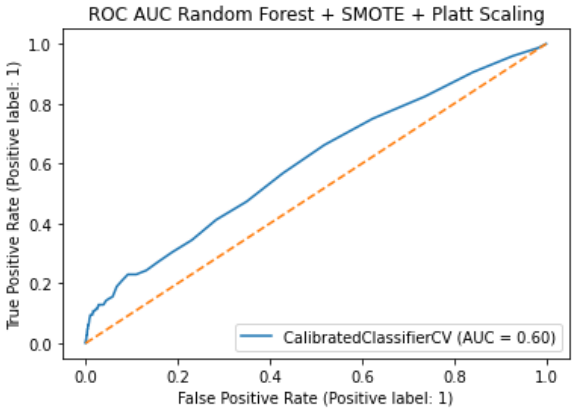
Dataset IV (imbalance proportion: 0.2%)

Random Forest

In the scenario of this highly imbalanced dataset, the combination of Random Forest with SMOTE used alone and in combination with Platt Scaling produces the highest ROC AUC. Similar to the cases presented above, the largest PR AUC was obtained by using Isotonic Regression alone. In this data, the best Brier Score, however, resulted from the application of Platt Scaling alone.

	ROC AUC	PR AUC	Brier Score	F-Score	Best Threshold	F-Score Best threshold
RF	0.5581	0.0676	0.0072	0.065	0.11	0.119
RF + SMOTE	**0.6026**	0.0186	0.0152	0.041	0.34	0.077
RF + Platt	0.5581	0.0676	**0.00712**	0.100	0.0518	0.119
RF + Isotonic	0.5565	**0.0699**	0.00715	0.088	0.1176	0.116
RF + SMOTE + Platt	**0.6026**	0.0186	0.00731	≈ 0	0.0274	0.077
RF + SMOTE + Isotonic	0.5677	0.0240	0.00732	0.088	0.0956	0.097

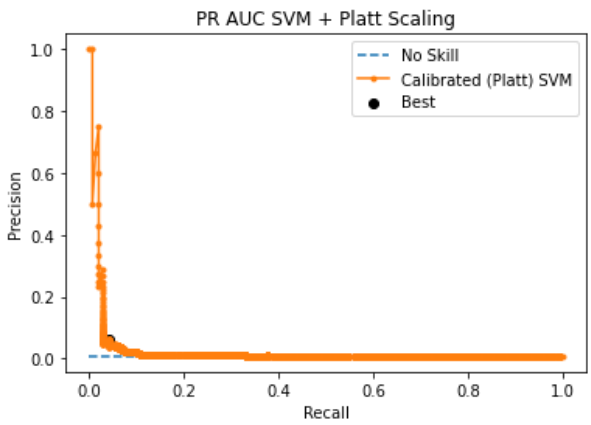


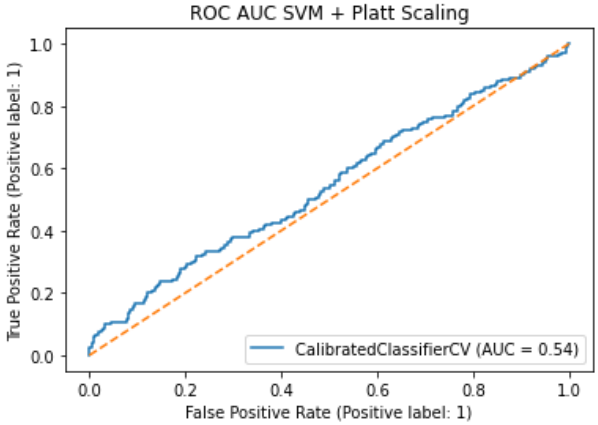


Support Vector Machine

The case in which SVM is used as classification model, the combination with Platt Scaling alone produces the best results overall. ROC AUC, PR AUC, and Brier Score are the best when this technique is applied.

	ROC AUC	PR AUC	Brier Score	F-Score	Best Threshold	F-Score Best Threshold
SVM	0.5444	0.0277	-	≈ 0	0.0166	0.050
SVM + SMOTE	0.5180	0.0155	-	0.018	0.0001	0.032
SVM + Platt	**0.5444**	**0.0277**	**0.00732**	≈ 0	0.0267	0.050
SVM + Isotonic	0.5297	0.0276	0.00736	0.39	0.043	0.047
SVM + SMOTE + Platt	0.5440	0.0104	0.007343	≈ 0	0.0124	0.033
SVM + SMOTE + Isotonic	0.5424	0.0098	0.007348	≈ 0	0.0195	0.027



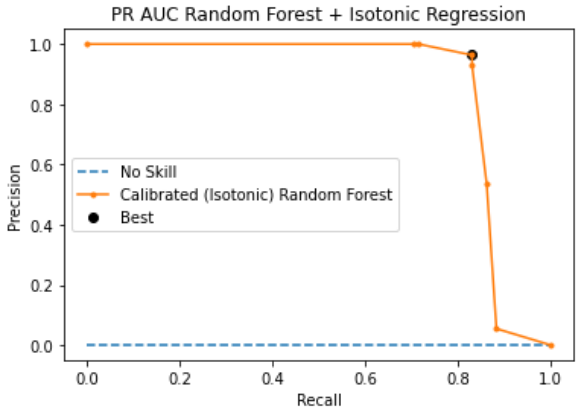


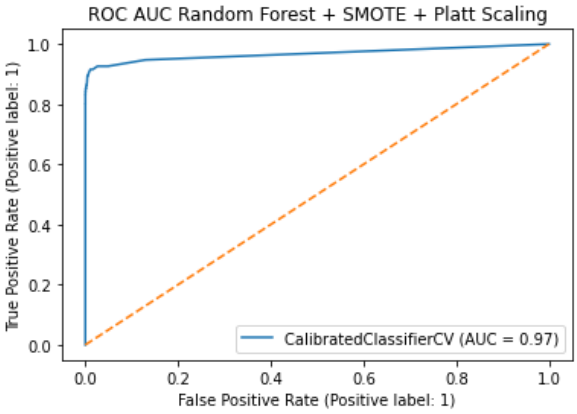
Dataset V (Real world - imbalance proportion: 0.17%)

Random Forest

When Random Forest is used, the best ROC AUC is achieved when SMOTE is used alone or in combination with Platt Scaling. Regarding the PR AUC, the best case again was the one in which Isotonic Regression was used alone. The latter also results in the best Brier Score.

	ROC AUC	PR AUC	Brier Score	F-Score	Best Threshold	F-Score Best threshold
RF	0.94	0.8596	0.00037	0.879	0.49	0.893
RF + SMOTE	**0.9678**	0.8614	0.00048	0.864	0.66	0.879
RF + Platt	0.9404	0.8596	0.00033	0.879	0.3412	0.893
RF + Isotonic	0.9403	**0.8621**	**0.00032**	0.893	0.75	0.893
RF + SMOTE + Platt	**0.9678**	0.86144	0.00036	0.879	0.5158	0.879
RF + SMOTE + Isotonic	0.9676	0.8626	0.00037	0.893	0.5	0.893

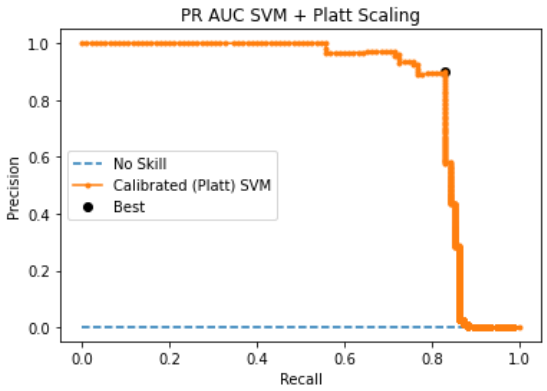


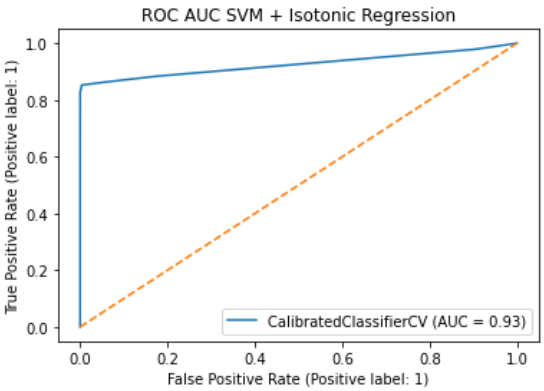


Support Vector Machine

In the case when SVM is used, the best ROC AUC and Brier Score are obtained when Isotonic Regression is used on its own. The best PR AUC, on the other hand, is generated when Platt Scaling is used alone.

	ROC AUC	PR AUC	Brier Score	F-Score	Best Threshold	F-Score Best Threshold
SVM	0.9087	0.8307	-	0.787	0.0034	0.863
SVM + SMOTE	0.9184	0.5901	-	0.152	0.9999	0.573
SVM + Platt	0.9087	**0.8307**	0.00048	0.819	0.0053	0.863
SVM + Isotonic	**0.9258**	0.8301	**0.00043**	0.821	0.0203	0.863
SVM + SMOTE + Platt	0.9176	0.5901	0.000988	0.511	0.2466	0.573
SVM + SMOTE + Isotonic	0.9257	0.5880	0.000980	0.488	0.3151	0.525


PR AUC SVM + Platt Scaling
Precision
Recall
No Skill
Calibrated (Platt) SVM
Best
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0


ROC AUC SVM + Isotonic Regression
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
CalibratedClassifierCV (AUC = 0.93)
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

Conclusion

This section will proceed with a general discussion in view of the results obtained above, as well as exposing the limitations embedded in this paper and possibilities for future research on the topics analysed throughout this study.

General Discussion

The task of classification under imbalanced data applied in credit card fraud detection framework was investigated on this research. The background related to the detection of frauds in credit card transactions was exposed, as well as previous approaches that were taken to tackle the issues that arise in classification tasks under imbalance, such as resampling techniques and probability calibration. Both approaches were analysed in this paper, by using SMOTE, Platt Scaling and Isotonic Regression. The machine learning models were chosen in accordance with previous studies conducted in the field. Support Vector Machine, as well as ensemble models, which includes the Random Forest used, were applied in multiple research projects on the area.

Simulations involving different class imbalances were used in this study so comparisons could be done regarding which combination of techniques would be more suitable for each case. The comparison of their results shows that datasets with different imbalance proportions present different combination of methods to reach the best performance.

The first dataset analysed was the one in which 10% of the transactions were labelled as frauds. On this dataset, the performance of the Random Forest classifier was generally superior to the SVM in any combination of techniques applied. It is worth noting that in both classification models, the use of Isotonic Regression alone produced the best PR AUC and Brier Scores, which suggests this technique is the most suitable to be used for this class imbalance.

On the second dataset analysed, in which the class imbalance was 5%, the Random Forest classifier also produced better results overall, when compared to SVM. In this case, the use of Isotonic Regression has also produced superior results in terms of PR AUC and Brier Score for both Random Forest and SVM. This also suggests that, under a class imbalance of 5%, using Isotonic Regression alone would be the best

course of action to tackle the issue of imbalance, under the framework considered in this paper.

For the generated dataset with 1% of transactions classified as frauds, Random Forest again showed overall better results than SVM. Similarly, the use of Isotonic Regression was the best choice to obtain good PR AUC and Brier Score measures, regardless of Random Forest or SVM being used.

The fourth dataset generated was highly imbalanced, with 0.2% of transactions being fraud. In this case, the Random Forest classifier has also been shown to generate better results than SVM overall. However, within each classification model, the best combination of techniques was different from the datasets with less class imbalance presented above. In this highly imbalanced scenario, Random Forest in combination with Isotonic Regression also produced the best result in terms of PR AUC, but the best Brier Score was obtained with Platt Scaling instead. In the case of SVM, however, the use of Platt Scaling alone produced the best results for all the performance measures considered.

The same combination of methods was applied to the real-world dataset considered in this paper. In this case, 0.17% of the transactions were classified as fraudulent. As in all the other cases examined, Random Forest presented an overall superior performance than SVM. When used, the combination of RF and Isotonic Regression produced the best PR AUC and Brier Scores. When SVM was used the Isotonic Regression produced the best ROC AUC and Brier Score, but the best PR AUC was obtained with Platt Scaling.

The results point that the use of Random Forest classifier is the most suitable in all the cases analysed. If the focus is on enhancing the performance of the model in classifying the minority class by using the PR AUC, as well as obtaining good probability calibration, then using Random Forest with Isotonic Regression is the best combination of methods to be used. If the focus is on enhancing primarily the ROC AUC and obtaining calibrated probabilities, however, the use of SMOTE and Platt Scaling is the most appropriate combination of techniques for obtained superior ROC AUC and Brier Scores for all the cases in which Random Forest was used.

In addition to that, the results also show that in most cases, it is better to consider changing the probability threshold used for the classification. Instead of the default of p = 0.5 used in the classification models, in most cases tweaking this threshold results in better precision-recall balance and higher F-Measure. It was observed that as the class imbalance proportion increases, the optimal decision threshold decreases. For instance, in the first dataset with 10% of imbalance proportion, the best threshold for the different scenarios is in the interval of p = [0.63, 0.65], whereas in the highly imbalanced case of the last dataset generated, p = [0.01, 0.04].

The report from the European Central Bank presented at the beginning of the paper showed that a very small proportion credit card transactions in the continent are fraud, representing less than 1% of overall transactions. That suggests the two generated datasets with high imbalance of 1% and 0.2% would match what is currently observed in the industry. Given that, and considering the pattern observed in the results obtained on these datasets, some advice and directions for researchers and practitioners aiming at implementing machine learning to detect fraud can be derived. Under the framework presented in this paper, Random Forest should be used instead of SVM, as the former showed superior performance than the latter in the datasets analysed. Furthermore, it is advisable to combine the use of Random Forest with Isotonic Regression for obtaining better calibrated probabilities and better PR AUC, as well as using the maximisation of the F-score to find the best probability threshold to use for the classification.

Limitation

One of the limitations of academic research on the context of credit card fraud detection is the scarcity of real-world datasets publicly available, which was mentioned in different sections of this paper. The few available datasets are often anonymized due to confidentiality issues, such as the one used in this research. The anonymous nature of the variables harms the interpretability of the model in case it is the interest to detect variable importance and how they affect the outcome.

Other limitation for this paper arises from the lack of interpretability from black box models, such as the Random Forest and Support Vector Machine used. Since the main objective was to detect a combination of models that would yield the best

performance under different imbalance scenarios, the interpretability was not addressed. Different approaches could be taken to make machine learning models more interpretable, both towards global and local interpretability.

Future Research

Credit card fraud detection is a field in constant change, which reflects not only the continuous emergence of different fraud methods created by fraudsters but also reflecting the advances in the way machine learning models are used and tuned. In that context, there is a vast room for exploration for future research on the area. First, the combination of other machine learning models, other resampling and probability calibration methods could be explored in the context of simulated datasets.

Furthermore, the exploration new methodologies for the simulation of data that are tailored for credit card fraud tasks is something that could bring great value for the academic research on this field, given the already mentioned scarcity of real-world data publicly available the field.

References

Adewumi, A. O., & Akinyelu, A. A. (2017). A survey of machine-learning and nature-inspired based credit card fraud detection techniques. *International Journal of System Assurance Engineering and Management*, *8*(2), 937-953.

Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, *6*(1), 20-29.

Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. *Decision support systems*, *50*(3), 602-613.

Boyd, K., Eng, K. H., & Page, C. D. (2013, September). Area under the precision-recall curve: point estimates and confidence intervals. In *Joint European conference on machine learning and knowledge discovery in databases* (pp. 451-466). Springer, Berlin, Heidelberg.

Breiman, Leo. "Random forests." *Machine learning* 45.1 (2001): 5-32.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, *78*(1), 1-3.

Chan, P. K., Fan, W., Prodromidis, A. L., & Stolfo, S. J. (1999). Distributed data mining in credit card fraud detection. *IEEE Intelligent Systems and Their Applications*, *14*(6), 67-74.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.

Chen, R., Chiu, M., Huang, Y. & Chen, L. (2004). Detecting Credit Card Fraud by Using Questionaire-Responded Transaction Model Based on Support Vector Machines. Proc. of IDEAL2004, 800-806.

Davis, J., & Goadrich, M. (2006, June). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning* (pp. 233-240).

Duman, E., Ekinci, Y., & Tanrıverdi, A. (2012). Comparing alternative classifiers for database marketing: The case of imbalanced datasets. *Expert Systems with Applications*, *39*(1), 48-53.

Elkan, C. (2001, August). The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence* (Vol. 17, No. 1, pp. 973-978). Lawrence Erlbaum Associates Ltd.

European Central Bank. 6th report on card fraud. August 2020. [Online; Last consulted 09-October-2020]. URL:

https://www.ecb.europa.eu/pub/cardfraud/html/ecb.cardfraudreport202008~521edb602b.en.html#toc2.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, *27*(8), 861-874.

Fotouhi, S., Asadi, S., & Kattan, M. W. (2019). A comprehensive data level analysis for cancer diagnosis on imbalanced data. *Journal of biomedical informatics*, *90*, 103089.

Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: adaptive synthetic sampling approach for imbalanced learning. In 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence), 1322–1328. IEEE.

Le Borgne, Y. A., & Bontempi, G. (2004). Machine learning for credit card fraud detection-practical handbook. *ACM SIGKDD explorations newsletter*, *6*(1), 1-6.

Maalouf, M., & Trafalis, T. B. (2011). Robust weighted kernel logistic regression in imbalanced and rare events data. *Computational Statistics & Data Analysis*, *55*(1), 168-183.

Maes, S., Tuyls, K., Vanschoenwinkel, B. & Manderick, B. (2002). Credit Card Fraud Detection using Bayesian and Neural Networks. Proc. of the 1st International NAISO Congress on Neuro Fuzzy Technologies.

Mani, I., & Zhang, I. (2003, August). kNN approach to unbalanced data distributions: a case study involving information extraction. In *Proceedings of workshop on learning from imbalanced datasets* (Vol. 126, pp. 1-7). ICML.

Mrozek, P., Panneerselvam, J., & Bagdasar, O. (2020, December). Efficient resampling for fraud detection during anonymised credit card transactions with unbalanced datasets. In *2020 IEEE/ACM 13th International Conference on Utility and Cloud Computing (UCC)* (pp. 426-433). IEEE.

Niculescu-Mizil, A., & Caruana, R. (2005, August). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning* (pp. 625-632).

Niculescu-Mizil, Alexandru, and Rich Caruana. "Obtaining Calibrated Probabilities from Boosting." In *UAI*, vol. 5, pp. 413-20. 2005.

Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119*.

Popat, R. R., & Chaudhary, J. (2018, May). A survey on credit card fraud detection using machine learning. In *2018 2nd international conference on trends in electronics and informatics (ICOEI)* (pp. 1120-1125). IEEE.

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, *10*(3), 61-74.

Priscilla, C., & Prabha, D. P. (2019, October). Credit card fraud detection: A systematic review. In *International Conference on Information, Communication and Computing Technology* (pp. 290-303). Springer, Cham.

Provost, F. (2000, July). Machine learning from imbalanced data sets 101. In *Proceedings of the AAAI'2000 workshop on imbalanced data sets* (Vol. 68, No. 2000, pp. 1-3). AAAI Press.

Robertson, T., Wright, F., & Dykstra, R. (1988). Order restricted statistical inference. New York: John Wiley and Sons.

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, *10*(3), e0118432.

Sisodia, D. S., Reddy, N. K., & Bhandari, S. (2017, September). Performance evaluation of class balancing techniques for credit card fraud detection. In *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)* (pp. 2747-2752). IEEE.

Tharwat, A. (2020). Classification assessment methods. *Applied Computing and Informatics*.

Tomek, I. (1976). Two modifications of CNN. *IEEE Trans. Systems, Man and Cybernetics*, *6*, 769-772.

Van Vlasselaer, Véronique, Cristián Bravo, Olivier Caelen, Tina Eliassi-Rad, Leman Akoglu, Monique Snoeck, and Bart Baesens. "APATE: A novel approach for automated credit card transaction fraud detection using network-based extensions." *Decision Support Systems* 75 (2015): 38-48.

Wallace, B. C., & Dahabreh, I. J. (2012, December). Class probability estimates are unreliable for imbalanced data (and how to fix them). In *2012 IEEE 12th international conference on data mining* (pp. 695-704). IEEE.

Wallace, Byron C., and Issa J. Dahabreh. "Improving class probability estimates for imbalanced data." *Knowledge and information systems* 41, no. 1 (2014): 33-52.

Wilson, D. L. (1972). Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics*, (3), 408-421.

Yen, S. J., & Lee, Y. S. (2009). Cluster-based under-sampling approaches for imbalanced data distributions. Expert Systems with Applications, 36(3), 5718-5727.

Zadrozny, B., & Elkan, C. (2002, July). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 694-699).

Sun, Y., Wong, A. K., & Kamel, M. S. (2009). Classification of imbalanced data: A review. *International journal of pattern recognition and artificial intelligence*, *23*(04), 687-719.