|  |  |
| --- | --- |
| Programme | MSc Management with Data Analytics |
| Module name | Applied Modelling and Visualisation |
| Schedule Term |  |
| Student Reference Number (SRN) |  |
| Report/Assignment Title | Applied Modelling and Visualisation |
| Date of Submission |  |
| Declaration of Original Work:  I hereby declare that I have read and understood BPP’s regulations on plagiarism and that this is my original work, researched, undertaken, completed and submitted in accordance with the requirements of BPP School of Business and Technology. The word count, excluding contents table, bibliography and appendices, is \_\_\_ words.  Student Reference Number: Date: | |
| By submitting this coursework, you agree to all rules and regulations of BPP regarding assessments and awards for programmes. Please note, submission is your declaration you are fit to sit. BPP University reserves the right to use all submitted work for educational purposes and may request that work be published for a wider audience. BPP School of Business and Technology | |

# Executive Summary

Credit card fraud continue to be a menace to cardholders. As much as there are safety measures put in place by financial institutions, still credit card fraud continues to remain prevalent. The widespread of these card fraud cases begs the question whether the financial instutions and government agencies are doing enough to curb the menace. Our study finding reveal that cross-country transactions with different country of transaction and “shipping address” are the leading transactions with credit card fraud. As a result, there is a strong correlation between “country of transaction” and “shipping address”. With the implementation of a XGBoost Classifier Model, organizations can benefit from low error model as compared to a Random Forest Classifier model.

Table of Contents

[Executive Summary 2](#_Toc102076036)

[Applied Modelling and Visualisation 4](#_Toc102076037)

[Business Understanding 4](#_Toc102076038)

[Types of credit card fraud 5](#_Toc102076039)

[Problem Statement 6](#_Toc102076040)

[Data cleaning 7](#_Toc102076041)

[Exploratory Data Analysis 9](#_Toc102076042)

[Modelling 17](#_Toc102076043)

[Evaluations 17](#_Toc102076044)

[Conclusion 18](#_Toc102076045)

[References 19](#_Toc102076046)

# Applied Modelling and Visualisation

To successful complete this project, we pursue the cross-industry process for data mining (CRISP-DM) methodology. The first step in the CRISP-DM process is to figure out what you want to accomplish from a business standpoint. Other subsequent steps include; data understanding, Business understanding, Data understanding, Data preparation, Modelling, Evaluation, and Deployment, (Smart Vision Europe, n.d). Hence, the CRISP-DM) methodology shall be our guide throughout the project.

## Business Understanding

With the onset of the internet a lot of things have move online. E-commerce has accelerated the use of online payments as well. As more people transact online, cybercriminals prowl the web to take advantage of card vulnerabilities and steal from cardholders. Many transactions online are settled through, e-wallets, credit card, debit cards and payment vouchers. Therefore, credit card fraud occurs when someone obtains funds falsely either through a debit card or debit card. The cyber criminals act with the primary focus of obtaining services or goods, or to transfer funds to another bank controlled by a criminal. Card fraud is complex and sometimes a genuine customer may unknowingly approve the transfers funds to another account controlled by a lawbreaker. Alternatively, a transaction can occur without the authorization of the cardholder for the payment to proceed and the payment is carried out by a third party.

According to "FRAUD THE FACTS 2019- The definitive overview of payment industry fraud" (2019), the United Kingdom’s unlawful fraudulent financial losses from card transactions and remote banking accounted for £844.8 million in 2018. Financial institutions suffered a loss of approximately £1.66 billion in 2018 from card fraud cases. As a result, for every £2 card fraud prevented, financial institutions suffered £3 of undetected fraudulent activity. Credit cards are much more protected than ever now, thanks to regulatory agencies, banking institutions, and banks putting in extra time and resources to work collaboratively with fraud forensic experts. Funds belonging to cardholder are protected by regulatory requirements that hold the card issuer and the concerned financial institution responsible. Online payment systems and safeguards are improving, making it much more difficult for criminals to steal money.

There are two types of card fraud: card-present theft and card-not-present theft, the latter of which is more common. Compromise can take place in a variety of ways, and it normally happens without the cardholder's knowledge. Database security lapses have become especially costly as a result of the internet; in some cases, massive amounts of transactions have been subverted. Cardholders can quickly report stolen cards, but the details of a compromised account may be kept by a cyber-criminal for months before being stolen, making it much more difficult to identify the source of the compromise. It is possible that the account holder will not notice fraudulent activity unless he or she acquires a statement. Cardholders can help decrease their cases of fraud by checking their accounts on a regular basis to ensure there are no suspicious transactions.

If a credit card number is compromised, it can be used for illegal transactions until the owner notifies the financial institution to suspend the account. Most banks provide around the clock contact numbers to encourage prompt reporting. Nonetheless, a cybercriminal may be able to make unapproved transactions on a card before it is suspended.

## Types of credit card fraud

**Counterfeit and skimming fraud**

According to Jolly (2019) counterfeit and skimming fraud accounted for $14,935,409 in Australia in 2018. Skimming is the stealing of private details from a transaction that appears to be normal. Basic methods, such as photocopying receipts, or more advanced methods, can be used to obtain a victim's card details. When a credit card skimmer is connected to an Automated teller machine or a vendor's terminal, it steals information from the magnetic stripe on your credit card. You could be skimmed if someone comes near you with a payment card skimmer. Skimming information can then be used to create a forgery card or commit card-not-present fraud.

**False application fraud**

This type of fraud occurs when a person’s commits a criminal act by using someone else personal details to open an account. A lawbreaker might steal or make fake documents such as identity cards, passport, utility bills or bank statement to create a personal profile with a financial institution (Ronchetti, 2020). Once the account is approved, criminals use it obtain cash or credit other a different person. As a result, the genuine persons suffers a financial loss.

Veda performed a credit card application analysis in 2014 and unearthed that $1.6 billion in credit card applications were identified as suspicious transactions (Jolly, 2019). Most financial institutions take application fraud seriously and have countless safeguards in place to prevent it, however this does not necessarily imply that the infrequent scammers will not fall between the cracks. It is important to k Stay on top of your financial transactions, conceal private information, and, most importantly, disclose any suspicious practices as quickly as possible.

**Social engineering fraud**

Get Indemnity (n,d), defined social engineering fraud as the process by which a criminal impersonates another person in order to obtain a consensual movement of funds or confidential data to the cybercriminal. Cybercriminals have become much more advanced in their techniques of committing fraud on people and business entities. Submitting fake email messages deceiving a senior member of staff and attempting to dupe staff members into sending funds to an illegitimate bank account is a common tactic.

Falsely claiming to be a financial institution or disbursement processor, cybercriminals can use a number of tactics to obtain sensitive information. Another most common cyber technique for gaining the victim's trust is phone line phishing. Business owners can ensure the avoid such traps through dual authentication procedures before transferring funds, which requires two people to give approval, as well as a call confirmation procedure to a pre-existing contact number instead of the contact details included in the payment request ("What is credit card fraud and how can I prevent it?", 2021). The cardholder’s bank is required to reimburse you for any non - authorized disbursement; nevertheless, they could perhaps deny a reimbursement if you can confirm you authorized the transfer of funds, or if you are at fault since you did act intentionally or did fail to safeguard sensitive data that allowed the transfer of funds to occur.

Taking personal responsibility as a cardholder by being cautious is part of the ways of deterring card fraud. It is the collective responsibility as well for the financial institutions and government regulatory agencies to ensure that consumers are protected from card fraud. The measures above may not be enough to completely curb card fraud. Therefore, in a study we aim to offer a machine learning solution that can offers real-time monitoring of transactions. The model should be able to classify transaction with the highest degree of accuracy whether they are fraudulent transactions or not.

## Problem Statement

Every year, card fraud tends to cost billions of pounds. While credit card fraud is expensive for the banking industry, so is detecting it. Financial institutions must consider the costs of fraud against the implications of burdensome regulations on consumer experience in a competitive market.  False positives in the detecting fraud procedure cause customer transactions to be rejected incorrectly. This emphasizes the importance of an effective predictive model with a low false positive rate. Therefore, this study focuses on developing a model that can predict fraudulent transactions with low error rates

*Success criteria*: To come up with a model with less than 20 percent false positive rate.

Data Understanding

The dataset obtained had a 100,000 records and 15 features. The features in the dataset included; (1) 'Transaction ID': represent a unique identifier for every card transaction with non-numeric data types; (2) 'Date' : dates with which the transaction took place stored as “objects.” , (3) 'Time': represents the hour in which the transactions occurred and had numeric values,(4) 'Type of Card': describes whether the card was a MasterCard or visa, (5)' Entry Mode’: describes how the transaction was finalised either through “Pin” or “Tapping”, (6) ‘Amount': states the value of the transactions,(7) 'Type of Transaction': describes whether the transaction was done “online”, ”Point of sale (POS)“ or ATM , (8) 'Merchant Group', merchants were classified into; 'Children', 'Electronics', 'Entertainment', 'Fashion', 'Food', 'Gaming', 'Products', 'Restaurant', 'Services', ‘Subscription', (9) 'Country of Transaction' included; 'China', 'India', 'Russia', 'USA', 'United Kingdom' , (10) 'Shipping Address' included; 'China', 'India', 'Russia', 'USA', 'United Kingdom', (11) 'Country of Residence' included; 'China', 'India', 'Russia', 'USA', 'United Kingdom', (12) 'Gender': classified users into male and female, (13) 'Age': numerical values describing the ages of users, (14) 'Bank' transaction from the following banks ; 'Barlcays', 'HSBC', 'Halifax', 'Lloyds', 'Metro', 'Monzo', 'RBS', and (15)'Fraud' numeric values of “0” and “1” representing yes and no.

## Data cleaning

The dataset was checked for missing values and none was identified. However, we checked for duplicates and found around 5,000 records with duplicate values. The 'Transaction ID' ought to be unique since it differentiates one transaction from another. Therefore, we cleaning the data by removing duplicates based on the 'Transaction ID' variable while maintaining the first item in every set of duplicates. The resulting data frame had a records of 95,687. In addition, we transformed the datatypes for “Date” into ‘datetime64[ns]’ format. The transformation was necessary to allow us process the data further by deriving certain periods. Moreover, the value of transactions had been stored as “string objects” since it contained the pound sign. The feature was transformed into numeric values by stripping the ‘£’ using the regular expression. The data set was further subdivided into numeric and categorical dataset for easy data manipulation and wrangling. The resulting dataset include; “num\_df” and “cat\_df” which were used to obtain the descriptive statics

*Descriptive statics*

The numeric data represent by “num\_df” dataframe included the following features; 'Date', 'Time', 'Amount', 'Age'. The out of descriptive data is a shown below.

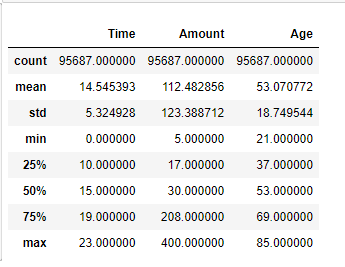


Figure 1: Numerical features

In general, the mean for Time was 14.54 hours, maximum: 23, and minimum: 0.00. the average value of a transaction was £112 while the maximum amount was £ 400 and the minimum value being £ 5. The cardholders were all above the age of 21 whereas, the oldest cardholder was 85 years. The average age of the cardholders is 53 years.

On the other hand, we undertook the descriptive statistics of categorical variables. The categorical features included; 'Transaction ID', 'Type of Card', 'Entry Mode', 'Type of Transaction', 'Merchant Group', 'Country of Transaction', 'Shipping Address', 'Country of Residence', 'Gender', 'Bank', and ‘Fraud'.

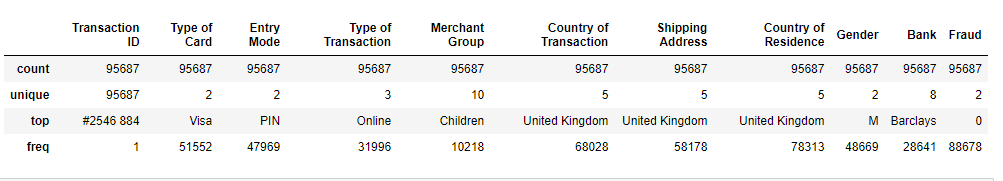


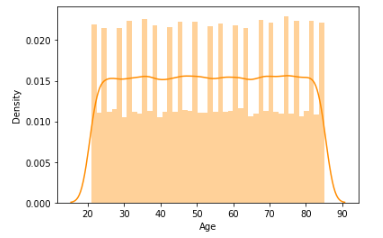
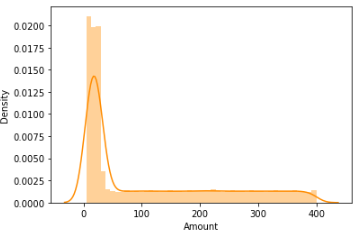
Figure 2: Categorical features

After the data cleaning the 'Transaction ID' had only single unique values, confirm that the duplicates had been eliminated. The descriptive statistics reveal that visa was the most popular type of card, pin was the most popular “Entry Mode”, online transactions were the most popular types of transaction, merchants classified as children were dominant, United Kingdom was the most common country of transaction as well as shipping address, and country of residence, most card holders were male, Barclays had the most transactions and “the dominant value for fraud was “0”.

## Exploratory Data Analysis

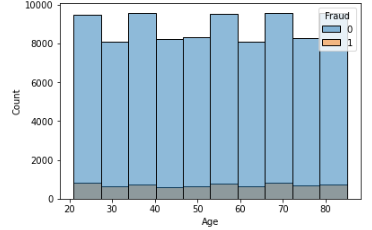
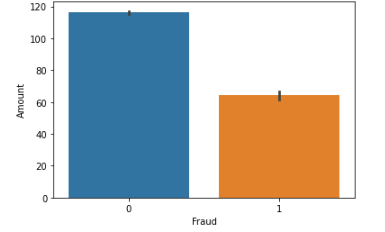
**Univariate analysis**

Univariate analysis is the analysis of a single variable. Univariate analysis does not explore relationships and focuses on descriptive statistics (Zach, 2021). The univariate analysis revealed the following;

* *

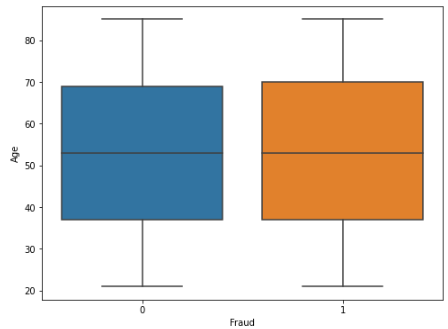
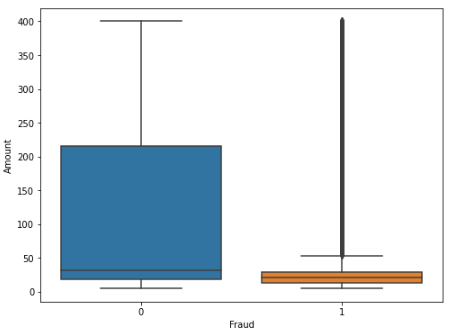
*Figure 3: “Age “ Distribution Figure 4: “Amount “ Distribution*

"Age" has a normal distribution, while “Amount” is skewed to the left. With most values being below 100 pounds

**

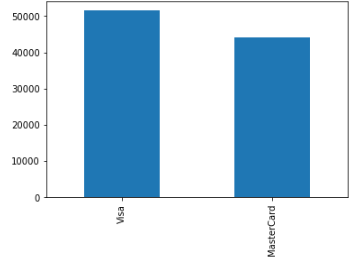
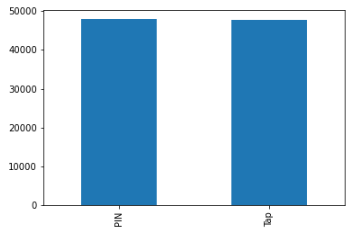
*Figure 5: Average “Fraud” figure 6: histogram “Age” vs “Fraud”*

The frequency of fraud was equal across all age groups; the mean amount of non-fraud transactions was around £118 while fraudulent transactions had a mean of around £60.

**

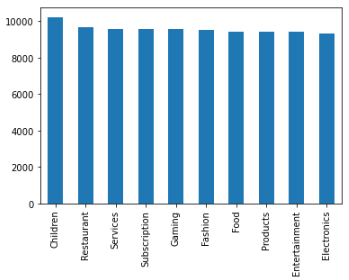
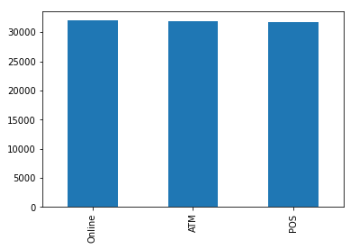
*Figure 7: Boxplot of “Amount” vs “Fraud” Figure 8: Boxplot of “Age” vs “Fraud”*

The box plot revealed outliers in the transactions with fraudulent activities, however, none was dropped since it is the nature of fraudulent activities to have outliers. The were no outliers for the “Age “vs “Fraud’.

* *

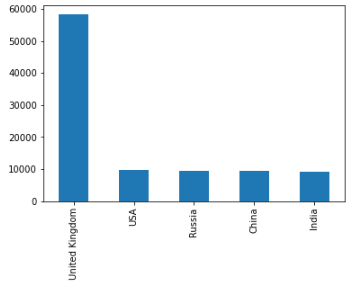
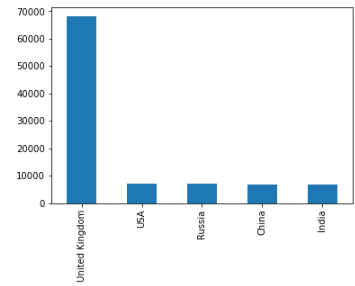
*Figure 8: 'Type of Card' Figure 9 : 'Entry Mode'*

Visa transactions were slightly more than MasterCard; approximately 50,000 and 45,000 respectively; the frequency for “Tap” and “Pin” were relatively equal.



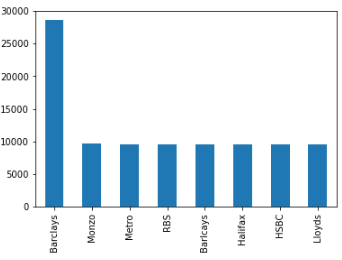
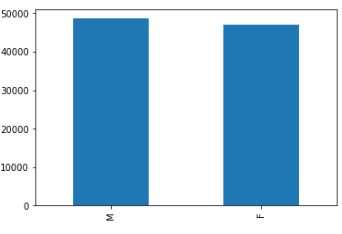
*Figure 10: 'Type of Transaction' Figure 12 : 'Merchant Group'*

“online”, “ATM”, and “POS” transaction have similar number of occurrence. Children merchant types are slightly dominating in the merchant groups while the rest of the merchant types are relatively equal on occurrence.



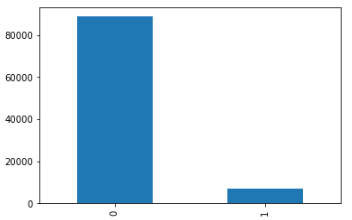
*Figure 13: ''Shipping Address' Figure 14 : 'Country of Residence'*

United Kingdom was dominant in all the three categories in terms of occurrence; 'Country of Transaction', 'Shipping Address', and 'Country of Residence'. While Russia, USA, China, and India had relatively equal.

 *Figure 15: 'Gender' Figure 16 : 'Bank*

Barclays transaction were profoundly dominant in occurrence compared to other banks that share equal frequency, and males’ transactions were almost as equal as females’

*‘*

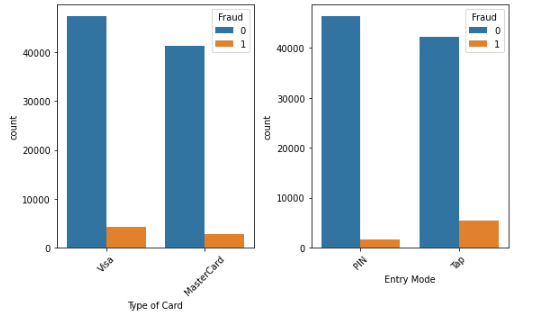


*Figure 17 : 'Fraud*’

Non-fraudulent cases were significantly more than fraudulent ones.

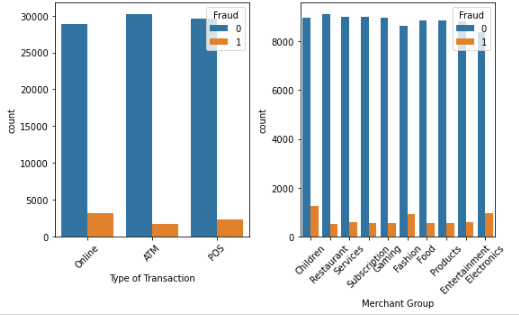
**Bivariate analysis**

The bivariate analysis was done in relation with reference to the target variable “Fraud”. The finding of the bivariate analysis includes;



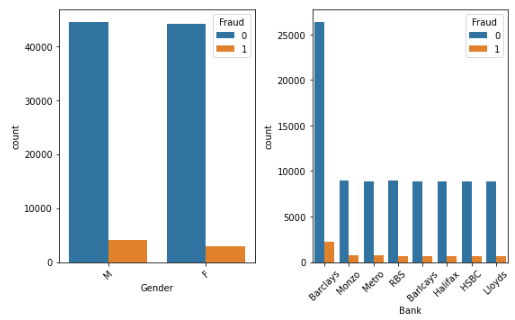
*Figure 18: “type of card vs Fraud” and “Entry mode” vs “Fraud”*

Fraud cases were slightly higher in Visa than MasterCard transactions; most fraud cases were observed in “tap” mode of entry more than PIN.



*Figure 19 “type of Transaction vs Fraud” and “Merchant” vs “Fraud”*

online transactions had slightly higher fraud cases followed by “POS” and lastly “ATM. Children, fashion and electronics merchants had higher cases of fraud more than other merchant types.



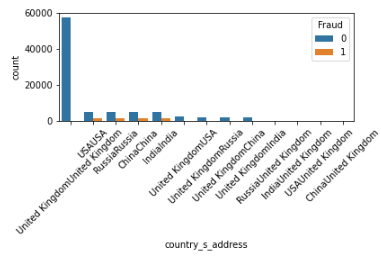
*Figure 20 “Gender” vs “Fraud” and “Bankt” vs “Fraud”*

Male transactions had more fraud cases than females’, (i) Barclays as expected had many fraud cases because it had the highest number of transactions, while the other banks faired on equally; all countries in “country of transaction” had relatively equally occurrence of fraud cases.

**

*Figure 21 “Country of transaction” vs “Fraud” and “Shipping adress” vs “Fraud”*

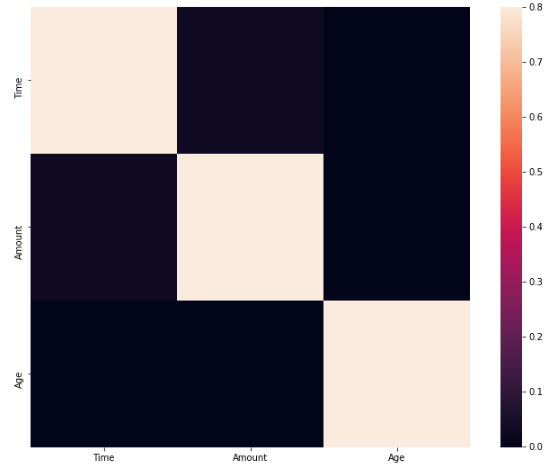
Despite the United kingdom being the most dominant shipping address, it had the least fraud cases compared to other shipping addresses; USA, Russia, China, and India.



*Figure 22 “Country \_s\_address” vs “Fraud”*

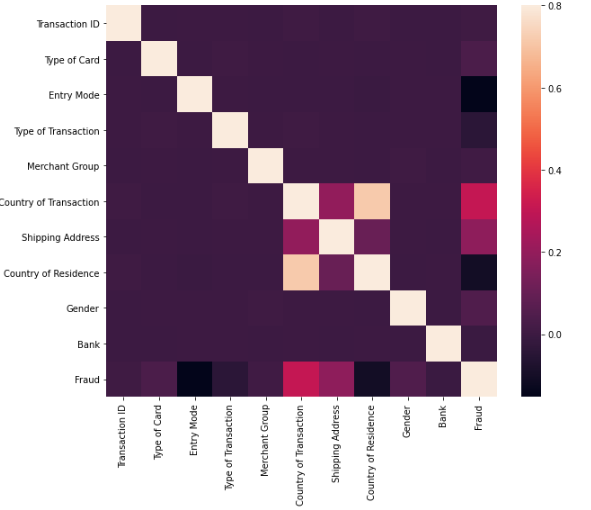
This new feature revealed that only transactions that had similar “country of transactions” and shipping address had fraud cases, but cross-country transactions did not involve any fraud.

**Multivariate analysis**



*Figure 23 “corrilation of numeric data”*

Pan (n.d.) argues that correlation analysis is the study of relationship between two or more variables. Hence, in this section we explore how the variables are correlated. The correlation analysis was carried out separately for the numeric and categorical dataset. In general, the numeric data revealed weak correlation between variables while the categorical data had interesting findings. The correlation analysis shows moderate positive correlation between Fraud and “country of transactions” and shipping address. In addition, the ‘country of residence’ had a positive strong correlation with ‘country of transaction”; as a result, it would be appropriate to drop either of the two variables.



*Figure 24 “corrilation of categorical data”*

**Data preparations**

The exploratory analysis gave us some much insight on the pattern of fraud occurrence. As a result, new features were engineered with the hope that it will post the performance of the classification model. The most significant find was the correlations between ‘Fraud’ and “country of transactions” and shipping address. Therefore, new feature that combines “country of transactions” and shipping address was developed. This new feature revealed that only transactions that had similar “country of transactions” and shipping address had fraud cases, but cross-country transactions did not involve any fraud.

Additionally, new features including names day of week and months were added. Since, fraud transactions had a lower mean of around £60 pounds, new feature that groups amount into “less than £60”, “between £60 and £120”, and above “£120” was engineered.

Further preparations include determining the target variable and the predictor variables. Target variables were label encoded and stored as “y” while the predictor variables were label-encoded and stored as “x”. Using the “train\_test\_split” pre-processing method available in the skit-learn library, each dataset was split into 80 percent, training dataset and 20 percent testing dataset.

## Modelling

For modelling we used two algorithms, XGBoost classifier and Random Forest classifier. Donges (2021) suggests that The random forest technique leverages a single tree model that outputs high variance with low predictive accuracy and turns them into a fairly accurate prediction model. Therefore, a random forest combines many weak trees and uses their predictive power to make an overall prediction, resulting to a more robust model. Whereas, according to NVDIA (n.d.), Gradient Growth Trees are an algorithm for training a group of decision trees for classification and regression, similar to a random forest. Collective learning algorithms combine several machine learning algorithms to create better models. with increasing the gradient level amplification gradual process Weak modelling is formalized as an algorithm for falling gradients on the target function. Gradient gain targets the result of the next model to reduce errors.

Default parameters were used for the XGBoost classifier while the Random Forest classifier number of trees parameter was set 300 trees. A confusion matrix and accuracy score was used to evaluate both models. The XGBoost classifier model had the following output

for the confusion matrix  . The matrix translates to True Positive Rate (TPR) of 99.72 percent and a False Positive Rate (FPR) of 15.22 percent. The overall accuracy was 98.66 percent. Random Forest classifier, on the other end hand, the following out; . The matrix translates to True Positive Rate (TPR) of 98.68 percent and a False Positive Rate (FPR) of 16.62 percent. The overall accuracy was 98.57 percent.

## Evaluations

The CRISP-DM methodology demands that the outcome of the model assessment is evaluated based on business objectives. The business criteria must be met for a model to proceed into deployment. Considering that business objective is to have a robust system that cuts save cost in detecting fraud cases effectively while not frustrating the customer experience with false positive alarms. The success criteria of the project were to come up with a classification model with a false positive rate of below 20 percent. Indeed, both our models passed the business success criteria. However, one model was better that the other. The model assessment outcome reveals that the Random Forest Classifier model had a false positive rate of 16.62 percent compared to 15.22 percent for the XGBoost Classifier (Kumar, 2020). The difference might be marginal but the economic impact of the difference is substantial. As a result, it would be appropriate to move the XGBoost Classification model into production.

# Conclusion

The onset of the internet has accelerated the adoption of ecommerce and electronic payment system. Popular electronic payment systems include Visa and MasterCard. Cardholders can finalise payment through “tapping” or entering “PIN”; as a result, personal information can be obtained during this stage by cybercriminals. However, other forms of credit card may happen through social engineering, false identity application, and Counterfeit and skimming fraud. Financial institutions including cardholders suffer huge financial because of credit card fraud. Security features and procedures put in place seems not to be enough to deter such kind of fraud. As an additional measure, we have adopted a data driven approach in developing a model that can work in real-time by assessing transactions and classifying them whether they are fraudulent or not. A Random Forest classifier algorithm proved to be a better classifier over XGBoost Classifier. This data-driven approach appears to be effective as it is able to achieve low error rate in identifying fraudulent cases. Hence, it would help a great deal in minimizing credit card fraud cases.

# References

*“FRAUD THE FACTS 2019- The definitive overview of payment industry fraud”*. Ukfinance.org.uk. (2019). Retrieved 24 April 2022, from https://www.ukfinance.org.uk/system/files/Fraud%20The%20Facts%202019%20-%20FINAL%20ONLINE.pdf.

Donges, N. (2021). *Random Forest Algorithm: A Complete Guide*. Built In. Retrieved 24 April 2022, from https://builtin.com/data-science/random-forest-algorithm.

Kumar, A. (2020). *ROC Curve & AUC Explained with Python*. vitalflux.com. Retrieved 24 April 2022, from https://vitalflux.com/roc-curve-auc-python-false-positive-true-positive-rate/.

Pan, L. CORRELATION ANALYSIS. *A-To-Z Guide To Thermodynamics, Heat And Mass Transfer, And Fluids Engineering*. https://doi.org/10.1615/atoz.c.correlation\_analysis

Ronchetti, G. (2020). *What is Application Fraud? How XTN can help - XTN Cognitive Security*. XTN Cognitive Security. Retrieved 24 April 2022, from https://xtncognitivesecurity.com/what-is-application-fraud-and-how-xtn-can-help/.

Smart Vision Europe. *Crisp DM methodology*. Smart Vision Europe. Retrieved 24 April 2022, from https://www.sv-europe.com/crisp-dm-methodology/.

*What is credit card fraud and how can I prevent it?*. MaPS. (2021). Retrieved 25 April 2022, from https://www.moneyhelper.org.uk/en/blog/scams-and-fraud/what-is-credit-card-fraud-how-prevent-it.

Zach. (2021). *What is Univariate Analysis? (Definition & Example) - Statology*. Statology. Retrieved 24 April 2022, from https://www.statology.org/univariate-analysis/.