

CET023 - AI and Analytics in Finance,

Credit and Related Risks



**Group Project Report --Team 1**

**Credit Card Fraud Prediction**

Group Members:

| Name | Matriculation Number |
| --- | --- |
| Balasubramanian Narasimhan | K2120089G |
| Chan Jing Wen | K2220994A |
| Gerald Sng | K2220995J |
| Tee Wen Shan | K2220993D |

# TABLE OF CONTENTS

[**Executive Summary**](#_smol1u4z7wi8) **3**

[**Introduction**](#_10mcot34rmts) **4**

[1.1 Problem Statement](#_1fob9te) 4

[1.2 Literature Review](#_3znysh7) 4

[1.3 Key Delivery](#_2et92p0) 5

[**Data description & processing**](#_yspa94k93qfi) **5**

[2.1 Overall Description](#_3dy6vkm) 5

[2.2 Data Exploration](#_t3vbc4z5elub) 6

[2.5 Data Pre-processing](#_17dp8vu) 9

[**Model Construction and comparison**](#_1ksv4uv) **9**

[3.1 Model Training](#_44sinio) 9

[3.2 Model Selection](#_qbztss88ptfm) 9

[Logistic Regression](#_80bnblupngw4) 10

[Decision Tree](#_2jxsxqh) 10

[Random Forest](#_aifksg9t3f9j) 10

[XGBoost](#_rtfo9lfl7pzt) 10

[Neural Network](#_1xqpwgkmamzp) 11

[3.3 Model Performance](#_nzgyp67yulvd) 11

[3.4 Feature Importance](#_4jf0nlv489u2) 12

[**Conclusion and Future studies**](#_1ci93xb) **13**

# 

# 

# **Executive Summary**

In this project, we aim to, with the use of machine learning algorithms, come up with a model which determines if a credit card transaction is a potential fraud case. The dataset we used consists of features such as Distance from Home, Distance from Last transaction, Ratio to median purchase price, Repeat retailer, etc. With these features, we seek to develop further understanding on how each variable affects the output with the use of 5 machine learning models: Logistics Regression, Decision Tree, Random Forest, XGBoost and Neural Network.

In data cleaning, we checked that there is no missing or null, or duplicated data in our dataset. Through data visualization, we found out that certain variables of our dataset are highly skewed. Hence, before we proceed to use the train test split, to train each model and apply on the test set, we transform the skewed data using log transformation to approximately conform to normality. In our analysis, we have also used both data with and without pre-processing to compare the accuracies.

For the classification problem, we applied the 5 machine learning models mentioned and compared their performance in terms of accuracy, Area under ROC curve, precision and recall. From the analysis, we found that Decision Tree has the best overall performance. Based on the result of Decision Tree we look at the feature importances, and observe that ratio-to-median purchase price is the most important feature in predicting credit card fraud.

# Introduction

## **1.1 Problem Statement**

Advancement in technologies and rise in e-payments have brought about opportunities for the credit card industry. With financial institutions offering more and more flexible payment options, rewards and customised benefits, etc., consumers are attracted to switch to card-based payments. According to the Credit Card Global Market Report 2022, the global credit card payment market is expected to grow from US$477.63 billion in 2021, to US$762.16 in 2027[[1]](#footnote-0).

The credit card industry is growing rapidly, so is the number of card fraud cases happening around the world. Cyber criminals are being increasingly sophisticated in their ways of obtaining personal information illegally and conducting fraudulent activities. As per the annual Nilson Report 2021, it is estimated that totalling US$408.50 billion will be lost to card frauds in the next 10 years.[[2]](#footnote-1) With the increasing prevalence, it is important that the financial institutions put in place effective fraud detection systems, to identify suspicious transactions and reduce financial losses to the company with early detection.

## **1.2 Literature Review**

In the increasingly digitalised world, it makes it easier for cyber criminals to conduct fraud. Methods such as phishing and identity theft have allowed them to access information even if they are in another geographical location, without gaining access to physical cards. In fact, credit card fraud is the top risk of identity theft as of 2022, reported by the Federal Trade Commission[[3]](#footnote-2).

Financial institutions need to continuously upgrade their systems as new fraud patterns emerge. Machine learning models have been increasingly utilised to tackle such risks nowadays. Before this, rule-based systems are commonly used. Like its name, rule-based systems pick up suspicious transactions with pre-programmed rules. However, rule-based models have limitations as they require significant manual intervention to identify the possible factors for rules creation. In addition, they are less scalable compared to machine learning models[[4]](#footnote-3).

On the other hand, machine learning models are able to process large datasets with minimal human intervention required. They have higher accuracy and support real time processing, which both are key to effective fraud detection.

## **1.3 Key Delivery**

Our study focuses on the application of the following supervised machine learning algorithms for credit card fraud detection, which include Logistic regression, Decision tree, Random forest and XGboost. Using real-world data, we aim to generate insights of Credit card Fraud by analyzing the past transaction details of the customers and extract certain behavioral patterns that could potentially help to flag out fraudulent transactions.

# Data description & processing

## **2.1 Overall Description**

The data was sourced from Kaggle[[5]](#footnote-4),Credit Card Fraud. The dataset consists of 1 million data points with the following attributes:

| Attributes | Description of the attributes |
| --- | --- |
| Distance from home  *(x-variable)* | The location of the identified transaction done with respect to location of the the cardholder’s home |
| Distance from last transaction  *(x-variable)* | The location of the identified transaction done with respect to the location of the last transaction done. |
| Ratio to median purchase price  *(x-variable)* | Ratio of the spending amount of identified transaction to median spending amount |
| Repeat retailer  *(x-variable)* | Is the identified transaction done with the same retailer? |
| Used chip  *(x-variable)* | Is the identified transaction done with a credit card? |
| Used PIN number  *(x-variable)* | Is the identified transaction done using the PIN of the credit card? |
| Online Order  *(x-variable)* | Is the identified transaction done online? |
| Fraud  *(Y-variable)* | Is the identified transaction fraudulent or legitimate?  **0 == Non-Fraudulent transaction**  **1 == Fraudulent transaction** |

## **2.2 Data Exploration**

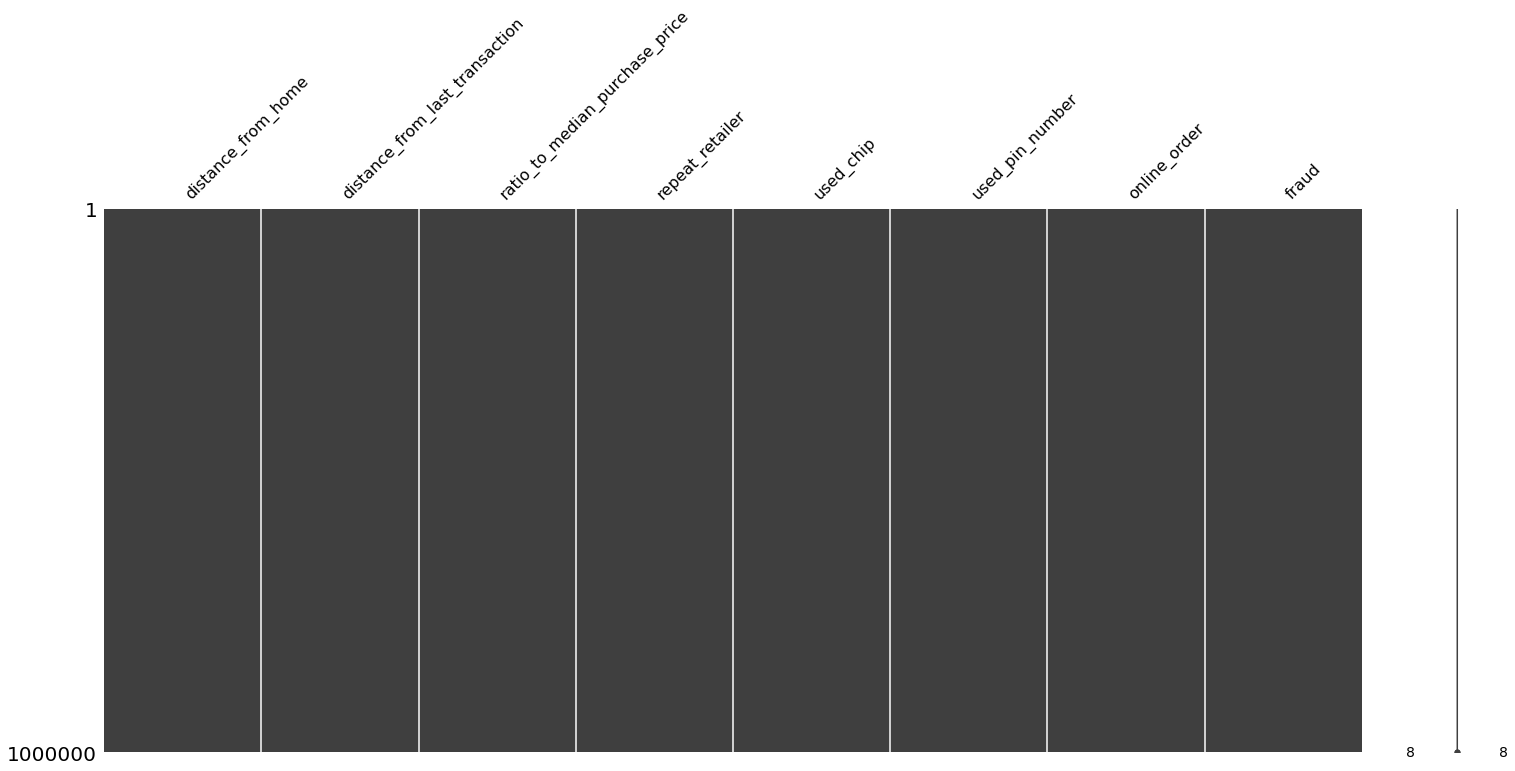


Figure 1

Before exploring the dataset in depth, we ensured that there are no missing or null values present in the dataset. It is important any missing data are identified and handled appropriately prior to further data analysis as they could potentially affect the accuracy of our prediction using Machine Learning algorithms.To better visualize the number of null values for each column, we used barplot (Figure 1) - a simple plot with each bar representing a column within the dataframe. The height of the bar indicates how complete a particular column is i.e. how many non-null values are present. From the diagram, we noted that the raw data obtained from the kaggle database is complete with no missing values. Hence, no data imputation is required.

We also identify that there is no duplicated data, as *df.duplicated().sum()* returns 0. Hence, data deduplication is not required.

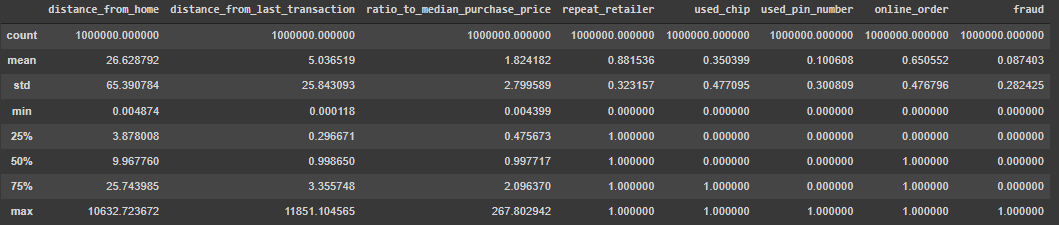
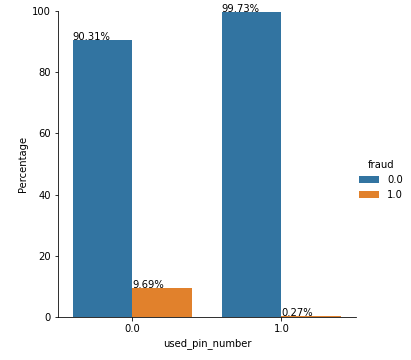
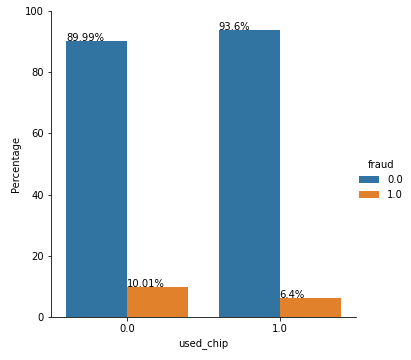


Figure 2

Next, we generated the descriptive statistics of the dataset to provide us some insights on the central tendency, dispersion and shape of the dataset’s distribution. As seen in Figure 2, 3 of the continuous variables: “distance\_from\_home”, “distance\_from\_last\_transaction” and “ratio\_to\_medium\_purchase\_price” are highly skewed, To verify our findings, we calculated the respective skewness values:

|  | distance\_from\_home | distance\_from\_last\_transaction | ratio\_to\_median\_purchase\_price |
| --- | --- | --- | --- |
| **Skewness** | 20.2 | 125.9 | 8.9 |

From the skewness values, we are able to confirm that the 3 variables above are highly skewed since the skewness should be about zero for normally distributed data. Normal distribution is important in regression models.



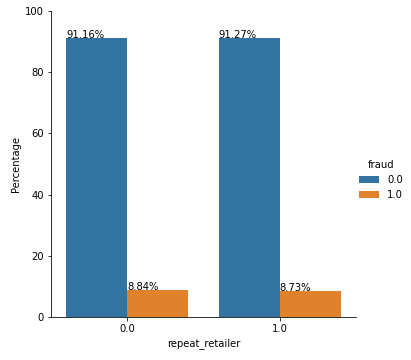
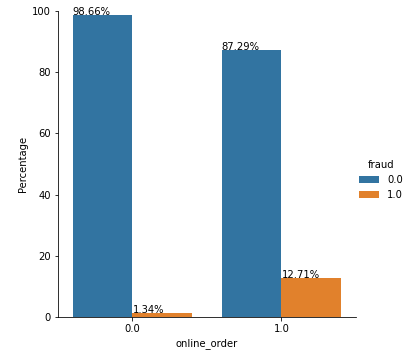


Figure 3

For the other 4 categorical variables: “Used chip”, “used pin number”. “Online order” and “repeat retailer”, we plotted bar graphs to show the relationship between “fraud” and each of the variables. From the bar graphs, we noted that fraudulent transactions are more likely to happen when a card is not used with a chip or a pin number. We also noted that fraudulent transactions are more likely to happen when transactions are done online. It seems like the probability of fraudulent transactions happening are equal for both repeat or non-repeat retailers.

**2.3 Feature Correlation**

## 

Figure 4

In view of optimizing our prediction performance, we look to eliminate any redundant or irrelevant data. A heatmap is used to help us gain a more intuitive understanding of the correlation between the variables. From the heatmap, we see that the feature‘ratio\_to\_median\_purchase\_price’ has the strongest correlation to fraud. There are no features that exhibit collinearity. We drop the 2 features with the lowest correlation, repeat\_retailer and used chip, both correlation significantly lower than 0.1 (0.001357 and 0.060975 respectively). Then, we used the remaining 5 variables with relatively higher correlation as the X-variables for our model training.

## **2.5 Data Pre-processing**

Before splitting the data into training data and test data, we transform the skewed data using log transformation to approximately conform to normality[[6]](#footnote-5).

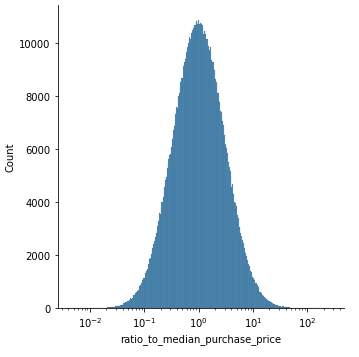
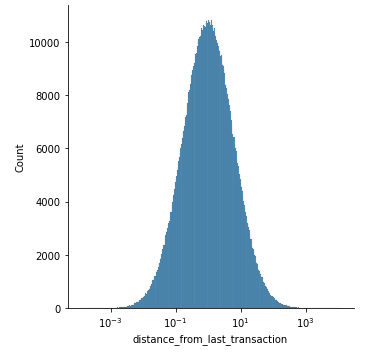
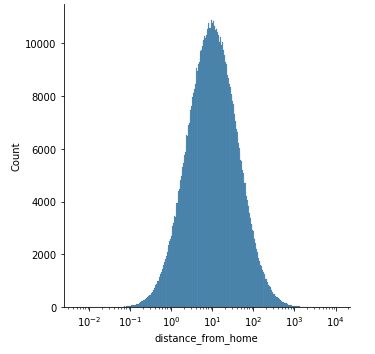


Figure 5

As seen in Figure 5, the distribution is more Gaussian-like after using a logarithmic scale on the x-axis. The skewness is seen to be reduced in the transformed data.

# Model Construction and comparison

## **3.1 Model Training**

After we have pre-processed the data and selected the models, we intended to run a grid search for hyperparameter tuning to improve model performance. However, we noted that if we were to predict every transaction as non-fraudulent, the accuracy score for the prediction is already as high as 0.913048. Hence, performing a hyperparameter tuning will not significantly improve the model performance further. Nonetheless, we will still use the dataset and select the most appropriate model by evaluating its accuracy, errors and also feature importance for further analysis and recommendations.

## **3.2 Model Selection**

Our target variable “Fraud” is a categorical variable with two possible values:

0 - Non-Fraudulent Transaction

1 - Fraudulent Transaction

With such binary features, we aim to classify our observations into 2 categories - customers who performed ‘fraudulent’ or ‘non-fraudulent’ transactions using the labeled dataset. In our study, we will assess which of the Machine Learning Algorithms will best classify the customer’s transactions into the ‘fraudulent’ and ‘non-fraudulent’ categories.

## **Logistic Regression**

Logistics Regression assumes linearity between the independent and dependent variables. We proceeded with logistic regression on both the preprocessed labels and one that is not preprocessed for high skewness in the numerical labels. It seems like the preprocessing did not help on model accuracy, AUROC, precision and recall. Thus we will not be using the preprocessed data for the training of other models.

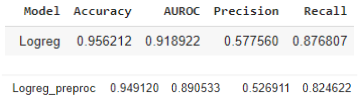


Figure 6

## **Decision Tree**

Decision tree provides different possible outcomes or consequences based on the related attributes in each branch. Relative to logistic regression, decision tree is able to capture non-linear relationships, and works with both continuous and categorical features.

## **Random Forest**

Random Forest makes use of multiple decision trees with each trained through bagging or boosting aggregating. Random Forest establishes the outcome based on the predictions of the multiple decision trees which generally improves the overall result.

## **XGBoost**

[XGBoost](https://xgboost.ai/) is a decision-tree-based ensemble Machine Learning algorithm that uses a [gradient boosting](https://en.wikipedia.org/wiki/Gradient_boosting) framework. It optimized the gradient boosting algorithm through parallel processing, tree-pruning, handling missing values and regularization to avoid overfitting or bias.

## Neural Network

Artificial neural networks are forecasting methods that are based on simple mathematical models of the brain. They allow complex nonlinear relationships between the response variable and its predictors.

## 3.3 Model Performance

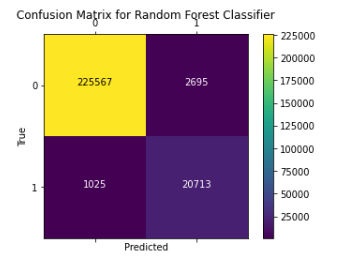
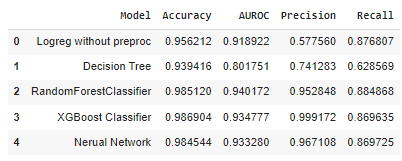
****

Figure 7

When comparing the performance of the various models, we take into consideration the following 4 metrics: Accuracy, Area under ROC Curve (AUC), Precision and Recall.

In terms of accuracy. XGBoost gives the best performance while Decision Tree gives the worst performance.

The Receiver Operating Characteristic (ROC) Curve plots 2 parameters: True Positive Rate against False Positive rate.[[7]](#footnote-6) The Area under ROC Curve (AUC) is a metric that measures the probability that an observation with a positive class will have a greater probability than an observation in a negative class. If AUC =1, it means there is perfect prediction by the model. If AUC = 0.5, it means the model is unable to discriminate between the classes.[[8]](#footnote-7) Hence, in terms of AUC, Random forest gives the best performance while Logistic Regression gives the worst performance.

Precision-Recall is a useful measure of success of prediction when the classes are very imbalanced.[[9]](#footnote-8) Precision quantifies the number of correct positive predictions made, with 1.0 for perfect precision.[[10]](#footnote-9) Recall quantifies the number of correct positive predictions made out of all positive predictions. A high precision relates to a low false positive rate, while a high recall relates to a low false negative rate. The precision rate and recall rate are therefore related to two types of error, which are costly and can have serious implications: (1) False positive - we wrongly predict fraudulent transactions which could result in terminating business relationships with customers and (2) False Negative - we fail to detect fraudulent transactions and put the institution at risk of facilitating bad actors to carry out crimes. Among the 5 models, XGBoost gives the best performance for precision, Random Forest gives the best performance for recall, while Logistic Regression gives a relatively poor performance for both precision and recall rates.

By computational time, a much longer time is required for XGBoost model as compared to Random Forest model. WIth no significant increase in accuracy, AUROC and the precision-recall, we thus conclude that Random Forest has the best overall performance for this data and we would analyze the feature importance based on its result.

## **3.4 Feature Importance**

As mentioned above, Random Forest Classifier performs the best among all five models. Therefore, we use the result of Random Forest Classifier to gain insights on significant features.

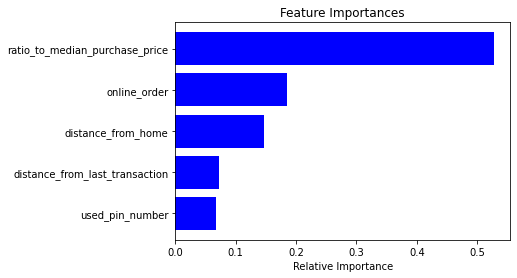


Figure 8

As per Figure 11 above, the ratio\_to\_median\_purchase\_price has the highest importance, it is the most predictive of the 5 factors. Then, followed by online order, and distance from home. It is evident that these factors can be important factors for consideration, as we have mentioned previously, with the technological advancements nowadays, cyber criminals can easily gain access to personal information, and are able to carry out fraudulent activities even when they are cross-border, through online methods.

Applying these factors in the real world, to reduce the risks from these factors, it is essential for credit card companies and banks to improve their authentication methods, especially for large volume transactions. There are current measures taken by banks, such as US.Bancorp offering opt-in for customers to allow them to track their cell phone locations[[11]](#footnote-10), this is insufficient with the rising crimes, and banks need to come up with even sophisticated solutions to tackle risk from cross-border payments. In addition, online merchants also need to be more vigilant and upgrade their payment process to improve the security for consumers.

# **Conclusion and Future studies**

To combat credit card fraud effectively, it is important to understand the modus operandi employed by fraudsters. Contrary to popular belief, merchants are far more at risk from credit card fraud than the cardholders. While consumers might have difficulties in reversing the fraudulent charges, merchants lose out on both the product that was sold and the cost of the product, having to pay a chargeback fee and even face far reaching consequences such as reputational damage or the risk of shutting down.

Even though our model accuracy is good without much hyperparameter tuning, this is usually a practice that is done in order to further optimize and improve model performance. We can explore Bayesian Search hyperparameter tuning since it balances between exploration and exploitation, thus giving us the best hyperparameters without using too much computation resources and time.

With increasing fraud patterns by cyber criminals, we can capture more of the patterns by including more factors/fraud indicators such as the “initial balance of the credit card” and ending balance of the credit card in the model, so that more comprehensive decisions can be made as well.

1. <https://www.businesswire.com/news/home/20220513005241/en/Global-Credit-Card-Payment-Market-2022-to-2027---Industry-Trends-Share-Size-Growth-Opportunity-and-Forecasts---ResearchAndMarkets.com> [↑](#footnote-ref-0)
2. <https://nilsonreport.com/upload/content_promo/NilsonReport_Issue1209.pdf> [↑](#footnote-ref-1)
3. <https://public.tableau.com/app/profile/federal.trade.commission/viz/TheBigViewAllSentinelReports/TopReports> [↑](#footnote-ref-2)
4. <https://www.altexsoft.com/blog/credit-card-fraud-detection/> [↑](#footnote-ref-3)
5. <https://www.kaggle.com/datasets/dhanushnarayananr/credit-card-fraud> [↑](#footnote-ref-4)
6. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4120293/#:~:text=The%20log%20transformation%20is%2C%20arguably,normal%20or%20near%20normal%20distribution>. [↑](#footnote-ref-5)
7. <https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc> [↑](#footnote-ref-6)
8. <https://towardsdatascience.com/illustrating-predictive-models-with-the-roc-curve-67e7b3aa8914#:~:text=Area%20Under%20the%20Curve%20(AUC)&text=This%20is%20the%20probability%20that,unable%20to%20discriminate%20between%20classes>. [↑](#footnote-ref-7)
9. <https://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html> [↑](#footnote-ref-8)
10. <https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-imbalanced-classification/> [↑](#footnote-ref-9)
11. <https://time.com/4247847/banks-tracking-cell-phone-fraud/> [↑](#footnote-ref-10)