Date of Entry: 3rd November

What I have worked on:

I worked on finding viable datasets on Kaggle for my project, which was to answer the research question of in which ways can we best detect fraudulent transactions on blockchain networks. Ultimately, I found 2 suitable datasets on Kaggle that provided labelled data of fraudulent transactions on Ethereum networks with other characteristics of the transaction.

What problems I have encountered:

It was quite challenging to find suitable datasets due to 2 key factors. Firstly, datasets surrounding fraudulent transactions on blockchain networks are sparse. Secondly, very few of these blockchain datasets actually contained labelled data, which was highly critical as unlabelled data would allow only for unsupervised learning which would only provide meaningless clusters.

What I learned:

I learned key differences between supervised and unsupervised learning and how in the context of my project, only supervised learning made sense.

What resources did I use:

https://www.kaggle.com/datasets/vagifa/ethereum-frauddetection-dataset

https://www.kaggle.com/datasets/gescobero/ethereum-fraud-dataset

Date of Entry: 4th November

What I have worked on:

I worked on planning the full pipeline of the project before working on any of the code. As a result, I have structured notes of my pipeline with the key steps and wrote down roughly what I intended to do in each step, and what the purpose of those steps were. This will help guide my work down the line and allow me to remain on the same track despite working on the project at different times across the next few weeks.

What problems I have encountered:

Initially, I had immediately started out coding the preliminary steps of the project. However, I realised that my work was rather messy and unstructured, often missing out certain steps along the way.

What I learned:

I learnt the importance of planning and designing the full pipeline before delving into the code as it ensures a more clearly structured process, as well as allowing for more robust practices.

What resources did I use:

NIL

Date of Entry: 5th November

What I have worked on:

I worked on the preliminary data exploration, finding out key variables that display heterogeneous characteristics between the fraudulent and non-fraudulent transactions, and using that to inform me which features I should retain in the training dataset. Through this process, the retained features gives the model a rich yet manageable feature space while eliminating unnecessary or sparsely populated columns thereby reducing dimensionality, mitigates overfitting and speeds up downstream model training without sacrificing predictive signal.

What problems I have encountered:

My initial approach to the exploration was to merely identify if any of the variables had significant missing data and were thus inappropriate for inclusion. On top of that, I was intending to choose features based on logical contextual reasoning of which factors are most likely to be related to fraudulent behaviours.

What I learned:

Although I still believe this is important, I realised something equally as important was using the dataset to identify what characteristics differ vastly between fraudulent and non fraudulent transactions. In this way, we are both applying contextual knowledge to enhance the performance of the eventual machine learning model, and also using the dataset to draw key insights.

What resources did I use:

NIL

Date of Entry: 7th November

What I have worked on:

I worked on the data pre-processing steps of my project.

What problems I have encountered:

Whilst undergoing the data pre-processing steps, I noticed that the dataset had no null values which was convenient as I did not have to do any imputation. However, upon inspecting the dataset, I noticed that most features had extreme outliers and thus how they are handled would have a significant impact on the efficacy of the model training later on.

What I learned:

I went to find out more about the 3 main ways to handle outliers, namely removing them, imputing them with mean values and winsorization. Ultimately, I felt that winsorization was the most appropriate way to handle the outliers in this case. In short, removing the outliers reduces the sample size which was something I wanted to avoid, whilst mean imputation was deemed inappropriate as I felt these values are not measurement errors, thus I wanted to retain the heterogeneity through winsorization.

What resources did I use:

https://medium.com/@heysan/understanding-and-handling-outliers-in-data-analysis-727a768650fe

https://cxl.com/blog/outliers/#h-3-change-the-value-of-outliers

Date of Entry: 9th November

What I have worked on:

I worked on feature engineering.

What problems I have encountered:

Although I did not encounter any problems, my main consideration of this step was to engineer only relevant features and ensuring that I was not overdoing it

What I learned:

NIL

What resources did I use:

NIL

Date of Entry: 11th November

What I have worked on:

I worked on the data validation of the project, ensuring that the dataset is now appropriate to use for model training.

What problems I have encountered:

After winsorizing the dataset at the 5th and 95th percentiles and doing a basic inspection of the min/max values of the pre- and post-winsorized datasets, I noticed that the outliers seemed to be a lot less significant. However, upon doing visual analysis through the use of distributions (histograms) and boxplots, I noticed that the outliers were indeed still significant. Thus, I adjusted the winsorization to be at the 10th and 90th percentiles instead to deal with the significant outliers which helped significantly.

What I learned:

Inspections through min/max values are often not very informative and it is best to do a visual inspection through plots to identify if outliers have been handled appropriately.

What resources did I use:

NIL

Date of Entry: 14th November

What I have worked on:

Training the models on both the balanced and unbalanced train sets

What problems I have encountered:

After I had trained the models, I encountered a significant issue where my Random Forest and XGBoost models were giving near perfect results. This was quite puzzling at first as I was certain that this should not happen, so I went to look through my code again. However, I kept looking over my code and could not find an issue with my pipeline, checking that my steps were sequentially sound and did not result in any form of data leakage during data cleaning and feature engineering.

What I learned:

It was only through inspecting the dataframes individually where I found that the problem lies in my conversion of the dataframe to a CSV file for export. This resulted in a sorted indexing which served as a perfect indicator for fraud transactions. This problem was more noticeable in the Random Forest and XGBoost model as compared to the Logistic Regression as Logistic Regression is a linear model, while the other two split the samples on thresholds.

What resources did I use:

NIL

Date of Entry: 17th November

What I have worked on:

I worked on the hyper parameter tuning of the models, as well as the subsequent evaluation of the models. After hyper parameter tuning, I used the best parameters of each model found on the validation set and evaluated their individual performance. Based on their performance on the validation set, I then chose the best model and ran it on the test set to get the final results

What problems I have encountered:

This portion was rather smooth besides the run-time issues. Originally, I wanted to do an exhaustive Grid Search but there were simply too many parameters and iterations that made it difficult to run the tuning locally. Thus, I swapped over to a Random Search with the same parameters but limited each model to 4 folds for 25 candidates, totalling to 100 fits.

What I learned:

The tuning can take a significant amount of time despite the data set in this project being relatively small.

What resources did I use:

NIL

Date of Entry: 20th November

What I have worked on:

Project Description

What problems I have encountered:

NIL

What I learned:

NIL

What resources did I use:

NIL