The use of machine learning algorithms in CCFD

Abstract

According to (Experian, 2022) by the end of the last quarter of 2022, credit card fraud rose by 18%, and the card fraud rate for 2022 was 0.65% of the sample of 10000 people that has been taken. Mainly there are two transaction fraud scenarios, Card-present (CP) and Card-not-present (CNP), according to (European Central Bank, 2020) 79% of the fraud detected in 2018 within SEPA was CNP, that's because of the development of EMV technology, while CP fraud detected transaction was categorized to three types: lost card, counterfeited card, and card not received as per (European Central Bank, 2020).

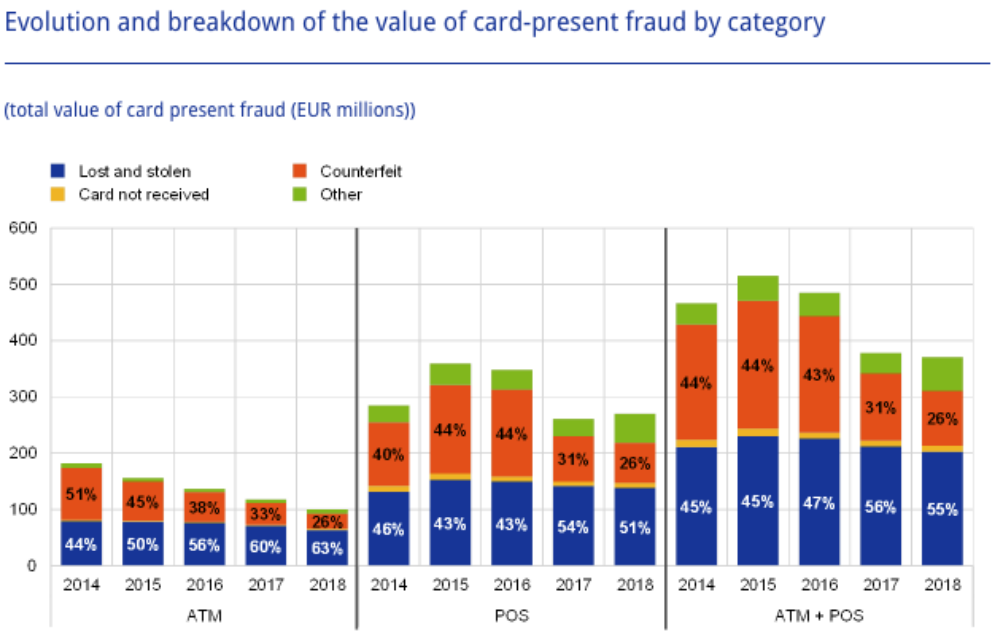


Fig. 1. Evolution and breakdown of the value of card-present fraud by category within SEPA

Using a machine learning algorithm finds patterns of behavior that can be detected through a training dataset that is supplied to the algorithm to build a model to get the possible fraud transaction, this essay implements three machine learning algorithms and demonstrate the effect of feature engineering techniques used to prepare the data after the exploratory data analysis (EDA) stage, more specifically the use of sampling, transformation and cost sensitivity techniques. The three learning algorithms selected to be applied in the Finance start-up are Random Forests, Logistic Regression, and Naïve Bayes.

1- Introduction

To have an effective fraud detecting system, it must have high accuracy and a high true negative rate, unlike a lot of other applications, the cost of wrong detection of a fraud transaction is high. For example, if a fraudster did a transaction with 10000 USD, and the ML system failed to detect it as fraud, that means the victim lost that money, and as the victim's numbers increase the bank will also start to face financial problems and accordingly face legal problems.

The selection of Random Forests is because it is considered to be one of the best algorithms to be used in classification problems because of its high performance and resistance to overfitting (Breiman, 2001), on the other hand, Naïve Bayes and logistic regression were selected because among several other classification techniques I tried before starting the experiment, they were the highest in performance among them, nevertheless, others algorithms performed well enough, but I wanted to study a

probabilistic machine learning algorithm along with performance proved Ensemble learning method to get the best results to provide to the start-up.

2- Dataset and analysis

The data used in this study is publicly available data on Kaggle (kaggle.com, 2022), it is synthetically generated via a deep learning model that was trained on credit card fraud detection, after the first look at the data in the EDA stage, the dataset consisted of 219129 instances with 32 attributes labeled as 'ID', 'Time', 'V1', …., 'V28', 'Amount', 'Class'. The dataset attributes were all numeric attributes, in addition, it was noticed that the target class was highly imbalanced 218660: 469.

Firstly, the ID attribute was removed as it is a numeric ordering for the instances, the attribute was irrelevant to the data and the scheme, accordingly the ID attribute was deleted from the dataset. as a classification problem the target class had to be nominal class, consequently, a single feature transformation had to be done to transform the target class from numeric to nominal value.

As mentioned above the data in the dataset was highly imbalanced, and supplying such a dataset as a training set to the algorithm will have a serious overfitting problem, in addition, unrealistic performance (Oded Maimon and Lior Rokach, 2010), in this essay Synthetic Minority Oversampling Technique (SMOTE) will be used to generate Synthetic values from the minority target value, as suggested by (Chawla et al., 2002), in order to increase the values of the minority class (oversampling the minority class) K nearest neighbors are randomly being used so that you can select that K number, the K number was selected by the researchers to be 5, a random

number between 2 and 5 then their values are being used to generate oversampling as follows: the difference between the feature vector and its nearest neighbor, then multiply this difference by random number ranges from 0 to 1 and add it to feature vector.

3- Method

The goal is to detect fraud within a set of unseen data and compare the efficiency of the three learning algorithms to determine which one to be used in production by the Start-up. the software used in the process is WEKA tool, it is a free software programmed by Waikato University team (Waikato University, 2019) that is available online for downloading, the tool contains a huge number of machine learning and filtration algorithms.

After the data is loaded into the system transformation filter is used to transform the values of the target class from numeric to nominal values, then ID attribute is removed from the dataset as it is irrelevant to the data and the scheme because it is just an ordering numeric value, at this point, the only remaining preprocessing problem is the high imbalance that will be solved using SMOTE to increase the number of the minority class, the minority class was increased to 30016 instances to have a ratio of nearly 9: 1 majority to minority respectively, now the data is ready for machine learning algorithms to be applied.

The data was split into 80: 20 training to testing sets, the 80% will be used for training, and 20% unseen data will be used for evaluating test. The model will be trained using stratified cross-validation and held-out, cross-validation is a technique in which the training dataset is split into N number of sets N-1 is being used for training and the

remaining set is used to validate the model, this process is repeated for N number of times, this option is available on WEKA tool in classify tab.

Random Forests (RF) is an ensemble learning algorithm that uses decision tree algorithms (Breiman, 2001), decision tree algorithm selects the best attribute according to information gain, make it as the root node for the tree then starts branching there (Bell, 2020), accordingly the RF tree uses the same technique, but instead of selecting only one attribute it selects an X number of best attributes, and randomly select one of them as a node for the tree then branch it, and do the same for the next node .. etc. in fig. 2. The parameters selected for the RF algorithms are shown, and the number of selected features is 5, therefore the RF algorithm will select 5 features among the best to randomly select 1 of them as the root node.

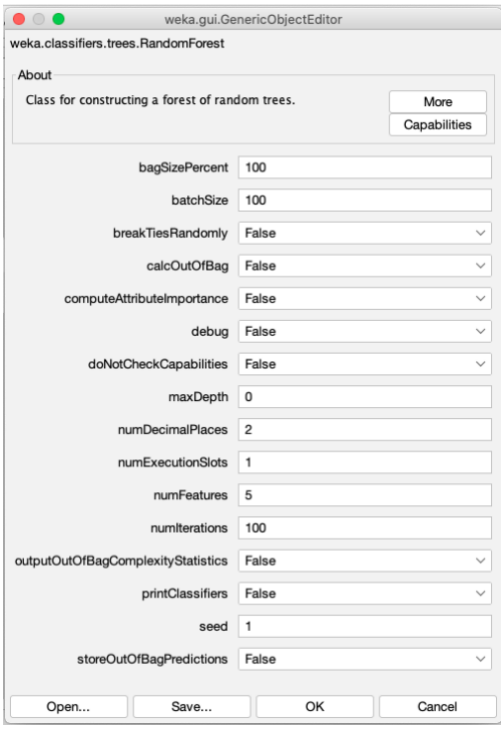


Fig. 2. RF Parameters used on WEKA tool

Whereas Naïve Bayes (NB) is a scheme-independent algorithm that uses a probability formula to reduce the number of attributes that is redundant to the dataset, therefore, the complexity of the problem as per equation 1 (Xue and Titterington, 2008).

$$Y \leftarrow \text{argmax}\ yk\ P(Y = yk) \prod i\ P(Xi\ |Y = yk) \qquad (1)$$

Moreover, Logistic regression (LogR), the third algorithm, implements a logit transform equation 2. Which choose weights to maximize the log-likelihood (Daniel and Martin, 2023).

$$Pr[1|\ a_1, a_2, \ldots, a_k] = 1/(1+\exp(-w_0 - w_0a_1 - \ldots - w_ka_k) \qquad (2)$$

Additionally, 1 more method has been used, a cost-sensitive method, as mentioned above the cost of miss classifying a fraud transaction is very high, therefore more weight is given to the false positive (Elkan, 2001) C_{10} in fig. 3.

	Y_0	Y_1
$\hat{Y}_0$	c_{00}	c_{01}
$\hat{Y}_1$	c_{10}	c_{11}

Fig. 3. Confusion matrix

4- Evaluating Measures

Confusion matrix: is a matrix where all the values are being put into the same as fig. 3.

Precision: $\frac{true\ positive\ \ (C00)}{true\ positive\ (C00)+false\ positive\ (C10)}$

Recall: $\frac{true\ positive\ \ (C00)}{true\ positive\ (C00)+false\ negative\ (C01)}$

F1: $\frac{2*precision*recall}{precision+recall}$

AUC ROC: the area under the curve of recall vs false positive rate.

5- Results & Discussion

In this section I will discuss in detail the evaluation of the models, the model was slow while training on WEKA, as it took around two hours to train the algorithm and generate the model, On the contrary, Naïve Bayes took 5 minutes and logistic regression took twenty minutes, among the learning algorithms as mentioned above Random Forests considered to be one of the best algorithms in performance yet the slowest to be trained, table 1 below shows the evaluation metrics of the machine learning algorithms.

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
RF	0.994	0.041	0.994	0.994	0.994	0.970	0.999	0.999
LogR	0.909	0.560	0.898	0.909	0.894	0.483	0.884	0.934
LogR with cost	0.888	0.279	0.904	0.888	0.895	0.544	0.889	0.934
NB	0.885	0.537	0.873	0.885	0.877	0.396	0.799	0.894

NB with cost	0.883	0.510	0.874	0.883	0.878	0.407	0.799	0.894

Table 1. Evaluation measures for the training data

As appears in the table the precision of RF is 0.994 and the AUC ROC is 0.999 compared to 0.898 in LogR and 0.873 in NB, the Metrics for RF seem high, but as per the procedure there is 20% of the dataset held-out as an unseen dataset to finally evaluate the model before giving it to production, the unseen data was loaded to WEKA tool as test set and each model was evaluated on the unseen data again, the table below shows the metrics of evaluation.

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area
RF	0.994	0.041	0.994	0.994	0.994	0.972	0.999
LogR	0.907	0.571	0.896	0.907	0.893	0.469	0.880
LogR with cost	0.888	0.280	0.905	0.888	0.894	0.541	0.885
NB	0.885	0.543	0.873	0.885	0.877	0.391	0.792
NB with cost	0.883	0.518	0.874	0.883	0.878	0.401	0.792

Table 2. Evaluation measures for the unseen test data

In Appendix you will find the ROC curves for the trained data and the screenshots from the unseen data evaluation.

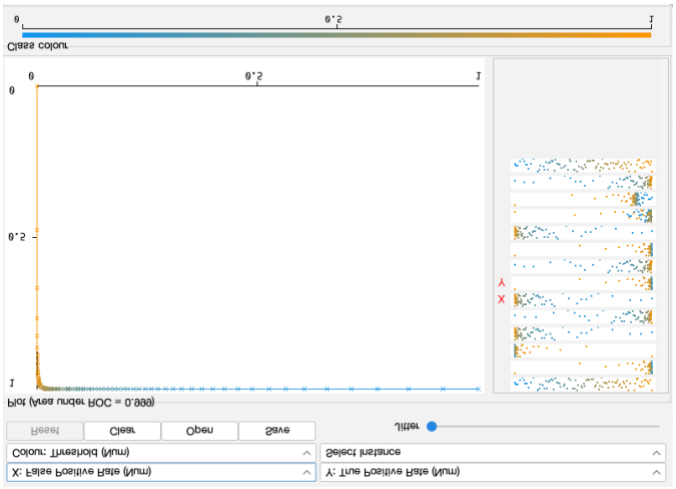
6- Conclusion

Among the three learning algorithms, RF was the highest precision, and AUC ROC and Naïve Bayes was the lowest ones even after applying the cost-sensitivity technique and slightly getting worse after applying it. Random Forests appeared to be the best solution for this start-up to protect them from the CCFD, now it is safe for this financial start-up to open a division that serves the credit card holders.

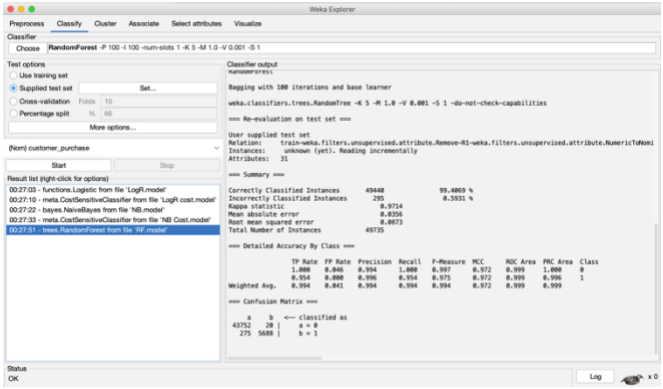
7- Appendix

a. Random Forests

i. AUC ROC for the training set

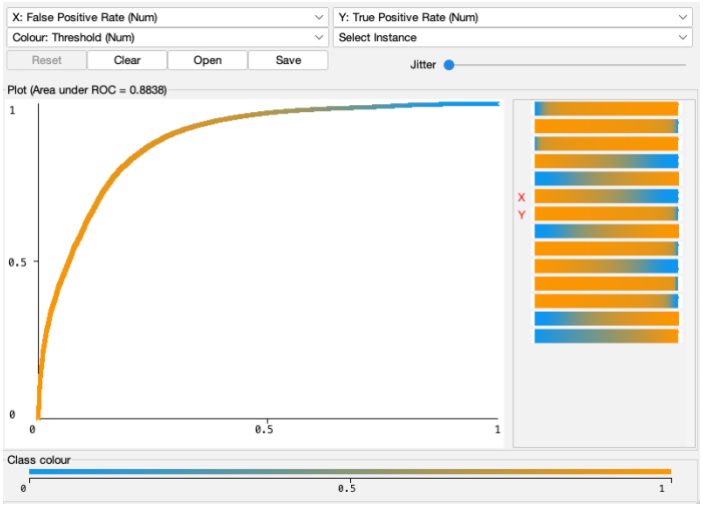


ii. Unseen data results

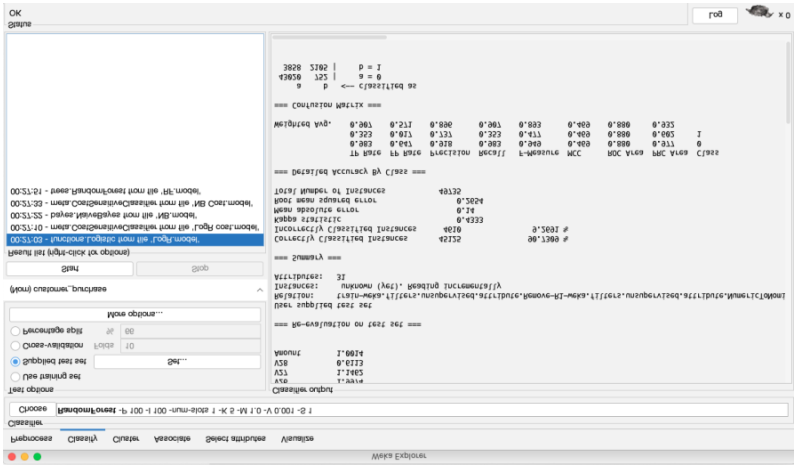


b. Logistic Regression

i. AUC ROC

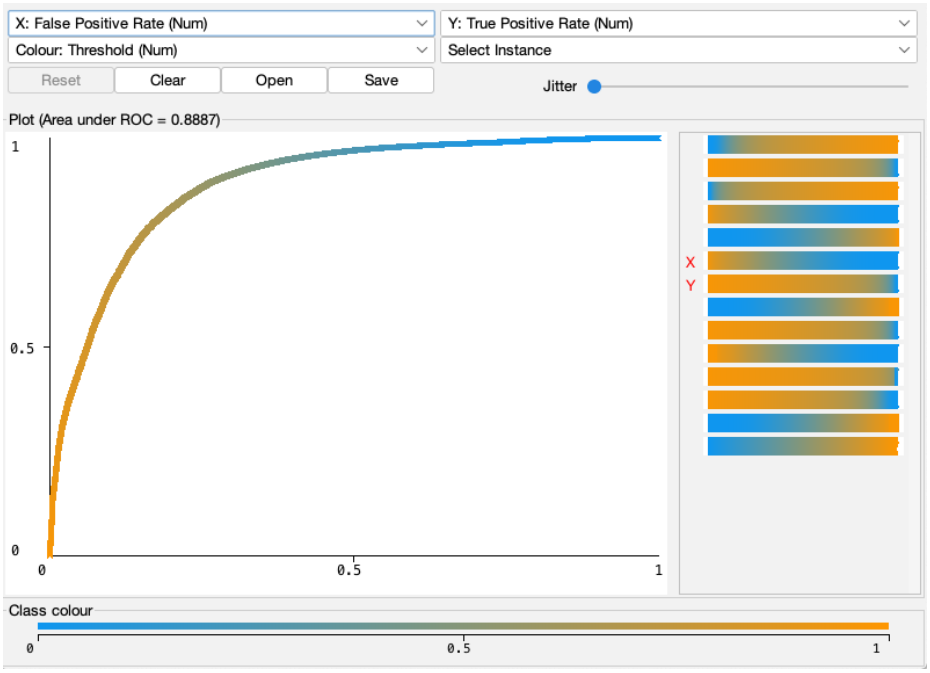


ii. Unseen data results

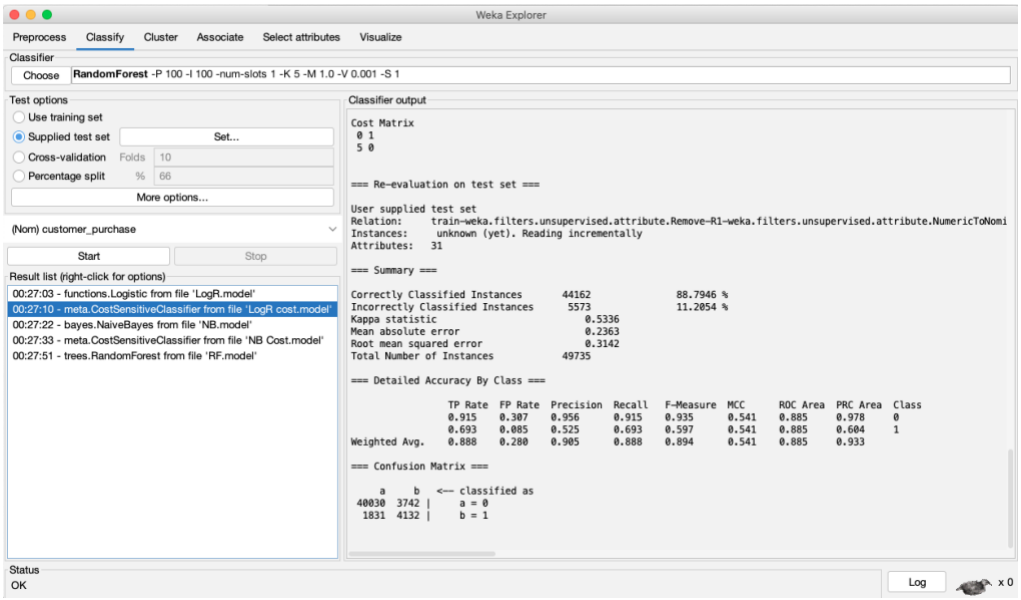


c. Logistic Regression with cost sensitivity

i. AUC ROC

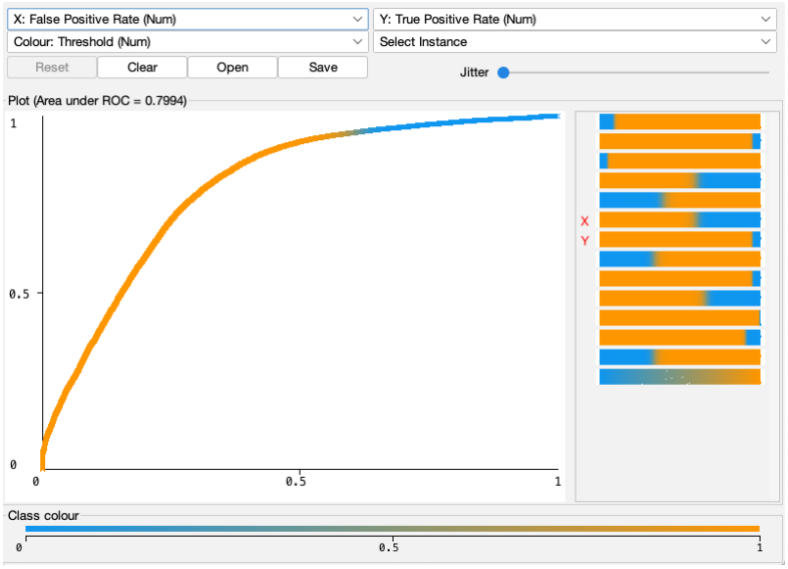


ii. Unseen data results

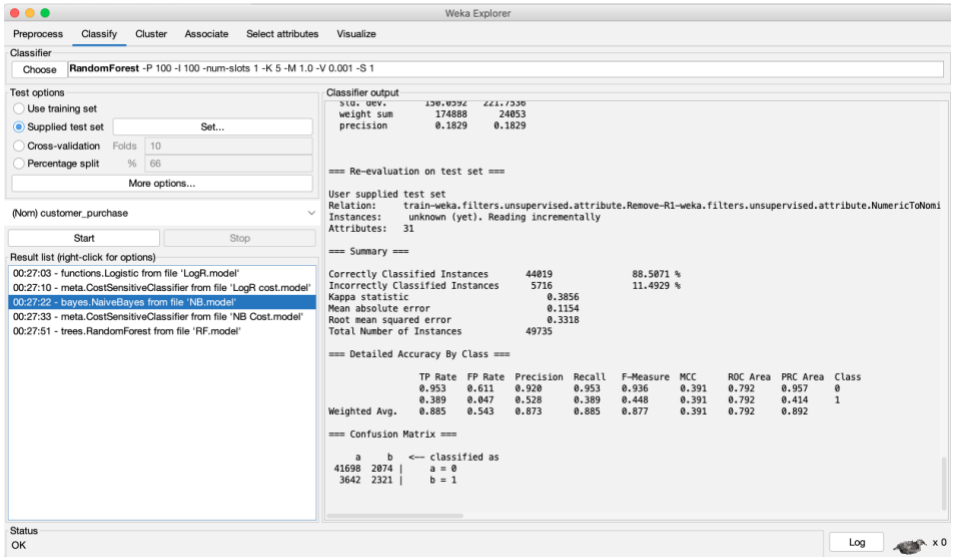


d. Naïve Bayes

i. AUC ROC

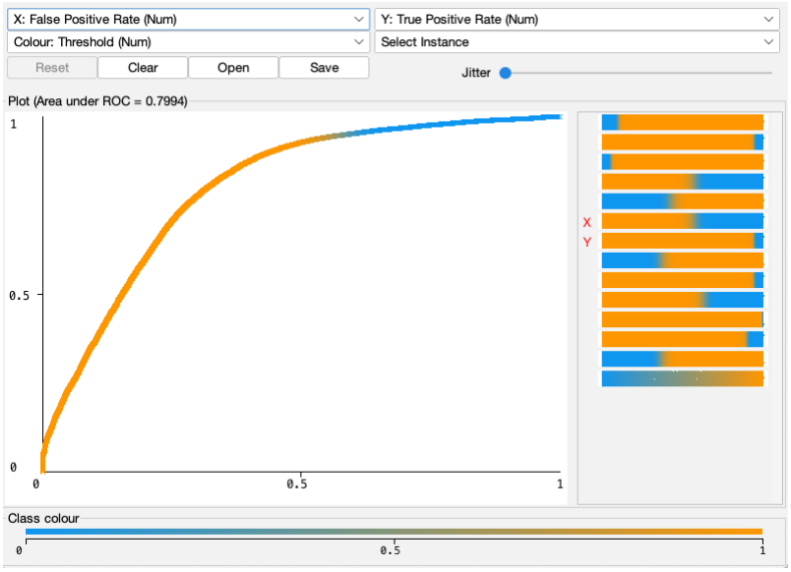


ii. Unseen data results

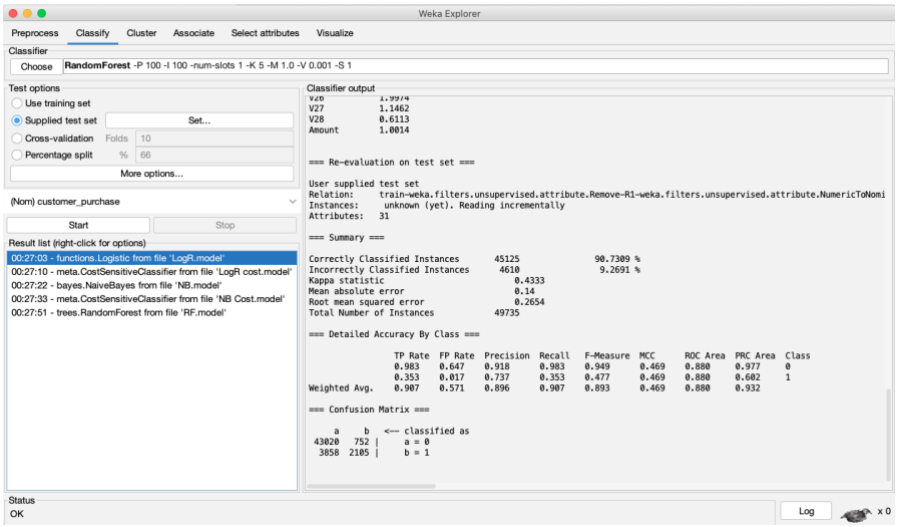


e. Naïve Bayes with cost sensitivity

i. AUC ROC



ii. Unseen data results



Reference list

Bell, J. (2020). *Machine Learning*. John Wiley & Sons.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), pp.5–32. doi:https://doi.org/10.1023/a:1010933404324.

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(16), pp.321–357. doi:https://doi.org/10.1613/jair.953.

Daniel, J. & Martin, J. (2023). *Speech and Language Processing*. [online] Available at: https://web.stanford.edu/~jurafsky/slp3/5.pdf.

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 (2020). Sixth report on card fraud. *www.ecb.europa.eu*. [online] Available at: https://www.ecb.europa.eu/pub/cardfraud/html/ecb.cardfraudreport202008~521edb602b.en.html#toc2 [Accessed 16 Apr. 2023].

Experian (2022). *Experian plc - Credit card fraud soars to 10-year high*. [online] www.experianplc.com. Available at: https://www.experianplc.com/media/latest-news/2023/credit-card-fraud-soars-to-10-year-high/ [Accessed 16 Apr. 2023].

kaggle.com. (2022). *Binary Classification with a Tabular Credit Card Fraud Dataset*. [online] Available at: https://www.kaggle.com/competitions/playground-series-s3e4/data [Accessed 16 Apr. 2023].

Oded Maimon & Lior Rokach (2010). *Data mining and knowledge discovery handbook*. New York: Springer.

Waikato University (2019). *Weka 3 - Data Mining with Open Source Machine Learning Software in Java*. [online] Waikato.ac.nz. Available at: https://www.cs.waikato.ac.nz/ml/weka/.

Xue, J.-H. & Titterington, D.M. (2008). Comment on ‘On Discriminative vs. Generative Classifiers: A Comparison of Logistic Regression and Naive Bayes’. *Neural Processing Letters*, 28(3), pp.169–187. doi:https://doi.org/10.1007/s11063-008-9088-7.