**Topic selection**

From the 3 topics that were brainstormed in the first week of this project, which were for :

1. Heart attack risk prediction
2. Credit Card fraud detection
3. Covid-19 modelling for prevention

I have decided to pursue the topic 2, in credit card fraud detection because I find the topics interesting as am currently in the same field so I could better relate to the topic and was also curious to put my new learnt data science skillset to solve work real world issues.

**Data Collection and Preprocessing**

Finding the right data set was a challenge for me as I really wanted to use actual data set from my current company but was rejected from extrapolating customers cards firstly due to PII reasons and secondly because the project itself required a large volume of card numbers (minimum of 10,000).

Instead, I searched and obtained the dataset from Kaggle was able to obtain a dataset however the data set was anonymized and it was difficult understand what each row meant and this seemed to be a common issue when searching for the data set because most Credit card companies may not release detailed transaction data due to privacy concerns, thus making it difficult to understand each data means. So what I did instead is making understanding of what the number spoke as and then used the Logistic regression to then used the findings of the values to describe the data instead.

The next challenge I had to deal with was the fact that the data was unbalanced as Credit card fraud datasets often have high class imbalance because fraudulent transactions are relatively rare compared to legitimate transactions. Fraudulent transactions are a small proportion of all transactions, making it a rare event. As a result, the number of positive instances (fraudulent transactions) is significantly lower than the number of negative instances (legitimate transactions).As a result, datasets may be biased towards the majority class (legitimate transactions), which is easier and less expensive to label. The high class imbalance in credit card fraud datasets poses significant challenges to machine learning models.

**Data Cleaning And EDA**

The imbalance dataset posed a challenge as either Smote, Undersamping and oversampling were the options available to reduce the unbalanced dataset.I had utilized the SMOTE , Synthetic Minority Over-sampling Technique however It had very minimal impact as copy of SMOTE as follows below:

Hoewever post discussion with my trainer we had decided to remove the balancing as data set is balanced with a 0.5 split.

For the EDA I had decided to start with df.describe as the data set was ambigious so I wanted to see if there were any outliners where it calculated the mean, Std Dev, Min, max, and the quartiles( Q1- Q4).

Next I worked on plotting a heat map Correlations of V1- V28 and noticed that V16, V17 and V18 had the stronger positive correlations with V9, V10

Finally I looked at the Z scores for all the variables as z score is a statistical measurement that describes a data point's position relative to the mean of a group of values. It is expressed in terms of standard deviations from the mean. Z-scores are commonly used in statistics for various purposes such as identifying outliers in a dataset. Since the data set is ambigious I thought it would be a good idea to include it in the EDA as to pick up on any such outliers

**Data Modelling and Evaluation**

**Findings:**

1. Logistic Regression achieved an accuracy of ~96.49%.
2. Decision Tree achieved an accuracy of ~99.93%.
3. Random Forest achieved an accuracy of ~99.95%.
4. Neural Network achieved an accuracy of ~99.94%.
5. Logistic Regression False positive was 1270
6. Decision Tree False positive was 180
7. Random Forest False positive was 19
8. Neural Network False positive was 232

**Evaluation:**

* Logistic Regression: High accuracy but may suffer from overfitting due to imbalance.
* Decision Tree: Similar accuracy to Logistic Regression, interpretable but may overfit.
* Random Forest: Best performance, robust to overfitting.
* Neural Network**:** Good performance but requires more computational resources.

**Inference:**

* Random Forest model performed the best among all models tested.
* Neural Network also showed promising results but requires tuning for better performance.

**Deployment**

**Implementation:**

Integrate the model into the existing transaction processing system and picking up charges that are fraud upfront and notifying customers to confirm if charges are valid or fraud before taking necessary action of blocking card or keeping it active for usage.