TEESSIDE ADVANCED PRACTICE

ACTION PROJECT

Think Pacific foundation

To create a detection software that can be utilized to detect credit card fraud.

DENEDO, ELOHOR (Student)

C2108436

TEESSIDE UNIVERSITY

UNITED KINGDOM

Submitted 21st November, 2023

Table of Content

1. Introduction:

A credit card is a type of payment card that enables its user to make purchases up to their credit limit or make advance cash withdrawals. The convenience of being able to postpone or paying off debt until the following billing cycle is just one more perk of using a credit card.

With the increase in credit card transactions in Think Pacific Foundation website recently, the use of credit card safely on their website has become worrisome and of concern. Machine learning models are widely used nowadays due to their ability to process vast amounts of data and make accurate predictions. This project aims to predict credit card frauds in Think Pacific website to forestall fraudulent activities. This project aims to compare the performance of three machine learning models, Decision Tree, Logistic Regression and random Forest. The report will evaluate the performance of the models and compare their accuracy.

2. Existing work:

Methods such as Artificial Neural Networks, Clustering Techniques, Genetic Algorithms, Decision Trees, Neural Networks, and Support Vector Machines are now utilised to detect this type of fraud (Chandorkar, 2022).

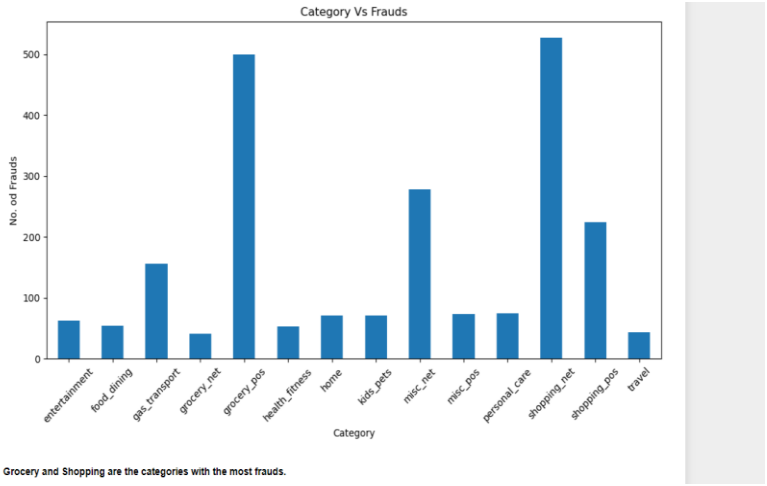
3. Methodology:

a. About the Data:

The necessary libraries were imported and the dataset was imported too. The credit card fraud dataset from Kaggle https://www.kaggle.com/code/renjithmadhavan/credit-card-fraud-detection-using-python/notebook was given but I could not relate to the features as most of them were hidden, probably due to the sensitivity of the information on credit cards. After discussion with my mentor, I used the below dataset https://www.kaggle.com/datasets/kartik2112/fraud-detection/data which is a simulated credit card transaction dataset . It comprises genuine and fraudulent transactions from the duration 1st Jan 2019 - 31st Dec 2020.

b. Exploratory data analysis

- Exploratory data analysis (EDA) is an essential pre-processing phase for gaining insight into the data and spotting trends, outliers, and other irregularities.



- **Grocery and Shopping are the categories with the most frauds.**

Fig 3: Distribution of Fraud by category

4. Data Pre-processing

In our analysis, we began by cleaning and preparing the credit card fraud dataset. The dataset contained no missing values, but the dataset was imbalanced and so the dataset was balanced by resampling the minority class. This was done to mitigate against bias towards the majority class.

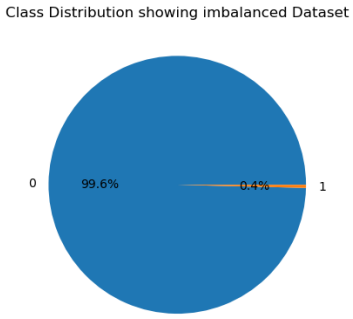


Figure 4.1 : Distribution showing imbalance in dataset

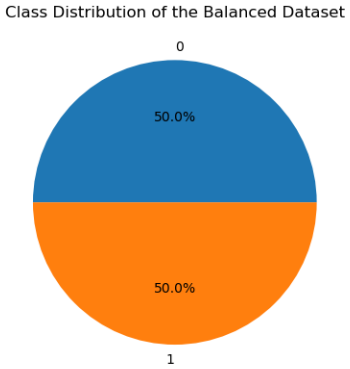


Figure 4.1 : Distribution showing balanced dataset

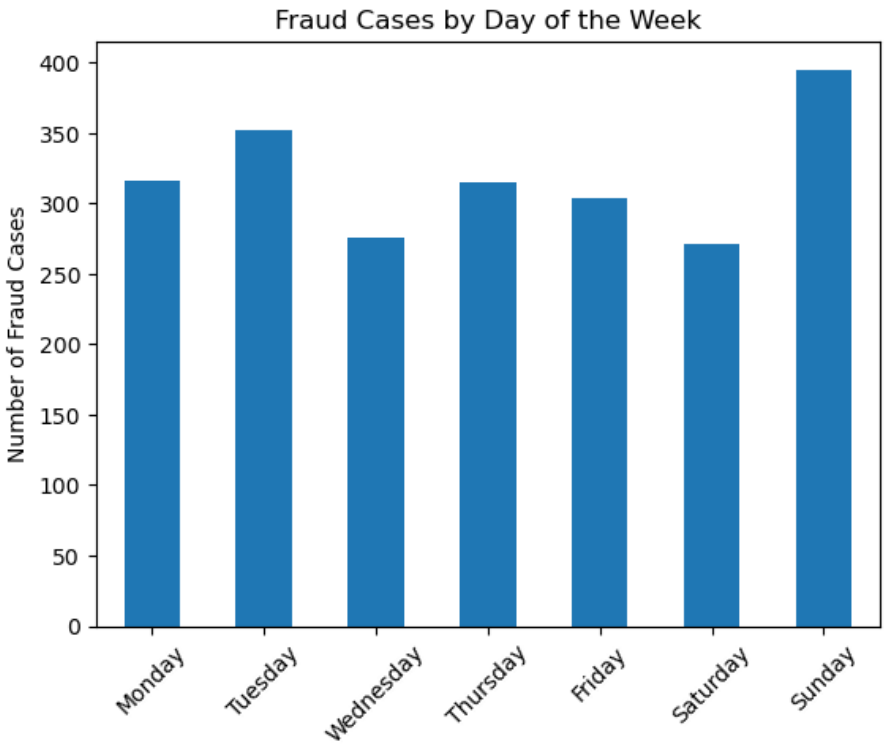


Figure 5 : Fraud cases by days of the week

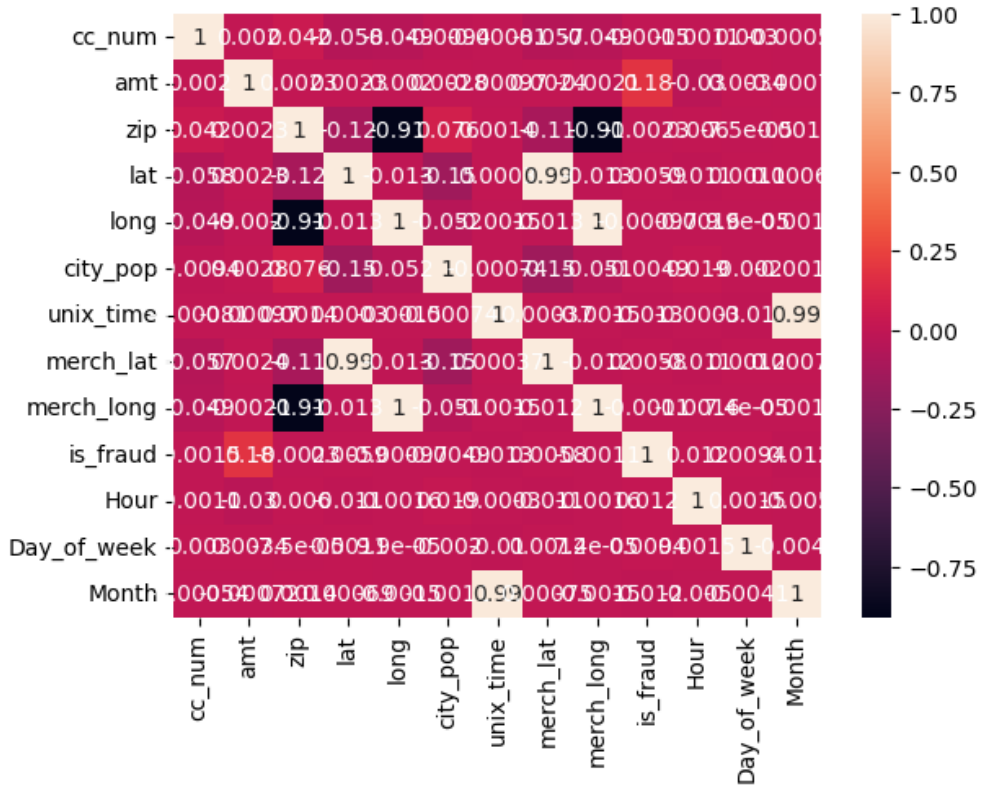


Fig 1: Correlation Heat Map of original data

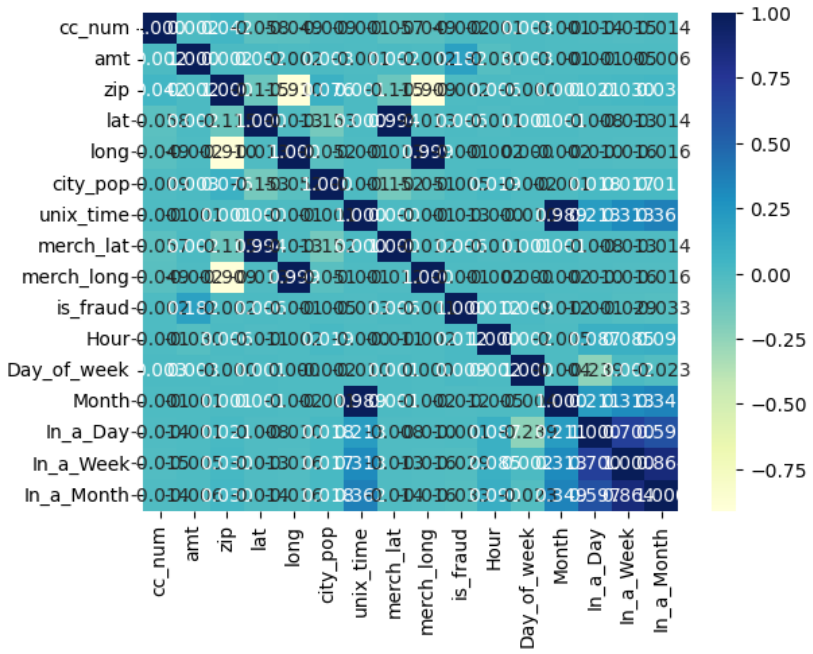


Fig 2: Correlation Heat Map of data with added features

- Correlation analysis: Pearson's correlation coefficient and other similar measures of correlation can shed light on the nature and direction of a link between two variables. As can be seen in fig. 1 and 2 above, a heat map was utilised to illustrate the connections between the variables.

5. Model Development and Evaluation

After the data was cleaned and prepared, it was divided into a training set and a test set. The models were trained on the training set and then tested on the test set to determine how well they performed. In this research, we employed three different algorithms: decision tree, logistic regression and artificial neural network. The models' performance was evaluated using metrics such as accuracy, confusion matrix, classification report, ROC score, and feature importance.

i. Model Development

1. Decision Tree

The decision tree algorithm is a tree-based method that uses a set of rules to make predictions. It is a well-known machine learning method that has found success in a variety of domains, including the financial sector. (Hastie et al., 2009). The model's performance was evaluated using various metrics, including accuracy, confusion matrix, classification report, and ROC score.

2. Logistic Regression

Logistic regression predicts binary outcomes linearly. It has been widely used in finance to model various phenomena, including credit risk (Altman, 1968). The model's performance was evaluated using the same metrics as in Decision Tree.

3. *Random Forest*

The random forest algorithm is an ensemble learning method that combines multiple decision trees to make predictions. It has been widely used in various fields, including finance, due to its high accuracy and robustness (Breiman, 2001). In our study, we used the random forest algorithm to develop a model to detect credit card fraud. The model's performance was evaluated using various metrics, including accuracy, confusion matrix, classification report, ROC score, and feature importance.

6. Evaluation Metrics:

In this work, we utilized model evaluation as part of our methodology to predict credit card fraud using machine learning algorithms. We evaluated the performance of the developed models using metrics such as classification report, accuracy, confusion matrix, and ROC score.

1. *Classification Report*

Machine learning metrics like classification reports summarise a classification model's performance. It reports several metrics that are used to evaluate the predictive accuracy of a model, including precision, recall, F1-score, and support.

According to Geron (2019), precision and recall are commonly used metrics in machine learning classification models to evaluate predictive accuracy. These metrics are used to assess the number of true positives and the number of false positives, respectively. The author further explains that the F1-score provides a balanced measure of a model's performance, while support is the number of instances in each class.

2. *Accuracy Score*

Accuracy measures the proportion of correctly predicted instances out of the total instances. It is a commonly used metric for evaluating binary classification models, where the goal is to accurately predict the class labels of instances (Bishop, 2006). We used accuracy to evaluate the models' overall performance in predicting credit card fraud.

3. *Confusion Matrix*

The confusion matrix is a table that summarizes the predicted and actual class labels for a binary classification problem. It contains four elements: true positives, false positives, true negatives, and false negatives. It is useful for understanding the model's performance in terms of sensitivity, specificity, precision, and recall (Géron, 2017). In our study, we used the confusion matrix to evaluate the models' performance on loan default predictions.

4. *ROC Score*

ROC scores measure the true positive/false positive trade-off. It estimates the AUC by plotting the true positive rate against the false positive rate at different threshold levels. (Fawcett, 2006). The ROC score is useful for evaluating binary classification models when the class distribution is imbalanced (Géron, 2017). In our study, we used the ROC score to evaluate the models' performance on credit card fraud detection.

Overall, our study aimed to develop and evaluate models to detect and flag credit card fraud using different machine learning algorithms.

7. Result

1. Results:

LOGISTIC REGRESSION

	precision	recall	f1-score	support
0	1.00	1.00	1.00	230360
1	0.00	0.00	0.00	858
accuracy			1.00	231218
macro avg	0.50	0.50	0.50	231218
weighted avg	0.99	1.00	0.99	231218

Confusion Matrix:
[[230360 0]
[858 0]]
ROC AUC Score:
0.5

Decision Tree:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	230360
1	0.74	0.67	0.70	858
accuracy			1.00	231218
macro avg	0.87	0.83	0.85	231218
weighted avg	1.00	1.00	1.00	231218

Confusion Matrix:
[[230157 203]
[285 573]]
ROC AUC Score:
0.8334754692260337

Random Forest:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	230360
1	0.98	0.53	0.69	858
accuracy			1.00	231218

macro avg	0.99	0.76	0.84	231218
weighted avg	1.00	1.00	1.00	231218

Confusion Matrix:
[[230350 10]
[404 454]]
ROC AUC Score:
0.9826292919039056

a. Result Interpretation:

The results were used to evaluate the two models (Decision tree and Logistic regression) on the whether a transaction is seen to be genuine or flagged as fraudulent. The Precision and Recall scores are measures of how well the model performs on both classes, where Precision measures how many of the predicted positive (Fraudulent/genuine) are actually positive and Recall measures how many of the actual positive cases are correctly predicted by the model.

8. Summary and Conclusion:

This action project utilized Python machine learning algorithms to create a software that can detect and flag fraudulent transactions in a website. The dataset was highly imbalanced and so was balanced during pre-processing. The trained dataset was split into train and test sets with a test size of 0.2. Decision tree, Logistic regression and random forest models were applied to the data and evaluated with a classification report, accuracy, and ROC score. These findings can help Think Pacific in securing their websites and thereby flagging likely fraudulent credit card transactions.

9. Further Recommendations:

- Real-time Monitoring: Put in place a system to track incoming transactions in real-time. The model should check each new transaction for the possibility of fraud as it happens.
- Anomaly detection: Spot unusual or out-of-the-ordinary financial dealings. Mark transactions for further inquiry if they differ significantly from usual legitimate transactions.
- Feedback Loop: In order to keep up with ever-changing fraud tendencies, it is necessary to regularly update and retrain the model.
- Privacy and Security: Maintain the confidentiality of any sensitive credit card information stored in the system in accordance with applicable data protection laws.
- Deployment of the software: Software should be deployed in a safe and extensible environment, such a cloud-based system.
- User Interface: A user-friendly interface should be developed so that analysts and investigators may evaluate flagged transactions and take the necessary measures.
- Maintenance: Set aside funds to keep your fraud detection system up-to-date and running well. Keep yourself updated of new fraud trends and work to improve your models and policies on an ongoing basis.

10. Consideration of Professional, Ethical, and Legal Issues

1. Dataset Selection: We carefully selected a dataset that did not contain any personal or sensitive information to ensure data privacy and protection.
2. Algorithmic Fairness: We considered fairness, transparency, and accountability in our approach to algorithmic decision-making to avoid any potential bias and discrimination against marginalized groups.
3. Legal Compliance: We reviewed regulations and laws surrounding the use of machine learning algorithms in credit card fraud to ensure legal compliance.

References

Altman, E. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance, 23(4), 589-609.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

Chandorkar, A., 2022. Credit card fraud detection using machine learning. *Int Res J Moderniz Eng Technol Sci*, *4*, pp.42-50.

11. Appendix

2. Visualisation of Attributes