COMM054 – Project Plan

To What Extent Can a Credit-Card Transaction Be Identified as Fraudulent?

Abubakar Jauro Sani (6769836)
Adarsh Manoj (6532510)
Piyush Patel (6760850)
Ramandeep Singh (6730378)

Problem Definition

The number of credit card fraud transactions has three folded from 2011 to 2020 based on the report by Merchant Savvy, accumulating a loss of greater than $32.39 billion (Whatman, 2021). Credit card fraud is any fraud committed by stealing /using someone's money to buy goods or services under someone else's credit or debit cards or by unauthorized means. In 2018 alone, the UK suffered a loss of approximately £678 million through acts of credit card fraud (Whatman, 2021). This leads to a series of concerns amongst the Credit Card issuing companies to protect their customers from these frauds hence they need an efficient way to handle these – leading to the credit card fraud detection systems. These systems help these companies in answering key questions, such as:

- What percentage of credit card transactions are frauds?
- How efficient and accurate the strategy is?
- How much amount of money can be saved by implementing an efficient strategy?

Credit Card Fraud Detection is a method of detecting fraud by implementing complex algorithms and machine learning knowledge to identify patterns in data that can protect customer's money from falling into the wrong hands. Traditional techniques for fraud heavily rely on manually analysing patterns regarding time, geolocation, and the principles of fraud – this is a laborious task, which machine learning-enabled credit card fraud detection can aide in.

The aim is to determine an efficient way to detect whether a particular transaction is a fraud or not by choosing various strategies to maximize the efficiency and accuracy of our models. Our model will examine the transactions using Business Analytics techniques including but not limited to the use of machine learning models to determine the patterns amongst the data.

The dataset used for this investigation is available on Kaggle and Is the Machine Learning ULB Dataset. The dataset contains the transactions of European cardholders in September 2013 (www.kaggle.com, 2017).
The dataset has transactions of 2 days, containing 284,807 transactions, of which 492 are labelled as fraud (www.kaggle.com, 2017).

The columns in the dataset are, V1 to V28, Time, Amount, and Class (Non-PCA values) due to data privacy issues. V1 to V28 are the numerical variables that are the result of PCA (Principal Component Analysis). The key variables will be the Time and Amount, due to investigation into industry standard practices (Chuprina, 2021).

The table below indicates the metadata for the dataset (www.kaggle.com, 2017):

Columns	Description	No. of Columns
Time	Seconds Elapsed between each successive transaction. (Not transformed by PCA)	1
V1 to V28	PCA Components	28
Amount	Amount spent by cardholder (in euros) (Not transformed by PCA)	1
Class	Contains only two values: 0 -Non-Fraud Transaction 1-Fraud-Transaction (Not transformed by PCA)	1

Business Analytics Tasks

The business analytics tasks will follow the Cross Industry Standard Process for Data Mining (CRISP-DM) approach, CRISP-DM is a set of guide rails that helps plan, organize, and implement a machine learning project (knowledgeburrow.com, n.d.). The key phases include Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment (Rodrigues, 2020). Using this knowledge, an investigation into the potential links, correlations, and patterns in the data.

Below is a breakdown of the BA tasks in series, to implement the desired credit card fraud detection system:

1. Set up the environment by importing data and libraries.
2. Explore the dataset to see its variables and their relationships.
3. Clean data of any missing values, outliers, and duplicates.
4. Implement various models for the purpose of analysis.
5. Address class imbalance to avoid overfitting.
6. Split data using the train test method and cross-validate.
7. Reduce dimensionality using the best-suiting method.
8. Implement the logistic regression machine learning model.
9. Hyper tune the parameters of the model to achieve ideal performance
10. Evaluate model performance.

The machine learning models that will be used to carry out this task include those in supervised learning logistic Regression, Decision Tree, and Random Forest. Logistic regression is a method used to predict a dependent variable given a set of independent variables such that the dependent variable is categorical (e.g., binary) (Brownlee, 2016). Decision Tree is a model where data is continuously split according to certain parameters it can be utilizes both for classification and regression task (IBM Cloud Education, n.d.). Random Forest combines the output of multiple decision trees to reach a single result (IBM Cloud Education, 2020).

Expectations

Through various implementations of neural networks, and establishing a CRISP-DM approach, ultimately a Credit Card Fraud detection system will be implemented. To evaluate, and understand the performances of the various models, key decisions must be made to identify a suitable metric for evaluating performance. The options available include, but aren't limited to – MAE, RMSE, F1 Scores. The key metric chosen is the Area under the Receiver-Operating Characteristics (AUC-ROC), which is a widely used metric for classification and regression tasks (Srivastava, 2019).

Project Plan

To structure and organize the project, the project management approach implemented was a Gantt chart, which helps to monitor and organize the structure of the project as it's being developed.

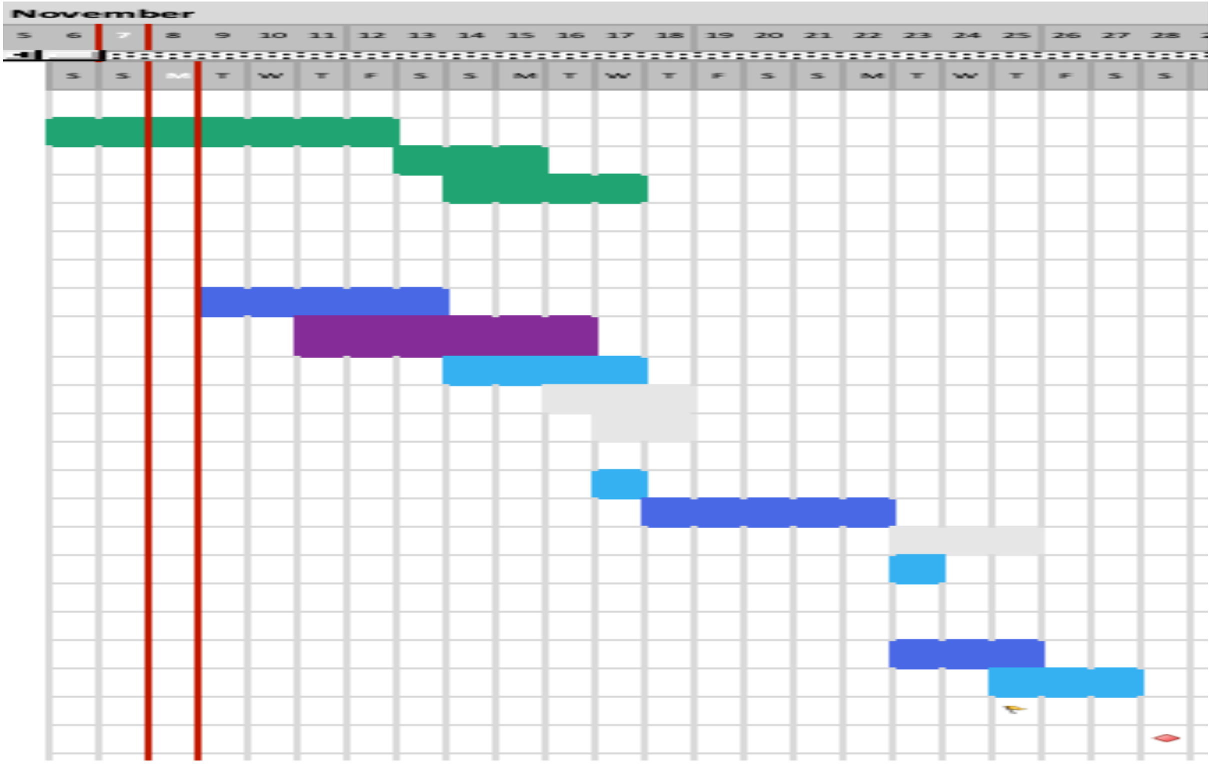
Gantt Table

On Track: item under consideration has been fully planned and allocated for.
Low Risk: task at hand hasn't been fully planned for but seems feasible.
Med Risk: task at hand hasn't been planned for and seems challenging.
High Risk: task at hand hasn't been planned for and seems considerably more challenging.
Report: Task at hand, is concerning only the report.

Milestone description	Category	Assigned to	Progress	Start	Days
Planning					
Discuss and choose working datasets	On Track	All	100%	05/11/2022	7
Identify steps to complete project plan	On Track	Abubakar, Adarsh	85%	12/11/2022	3
Discuss and prepare report structure	On Track	All	50%	13/11/2022	4
Data Exploration and Pre-processing					
Explore dataset and identify patterns	Med Risk	All	60%	08/11/2022	5
Using identified knowledge, generate base model assumptions	High Risk	Adarsh, Piyush	10%	10/11/2022	6
Pre-process and split the dataset	Low Risk	Abubakar, Piyush	0%	13/11/2022	4
Describe in the report exploration, pre-processing steps, and findings	Report	Adarsh, Raman	0%	15/11/2022	3
Identify suitable metrics for analysis based on initial model designs	Report	Piyush, Raman	0%	16/11/2022	2
Modelling					
Start instantiating the model	Low Risk	All		16/11/2022	1
Achieve optimum model performance	Med Risk	All		17/11/2022	5

Describe in report the model summary and architecture.	Report	Abubakar, Raman	22/11/2022	3
Ensure the code architecture and commentary are up to standard.	Low Risk	Adarsh, Piyush	22/11/2022	1
Evaluation				
Gather data for evaluating model performance	Med Risk	Piyush, Raman	22/11/2022	3
Plot, explore and analyse results	Low Risk	Piyush, Raman	24/11/2022	3
Describe and analyse findings in report.	Report	Abubakar, Adarsh	24/11/2022	1
Finalise report	Report	All	27/11/2022	1

Gantt Chart



Reference list

Brownlee, J. (2016). *Logistic Regression for Machine Learning*. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/logistic-regression-for-machine-learning/.

Chuprina, R. (2021). *Credit Card Fraud Detection: Top ML Solutions in 2021 - SPD Group Blog*. [online] Full-cycle Software Development Solutions. Available at: https://spd.group/machine-learning/credit-card-fraud-detection [Accessed 7 Nov. 2022].

IBM Cloud Education (n.d.). *What is a Decision Tree | IBM*. [online] www.ibm.com. Available at: https://www.ibm.com/uk-en/topics/decision-trees.

IBM Cloud Education (2020). *What is Random Forest?* [online] www.ibm.com. Available at: https://www.ibm.com/cloud/learn/random-forest.

knowledgeburrow.com. (n.d.). *What is CRISP-DM model? – KnowledgeBurrow.com*. [online] Available at: https://knowledgeburrow.com/what-is-crisp-dm-model [Accessed 7 Nov. 2022].

Rodrigues, I. (2020). *CRISP-DM methodology leader in data mining and big data*. [online] Medium. Available at: https://towardsdatascience.com/crisp-dm-methodology-leader-in-data-mining-and-big-data-467efd3d3781.

Srivastava, T. (2019). *Evaluation Metrics Machine Learning*. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/.

Whatman, P. (2021). *Credit card statistics 2021: 65+ facts for Europe, UK, and US*. [online] blog.spendesk.com. Available at: https://blog.spendesk.com/en/credit-card-statistics.

www.inscribe.ai. (n.d.). *Credit Card Fraud Detection: Everything You Need to Know*. [online] Available at: https://www.inscribe.ai/fraud-detection/credit-fraud-detection [Accessed 7 Nov. 2022].

www.kaggle.com. (2017). *Credit Card Fraud Detection*. [online] Available at: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud.