https://gitlab.uwe.ac.uk/sjr2-irons/ml-coursework

**Title: Machine Learning Credit Card Prediction**

**Abstract**

The widespread availability of online banking services has greatly facilitated many people's day-to-day lives by facilitating cash transactions, bill payments, and online shopping. While there are many benefits to conducting business online, users nonetheless face a serious challenge in the form of financial theft. If an unauthorised third party is able to bypass the bank's security measures and pose as a legal customer, this is considered fraud. The effects of financial fraud are widespread and growing, with serious consequences throughout the sector.

The credit card is a popular kind of consumer credit since it allows consumers to make purchases even when they don't have the cash on hand to pay for them. Use it for anything from petrol and groceries to gadgets and travel costs to regular old retail therapy. When utilised for a variety of purchases, credit cards' reward points can add up to great value.

Source data and Jupyter Notebooks for running the analysis are available here:

Github link

**Introduction**

Finance, insurance, e-commerce, and healthcare have fraud detection issues. Traditional rule-based systems cannot keep up with sophisticated fraud trends. Thus, machine learning has garnered attention for its potential to detect and prevent fraud. This literature review addresses recent machine learning algorithm fraud detection research, including methodology, datasets, evaluation metrics, and obstacles.

Credit card fraud threatens banks and consumers. Detecting fraud accurately reduces losses and protects customers. We analyse and simulate credit card fraud using a public dataset in this research. We want to build a model that detects fraudulent credit card transactions with few false positives.

**Problem Definition**

The objective of this project is to utilise various machine learning algorithms to identify instances of credit card fraud based on transaction time and amount.

**Dataset Description and Analysis**

UCI Machine Learning dataset repositories http://archive.ics.uci.edu/dataset/27/credit+approval provided the dataset.

This analysis utilises Addis Ababa Sub city police data from a Master's thesis. The dataset includes 32 features and 12,316 road incidents from 2017 to 2020. Preprocessing encoded attributes and removed personally identifiable information.

This dataset simulates valid and fraudulent credit card transactions from January 1, 2019, through December 31, 2020. It safeguards 1000 customers' credit card data from 800 merchants.

Brandon Harris' Sparkov Data Generation and Github app created this. This simulation ran January 1, 2019, through December 31, 2020. Files were combined and formatted.

The simulator has standard businesses, customers, and purchases. The "faker" python module and your simulation's clients and business owners build an intermediate list.

**Exploratory data analysis**

The dataset was explored and some of the following visuals and insights were gathered.

Fig 1: A glimpse of the dataset

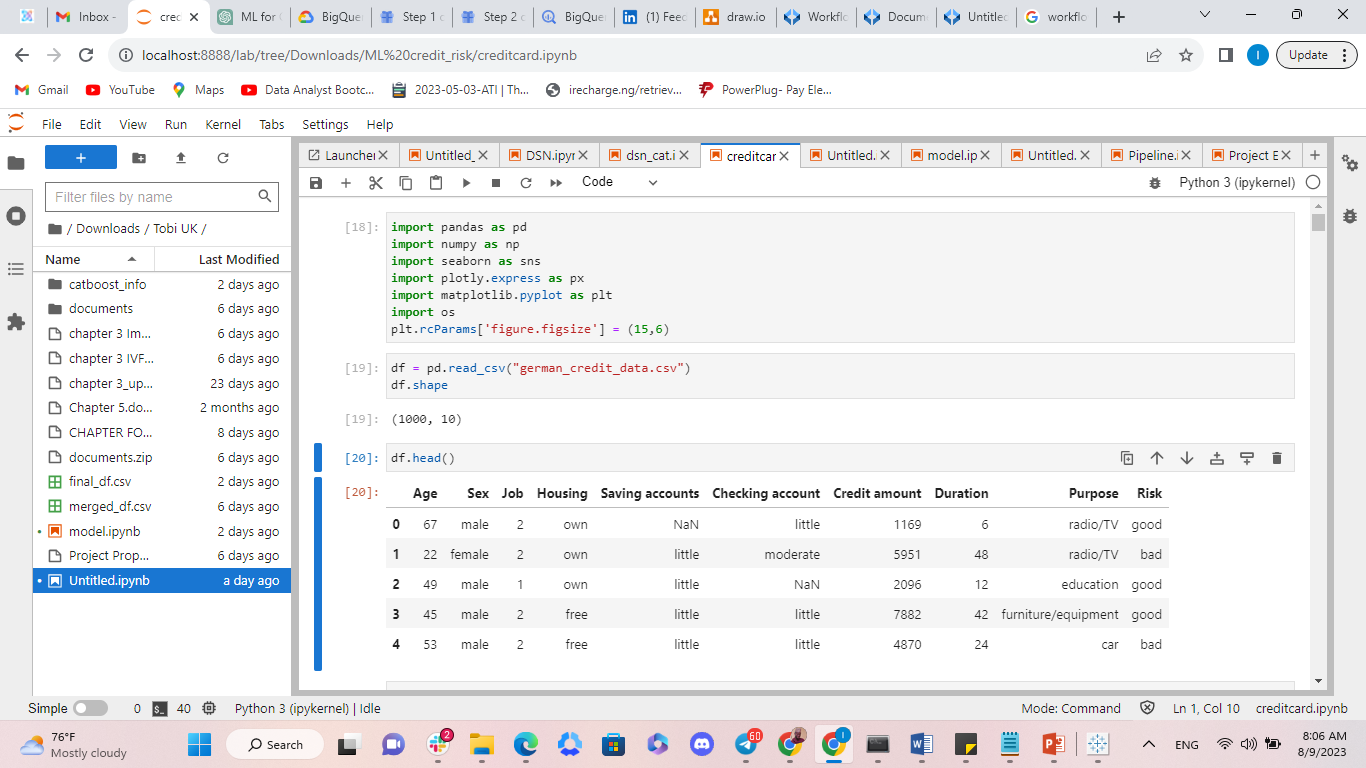
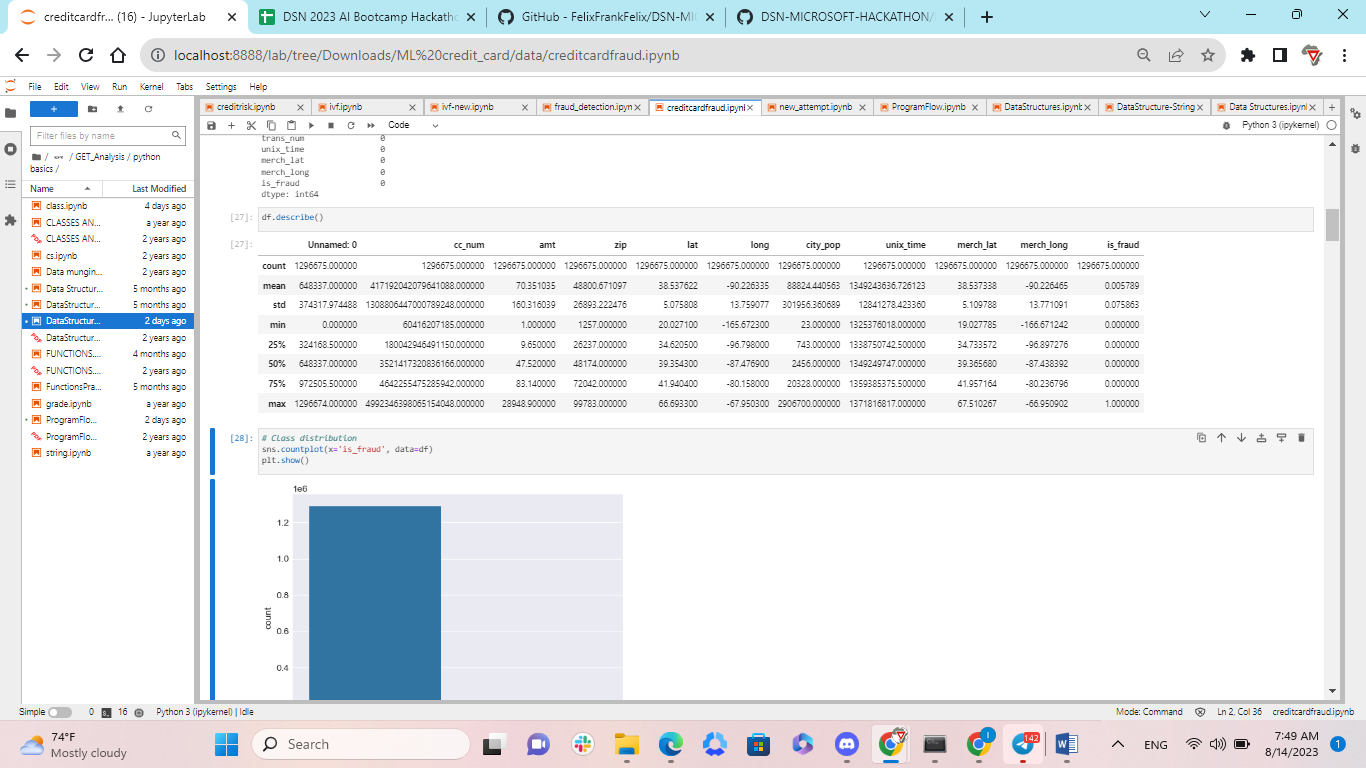


Fig 2: A descriptive analysis of the dataset



The descriptive result, which summarises the dataset, includes count, mean, standard deviation, minimum, maximum, 25%, 50%,70%.

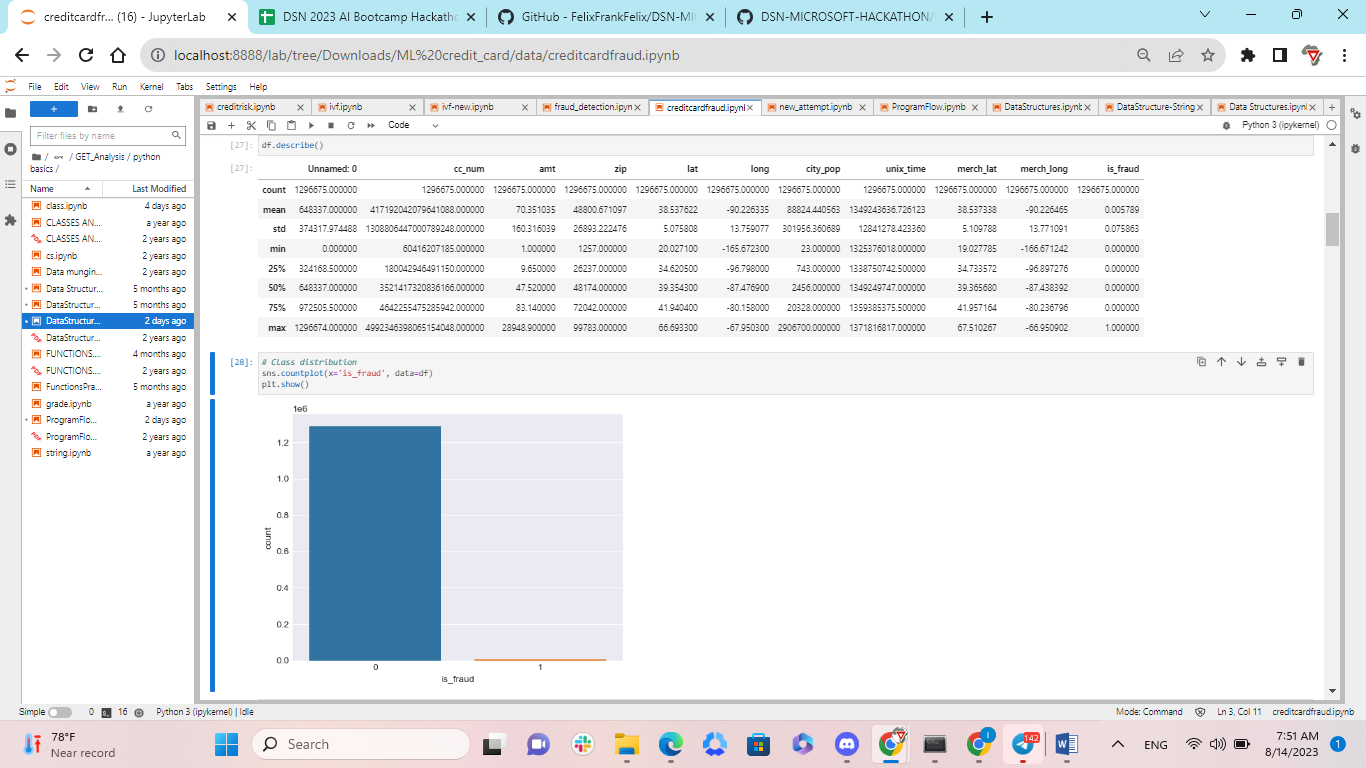


Fig 3: Count for is\_fraud target

The dataset's is\_fraud imbalance class demonstrates that class 0 has higher value than 1.

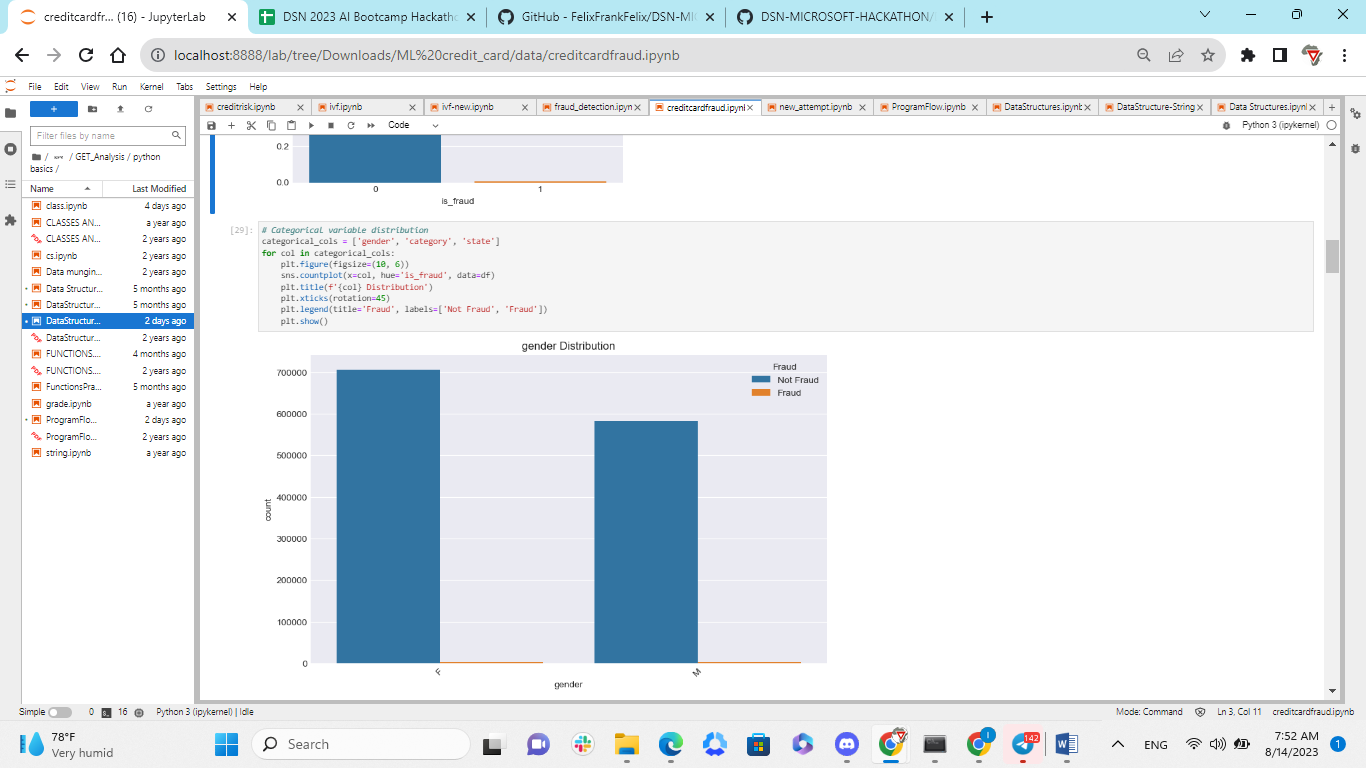


Fig 4: Gender distribution with the Is\_fraud

The result above shows that most gender transactions are recorded as not fraud with a little portion identified as fraud in the dataset.

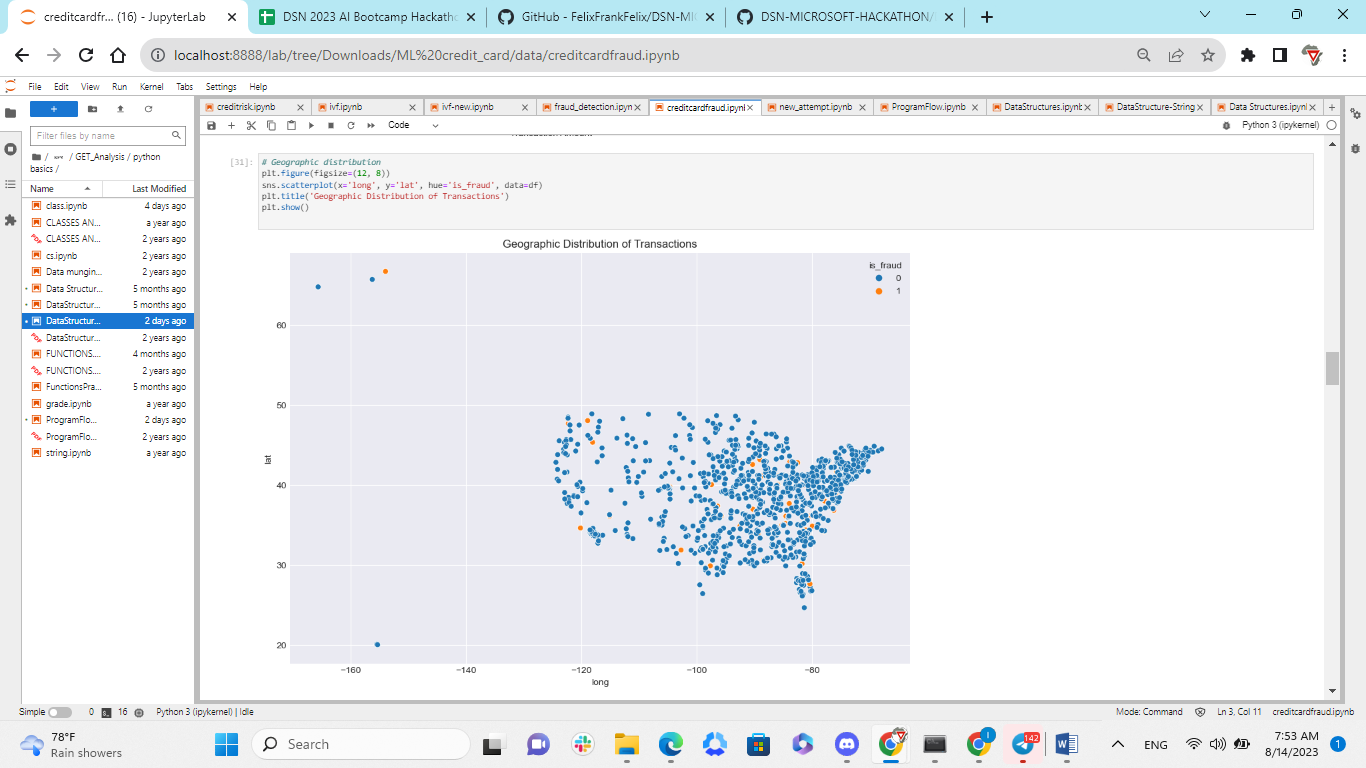
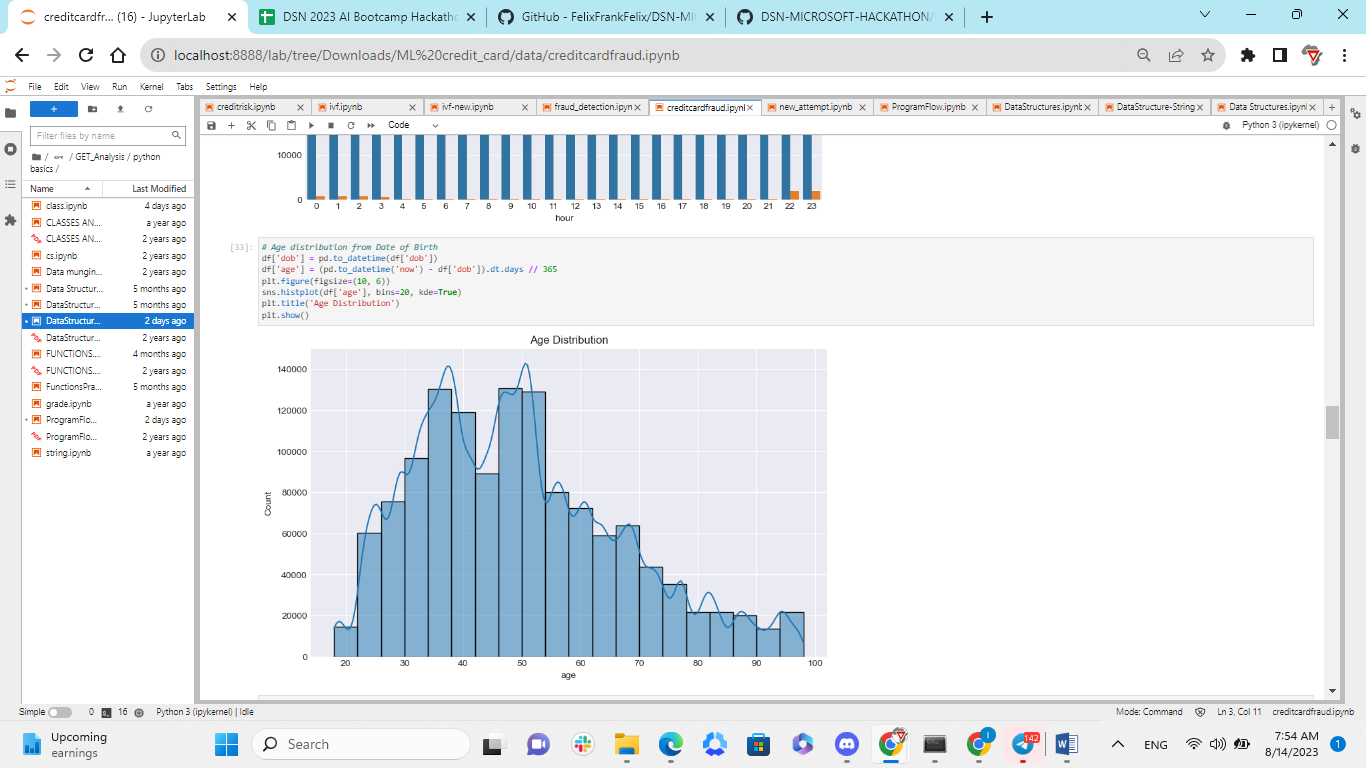


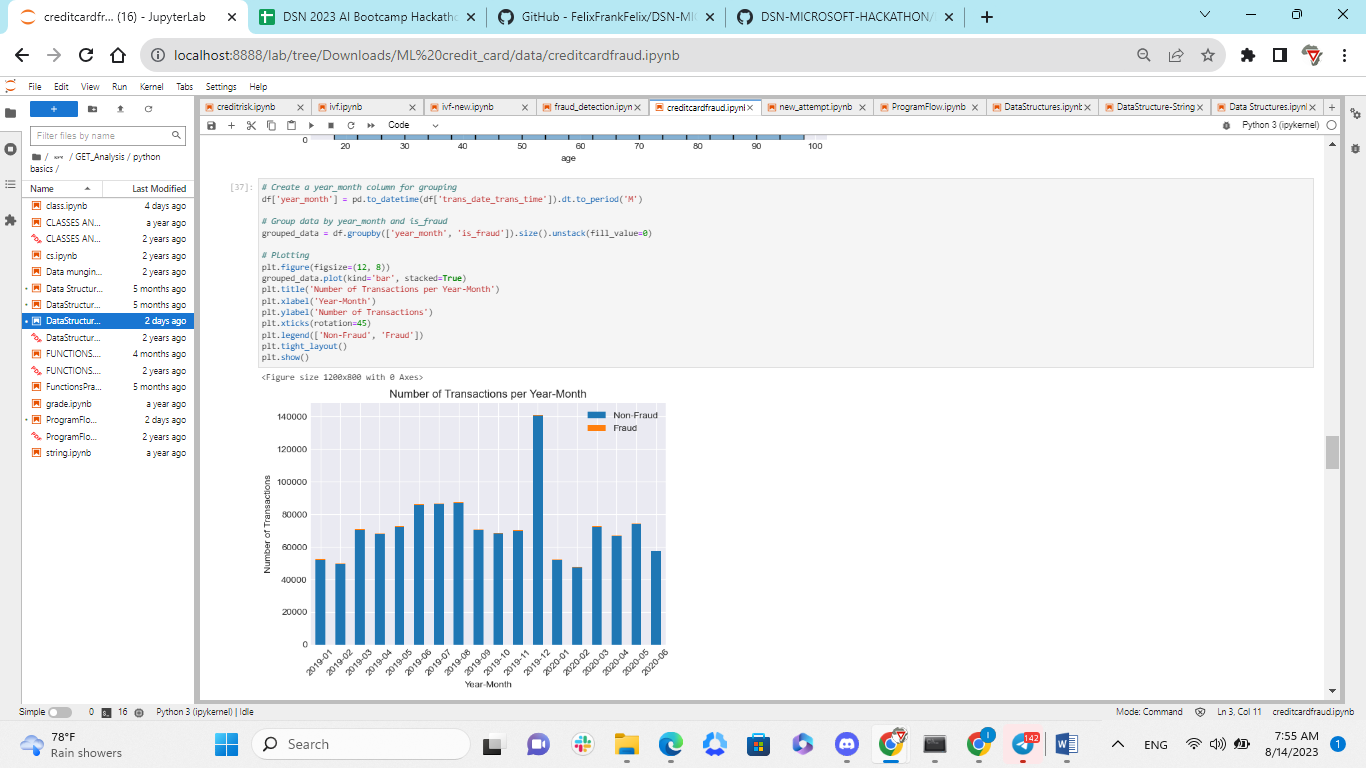
Fig 5: Geographic Distribution of Transactions

The result shows that based on the geographical latitude and longitude of customer, there are non fraud transactions recorded and a few portion of fraud recorded.



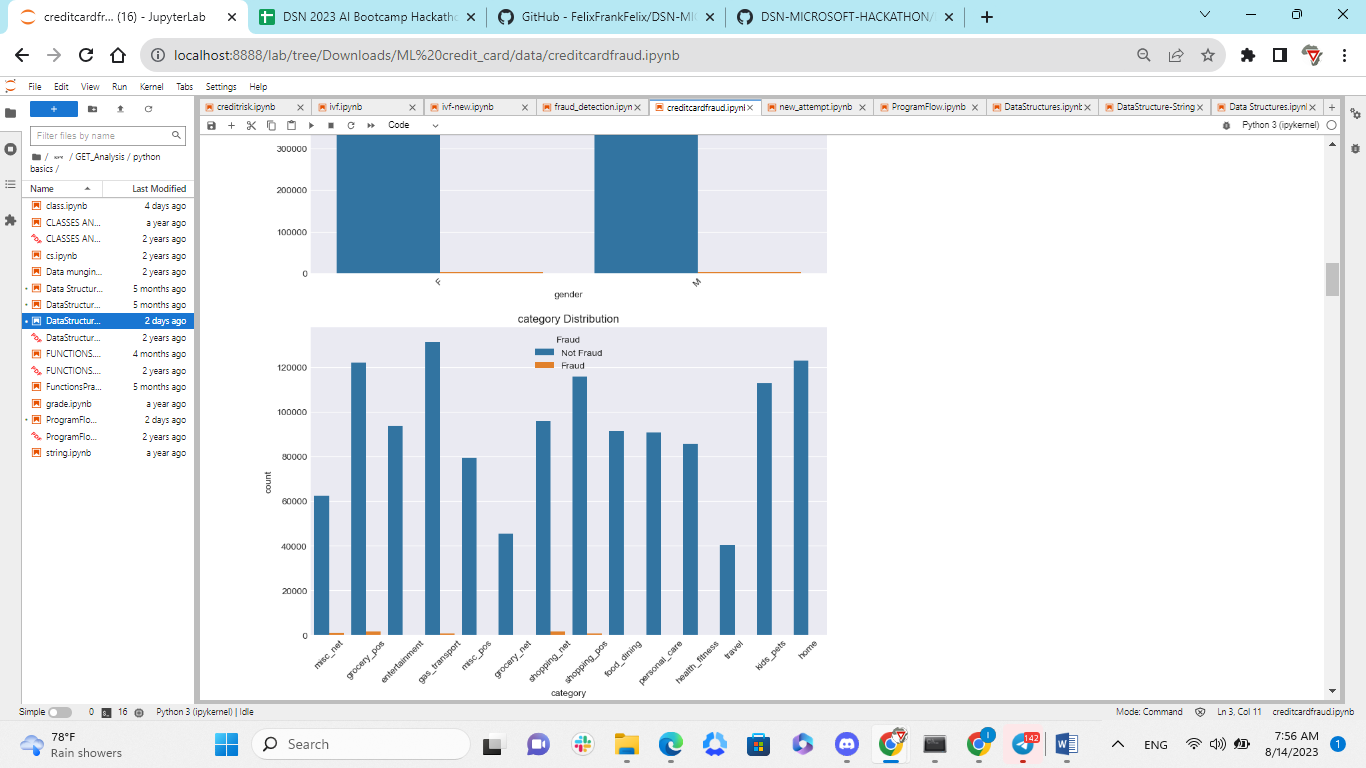
**Fig 6: Age distribution**

The result above shows that the age range of most people making most transactions are within the range of 30 to 50.



**Fig 7: Number of Transactions per Year-month**

The result above shows that are some portion of fraud detected over every months in 2019 and 2020, while a higher transactions recorded are not identify as a fraud.



**Fig 8: Category distribution over fraud**

The result above reveals that grocery\_pos, shopping\_net, shopping\_pos, misc\_net, and gas\_transport are the categories where fraud has been found, whereas others have no fraud and the majority of transactions in the 5 identified categories have no fraud.

**Data preprocessing**

**Feature Selection**

I initiated the data preprocessing phase by selecting a subset of features that were deemed relevant for credit card fraud detection. The features chosen included:

'category': Representing the category of each transaction.

'amt': Indicating the transaction amount.

'job': Reflecting the occupation of the individual involved in the transaction.

'merch\_lat' and 'merch\_long': Denoting the latitude and longitude of the merchant location.

The target variable, 'is\_fraud', was present to indicate whether a transaction was fraudulent or not.

**Label Encoding**

Categorical features such as 'category' and 'job' were originally in a non-numerical format. To make these features suitable for modeling, I applied label encoding. This process involved assigning a unique numerical label to each category. This conversion ensured that categorical variables could be effectively utilized as inputs for machine learning models.

**Train-Test Split**

To evaluate the model's performance on unseen data, I executed a train-test split on the dataset. Employing an 80-20 split ratio, 80% of the data was allocated to the training set, while the remaining 20% was assigned to the test set.

**Feature Scaling**

Recognizing the importance of ensuring consistent scales across features, I applied standard scaling to the numerical features. This process involved transforming the features to have a mean of 0 and a standard deviation of 1. By normalizing the features in this manner, I aimed to facilitate improved model performance and faster convergence during training.

**Handling Class Imbalance**

The dataset exhibited a noticeable class imbalance, with a significantly greater number of non-fraudulent transactions compared to fraudulent ones. To counter this challenge, I employed a resampling technique. Specifically, I employed the RandomOverSampler technique to oversample the minority class, which in this case was the class representing fraudulent transactions. This oversampling created a balanced training set and addressed the potential bias that could arise from the class imbalance. The objective was to enhance the model's ability to accurately detect instances of fraud.

**Finalized Training Set**

Upon completing the preprocessing steps, I derived the preprocessed training set, which I designated as 'X\_train\_final'. This training set incorporated both the original data associated with fraudulent transactions and the oversampled data, ensuring that the model was trained on a balanced dataset. Correspondingly, I adjusted the target variable, 'y\_train\_final', to align with the oversampled training data.

**Bivariate vs. Multivariate Analysis:**

Due to the complex nature of the relationships between the various factors that can impact the credit card evaluation as a whole, multivariate analysis is the method of choice for this prediction model. It is possible that the intricate interactions between the variables will be missed by bivariate analysis, resulting in less-than-ideal prediction performance.

**Model building**

**Choice of Learning Algorithms**

Supervised learning techniques are suitable because the problem has been labelled and classification is required.

Using cross-validation, we compared the efficacy of three distinct models: Random Forest, Support Vector Machine (SVM), and Logistic Regression.

Since it can accommodate intricate connections and possible feature interactions, the Random Forest model served as our primary model of choice.

GridSearchCV was used to fine-tune the Random Forest model's hyperparameters in order to get the highest possible F1 score, a value that takes into account both accuracy and recall.

Based on the F1 score, the top-performing model was chosen for further analysis.

**Analytical Evaluation of Chosen Solution:**

We compared the accuracy scores of all models as an extra performance parameter before making our final model selection using cross-validation and model scores.

Taking into account both precision and recall, the model with the highest accuracy score was selected as the best model.

The chosen model was trained using all of the training data and then tested using data that was never before seen.

**Less Suitable Learning Algorithm:**

Unsupervised methods like clustering may not work for this predictive model. Clustering groups data points without considering established classifications, hence it does not directly handle the binary classification problem of credit risk prediction. Clustering would not offer credit card evaluation labels and could lead to inaccurate conclusions.

**Model performance metrics**

**F1 score**

The F1 score measures precision-recall balance. In imbalanced datasets, where one class outnumbers the other, it is very useful. The F1 score evaluates a model's performance in credit card fraud detection, where non-fraudulent transactions often outnumber fraudulent ones.

F1 scores range from 1 (excellent precision and recall) to 0. It balances precision and recall to quantify a model's capacity to identify genuine positives while minimising false positives and false negatives.

**Classification Report**

A detailed classification report is generated, providing a comprehensive overview of the model's performance. This report includes metrics such as precision, recall, F1-score, and support for both classes. Precision signifies the proportion of correctly predicted positive instances, while recall indicates the proportion of actual positive instances that were correctly predicted by the best model.

**Model Result**

|  |  |
| --- | --- |
| Model | F1 score |
| Random Forest | 0.9608 |
| SVM | 0.8467 |
| Logistic Regression | 0.8371 |

Comparing the three models' F1 scores shows:

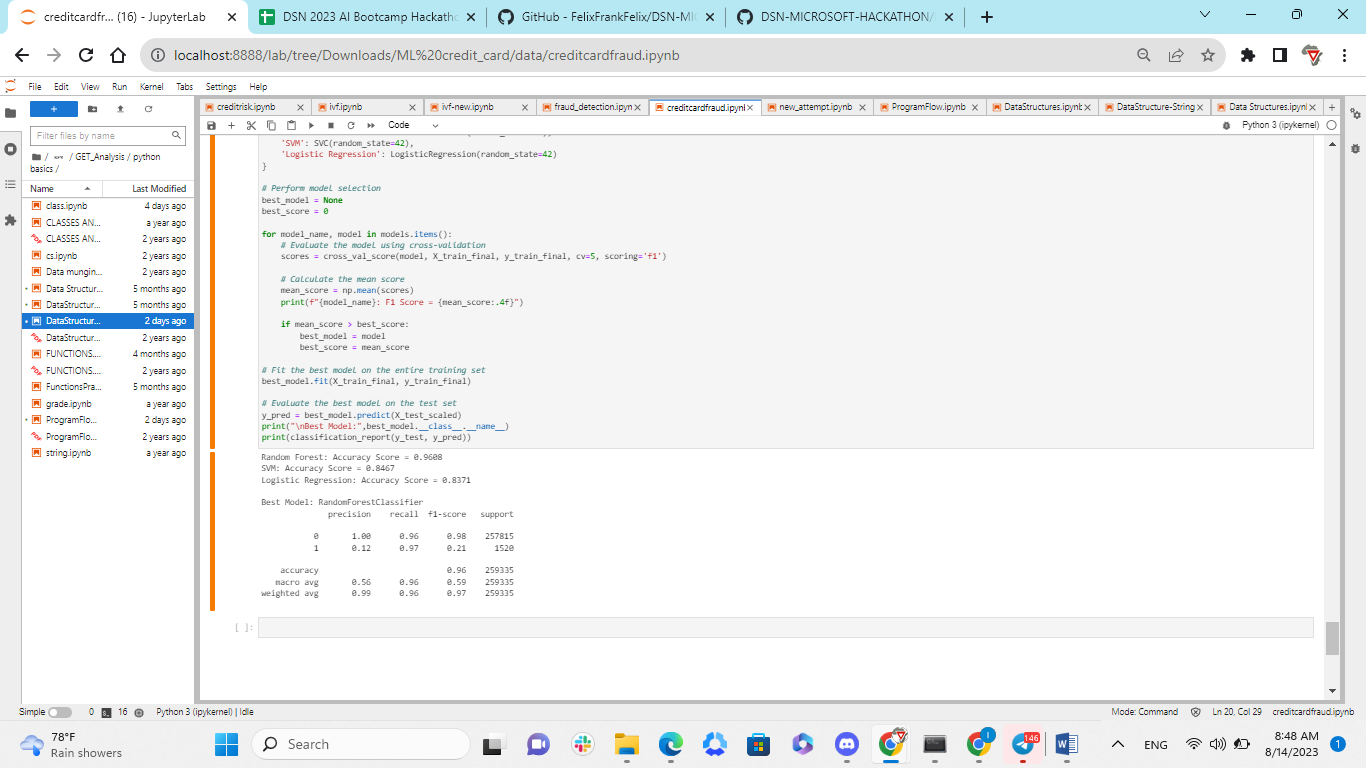
Random Forest: This vehicle scored the highest F1. A high F1 score indicates that the Random Forest model identified fraudulent and non-fraudulent transactions well. It balances precision and recall to accurately identify fraudulent transactions with few false positives and negatives. Since it detects credit card fraud best, the Random Forest model may be the best of the three.

SVM (F1 Score=0.8467): SVM scored F1. Though lower than the Random Forest model's F1 score, the SVM model can detect fraudulent transactions from non-fraudulent ones. SVM may offer slightly higher precision-recall trade-offs than Random Forest.

The Logistic Regression model had the lowest F1 score. Despite its excellent F1 score, the model may have had trouble spotting fraudulent transactions. The Random Forest model's precision-recall trade-offs may be better.

The F1 results indicate that the Random Forest model is the best credit card fraud detection model in this case. Its high F1 score suggests accurate fraud detection with low false positives and negatives. Real-world credit card fraud detection applications could benefit from the Random Forest model's effectiveness and robustness.

**Classification report**



The classification report for the RandomForestClassifier, the best model, details its credit card fraud detection ability. Interpreting metrics:

Precision: The model predicts non-fraudulent transactions with 1.00 precision. The bogus class (1) has just 0.12 precision. Only 12% of the model-predicted fraudulent transactions are fraudulent. Class 1 has low precision, suggesting the model labels some non-fraudulent transactions as fraudulent.

Recall (Sensitivity): The model properly identifies most non-fraudulent transactions with a recall of 0.96. The model captures a large percentage of fraudulent transactions in class (1) with a recall of 0.97. Class 1's strong recall shows that the model reduces false negatives by recognising most fraudulent transactions.

F1-Score: Class 0 has a 0.98 F1-score, indicating a good balance between precision and recall for non-fraudulent transactions. However, class 1's F1-score is 0.21, indicating that the model's precision and recall for fraudulent transactions is less effective. Class 1 has higher precision and lower recall.

Support: Support values indicate class occurrences. Class 0 (non-fraudulent) outnumbers class 1.

Accuracy: The model's accuracy is 0.96. Due to class imbalance, accuracy is high yet may be misleading. This accuracy is due to the model's fraud detection.

Macro and Weighted Avg: These measures average both classes. The macro average F1-score is 0.59, balancing the two classes. Considering class imbalance and class sizes, the weighted avg F1-score is 0.97.

The RandomForestClassifier accurately identifies valid transactions due to its high precision for non-fraudulent transactions. The model's low precision for fraudulent transactions suggests space for false positive reduction. The high recall for fraudulent transactions means that the model captures a considerable fraction of genuine fraud cases, while the overall F1-score of 0.21 for class 1 indicates that accuracy and recall need more optimisation.

Since the model excels at spotting non-fraudulent transactions, the class imbalance may exaggerate its accuracy. The model could be tuned to increase fraud detection and reduce false alarms.

The RandomForestClassifier's fraud detection performance can be improved by improving its precision for fraudulent transactions while keeping high recall.

**Conclusion**

Our research was all-encompassing, covering data pretreatment, feature selection, modelling, and evaluation as they pertained to the field of credit card fraud detection. Our analysis of the dataset's characteristics revealed distributions and patterns that are indicative of illicit financial dealings. We oversampled certain classes to compensate for underrepresentation and used label encoding to standardise the data before running the models.

By providing a strong model and a thorough understanding of the intricacies in data and approaches vital to accurate detection, our method contributes to the continuous fight against credit card fraud.

**Bibliography**

A Thennakoon, C Bhagyani, S Premadasa, S Mihiranga and N Kuruwitaarachchi. *Real-time credit card fraud detection using machine learning*. In *2019 9th International*

Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. Statistical Science, 17(3), 235-255.

Breiman, L., 2001. Random forests. Machine learning, 45, pp.5-32.

C Soulé-Dupuy, E Gaussier, M Lux, G Gianini, S Calabretto, M Granitzer, P.E Portier. *Credit Card Fraud Detection using Machine Learning with Integration of Contextual Knowledge* (Doctoral dissertation, INSA Lyon).2019

*Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 488-493). IEEE. , 2019

D Varmedja, M Karanovic, S Sladojevic, M Arsenovic, A Anderla. *Credit card fraud detection-machine learning methods*. In2019 18th International Symposium Infoteh-Jahorina (INFOTEH) 2019 Mar 20 (pp. 1-5). IEEE.

D.N Anowu, T Nyor, S.E Agbi, A.I Nelson, A.N Saliu. *Financial Forensic Analysis And Fraud Deterrence In Listed Deposit Money Banks In Nigeria*. **Gusau Journal of Accounting and Finance**. 2021 Oct 1;2(4):18

Dal Pozzolo, A., Caelen, O., Le Borgne, Y. A., Waterschoot, S., Bontempi, G. (2015). Learned lessons in credit card fraud detection from a practitioner perspective. Expert Systems with Applications, 42(10), 4643-4658.

G.K Singh, A Bhayye, S Dhamnaskar, S Patil and S.V Phulari. *Credit card fraud detection using isolation forest*. **International Journal of Recent Advances in Multidisciplinary Topics**, *2*(6), pp.118-119.2021.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of machine learning research, 3(Mar), 1157-1182.

H Nozari, M.E Sadeghi. *Artificial intelligence and Machine Learning for Real-world problems (A survey)*. **International Journal of Innovation in Engineering**. 2021 Oct 7;1(3):38-47)