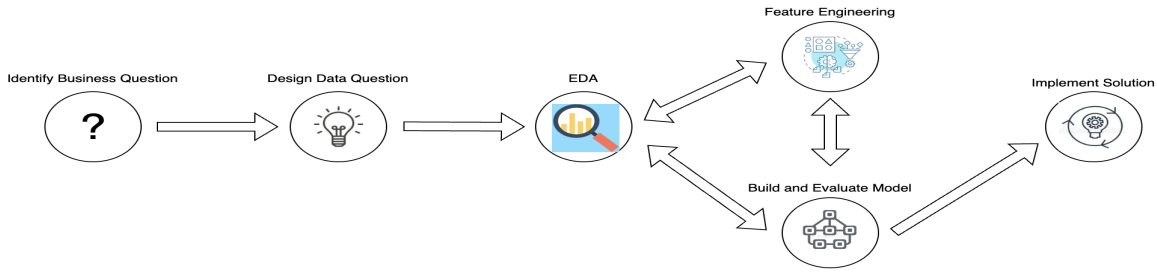
# Capstone Project Document
# Max Kim

## Process overview




## Problem statement

- What is the problem or the opportunity that the project is investigating?
    - Problem: Increase of credit card fraud affects credit card companies a lot of cost and time to deal with investigating the fraud, managing call-centre for customers and so on.
    - Opportunity: If detecting fraud transactions effectively, it will give credit card companies a lot of cost and time saving which leads to increase of their revenue and business performance.
- Why is this problem valuable to address?
    - This problem is valuable to address because, in 2020, there is about 393,000 credit card fraud reported to the Federal Trade Commission. This number shows a 44.7% increase from 2019 due to COVID-19. Increase of the fraud transaction meaning that there are more cases for credit card companies to deal with. This indicates that the amount that companies lose due to fraud is increasing but also the time and additional cost in it increase as well.
    - Lost $28.58 billion due to card fraud and the total amount loss from fraud is increasing every year. Projecting it to reach $49 billion in 2030. Being able to detect these fraud transactions effectively can be a valuable asset for the company as they can lose less money than before.
- What is the current state (e.g. unsatisfied customers, lost revenue)?
    - Currently a lot of credit card companies and banks are spending time and money to detect fraud effectively. They also have various ways to detect the fraud transaction.
    - As the number of credit card transactions are increasing, it is critical for companies to detect fraud as precisely and quickly as possible to minimise the loss. Every year, their losses are increasing.

- What is the desired state?
    - Desired state from this project is to build a machine learning model that can effectively detect fraud from credit card transactions. By implementing the model, the company can effectively and efficiently detect the fraud. Ultimately this will save cost and time for companies to prevent fraud to happen or can check with the cardholder when suspicious transactions are detected.

# Industry/ domain

- What is the industry/ domain?
    - The industry / domain is for credit card company and potentially can be used by banking
- What is the current state of this industry? (e.g. challenges from startups)
    - Credit card fraud is one of the most common forms of identity theft in 2022. Not only the small companies but also large credit card companies like VIsa or Mastercard do numerous methods to detect and prevent card fraud.
- What are the key concepts in the industry?
    - There are different concepts that help companies to detect fraud. One of the companies uses an anti-fraud detection system using AI and machine learning to analyse the transaction and produces a risk score to indicate the probability of the fraud. Also, there are card identification and security codes to prevent fraud, advanced identity verification to combat credit card fraud as well.
- Is the project relevant to other industries?
    - Although this project is relevant with credit card companies, this project can be used to apply in various different industries. Banking & financial services, IT companies and e-commerce can use this detection program so that they can detect the fraud and respond or prevent it from occurring.

# Stakeholders

- These people are key stakeholders for this problem. Since credit card company is spending lot of money and time dealing with fraud, those are the key stakeholders that we should persuade
- Who are the stakeholders? (be as specific as possible)
    - The main stakeholders for this project is the executive of the credit card company and head of the fraud department or managers in the fraud team.
- Why do they care about this problem?
    - These people are directly relevant to the fraud detection tasks and they know the most about it. As detecting and preventing fraud is one of the key KPI for credit card companies to concern, they will care about this problem seriously and will discuss how to improve their detection method to improve their performance.

- What are the stakeholders' expectations?
    - Stakeholders' expectation from this project is to be able to develop a model that can detect fraud precisely and quicker than the current system.
    - By implementing the solution, they are expecting to decrease their cost and time invested on fraud detection.

# Business question

- What is the main business question that needs to be answered?
    - In order to meet the stakeholder's expectations, our main business question that needs to be answered is "How to detect as many fraud transactions as possible to reduce the time and cost invested in the transactions?"
- What is the business value of answering this question? (quantify value and make necessary assumptions)
    - In 2020, due to credit card fraud, companies lost $28.58 billion worldwide. Assuming that business can build a model to improve the accuracy of the model or detection rate by 0.1%, that is already $0.3 billion dollars lost to be prevented. The actual cost saving for the company will be more since there are additional costs involved when dealing with lost money by card fraud.
    - There can be more to be concerned when quantifying the value of actual savings from improving the fraud detection system, the rough amount indicates that companies can save a lot of money by improving the model.
- What is the required accuracy? What are the implications of false positives or false negatives?
    - Compared to the number of normal transactions, the number of fraud transactions are very small. Thus, typically the data that will be used are imbalance data. Thus, the accuracy of the model is expected to be very high (0.95 and more). This is due to baseline accuracy of the dataset being high initially. Thus, not just the accuracy, but also needs to consider false positives and false negatives.
    - False positives indicate that the model predicted it is the fraud transaction where actually it was not. By having a high false positive rate, companies can save their time investigating the wrong transactions.
    - False negatives indicate that the model predicted it is a normal transaction where actually it was fraud. In this project, having high false negatives is crucial since wrong prediction is directly relevant to money loss to the company. Thus, when implementing the solution, it is required to have a high recall rate.

# Data question

- What is the data question that needs to be answered?
    - Data question that needs to be answered is “What model can be used to predict fraud transactions and what is the accuracy rate of it?”. By answering this data question, we will be able to answer the business question that we have come up in above.
- What is the data required to answer the question?
    - The required dataset will be the credit card company’s card transaction histories. The dataset may include the transaction time, amount, location, name of the store and so on.
    - The more details and more features available, it may help to build a model that has higher performance.

# Data

- Where was the data sourced?
    - The data sourced from kaggle call “Credit Card Fraud Detection”
    - https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud?datasetId=310&sortBy=voteCount
- What is the volume and attributes of the data?
    - This dataset has 284,807 transactions with 31 columns including target variables.
- How reliable is the data?
    - This dataset is coming from credit card transactions of European cardholders in 2013 September. This dataset has been collected and analysed during a research collaboration of Worldline and the Machine Learning Group of ULB. Thus, it is relatively reliable data.
- What is the quality of the raw data?
    - Quality of the raw data is well structured. The dataset has already been PCA transformed. Due to confidentiality issues, there are no details on original features or background information. However, since PCA was done, there was no row that has null value and entire columns are numerical.
- Is this data available on an ongoing basis?
    - This data will be available in the kaggle link above. However, the data will not be updated on an ongoing basis as the dataset is about two days of transactions in 2013 September.

# Data science process

## Data analysis

- This dataset has a total of 284,807 transactions with 492 fraud transactions and 284,315 non-fraud transactions. This ratio is 0.17% to 99.83%, indicating a very imbalanced dataset. Due to an imbalanced dataset, the first thing that needs to be done is to balance the dataset by using the SMOTE method.
- When doing raw data wrangling, since the dataset is already PCA transformed, it indicates that the dataset is already scaled. Apart from PCA transformed features, there are other features (amount, time) that are not scaled. These features has been scaled like other features.
- During the EDA process, the average amount for non-fraud transactions is $88.29 where fraud transactions were $122.21. This indicates that there are some relation between amount and fraud/non-fraud transactions. However, looking at 50% quantile, the non-fraud amount was $22 where fraud transaction was $9.25. This indicates that the transaction amount for fraud is skewed distributed then non-fraud.
- Looking at feature density for each feature, most features for non-fraud transactions seem to be distributed with 0 at the centre where fraud transactions are a bit skewed.
- This data pipeline is reusable to process the future data. However, this pipeline is most effective for reusing the PCA transformed datasets.

## Modelling

- What are the main features used?
    - All the features were used to build the model.
- Did you find any interesting interactions between features?
    - Between V1-V28 features, there is no noticeable correlation with each other. There are some negative correlation between Time vs V3 and Amount vs V2 as negative correlation. Positive correlation between Amount vs V7 & V 20.
- Is there a subset of features that would get a significant portion of your final performance? Which features?
    - Based on the feature importance of the model, amount V1, V4, V5 and V2 shows highest top 5 scores. This indicates that these are significant portions that affect predicting the fraud.
- What are the models used?
    - The models used in this project are logistic regression, K-NN, Random Forest, XGBoost and AdaBoost.
    - These models were used to work with binary classification models.
- How long does it take to train your model?
    - Depending on the model, the time taken to train the model was different. For logistic regression and K-NN, it took a relatively short period of between 5 to 10 minutes. For Random Forest, it took about 20 to 30 minutes, lastly for XGBoost and AdaBoost it took around 15 to 20 minutes to train the model.
- What are the model performance metrics?
    - KNN has accuracy of 0.99, precision of 0.71, recall of 0.77 and ROC AUC of 0.88.

    - Random Forest has accuracy of 0.99, precision of 0.87, recall of 0.78 and ROC AUC of 0.96
    - XGBoost has accuracy of 0.99, precision of 0.77, recall of 0.79 and ROC AUC of 0.977
    - AdaBoost has accuracy of 0.98, precision of 0.06, recall of 0.86 and ROC AUC of 0.96
    - For the selected model, logistic regression, it had accuracy of 0.98 with recall rate of 0.89. The model has ROC AUC of 0.98 which are all relatively high scores. However, it had a very low precision record of 0.06.
- Which model was selected?
    - The selected model is logistic regression. The background of the model selection was focusing on accuracy and recall rate. Logistic regression showed high scores for both factors out of 5 different models.

# Outcomes

- In conclusion, we have chosen logistic regression for our binary classification problem. We had a relatively good accuracy of 0.98 with 0.89 recall rate. Looking at the EDA and modelling, we will be able to conclude that credit card companies can focus on the amount of each transaction along with other key important features to detect fraud more effectively.
- Building and deploying a machine learning model that has focus on high recall will distinguish non-fraud and fraud transactions well. These models can continuously improve since the company has large number of transaction dataset to use for train the model.
- In this data science process, there can be further improvement with the model. For this logistic model, although precision is less important than recall or accuracy, having low precision score may cost extra money and time due to investigating transactions that are not actually fraud.
- Furthermore, we can develop our model in different process workflow stages. We can improve the data process by working on feature engineering more, apply deep learning and machine learning to build more precise and faster models.

# Implementation

- What are the considerations for implementing the model in production?
    - When implementing the model in production, things that need to be considered are that the model needs to be able to detect a wide range of frauds by using machine learning and A.I.
    - Also, when building the model, it is important to consider the out-of-box fraud as well. Only detecting the current existing fraud may not be enough. When building the solution, it will be crucial to build a model that can sufficiently prevent out-of-box fraud as well.

# Data answer

- Our data question was “What model can be used to predict fraud transactions and what is the accuracy rate of it?”. We have built several different models and compared the accuracy and recall rate. The Logistic Regression model was selected since it has high accuracy rate and recall rate. These data questions are answered with satisfaction, however, there is still room to improve the answer. Using different machine learning or deep learning methods can potentially provide a model with better accuracy and recall rate.

# Business answer

- Our business question was “How to detect as many fraud transactions as possible to reduce the time and cost invested in the transactions?”. By approaching this question by answering data question, we were able to answer this business question with some portion of the answer.
- Build a machine learning model to detect as many fraud transactions as possible. Having a well-built model to detect the fraud can help companies to reduce their time and cost on fraud investigation. The model that we have selected in this project does provide a good score, there can always be room for improvement when applying deep learning into the model.

# Response to stakeholders

- What are the overall message and recommendations to the stakeholders?
    - As time passes, there will be a continuous increase in the total fraud amount and it will be more important to have an effective and fast way to detect and prevent the fraud transaction. One of the ways to effectively and quickly detect fraud is by building a machine learning model that can use the credit card transaction history to detect the fraud automatically.
    - In this project, we have introduced some of the different models that can be used to detect fraud transactions. However, further improvement can be made in various aspects within workflow processes. Our recommendation is that by continuously updating the detecting model, they can achieve their goals of minimising the loss from fraud.

# End-to-end solution

- What is the overall end-to-end solution to use the model developed in the project?
    - There are several things to consider when using the model developed in the project. First, we need to gather the raw data and change the data structure to fit and be able to run with the model. Preferably to do the PCA transform like our dataset in the project. Once the data is wrangled and cleaned, we can

input the data into the developed model to get the answer. Once the model is run, what has been detected as fraud transactions can be sent to managers or fraud teams for the investigation. During this process, we can build a UI that shows statistical records such as total transactions, fraud detection score and details of transactions that model detect as a fraud.

- When developing the model, there are various libraries and tools used. Train_split_test to split the dataset into train and test data for learning. Sklearn.metrics to retrieve metrics scores to examine the performance of the model. Furthermore, sklearn.model_selection GridSearchCV to find the optimal parameters for each model.
- In order to implement the model as a solution, there are various different skills and efforts required. Technical skills to build the model, understand how each algorithm works are essential. Also, being able to understand the business context and dataset are crucial in order to build a model. It is very important to have skilled data science to be in a team when building models.

# References

- Where is the data and code used in the project? (show a simplified list of main items: notebooks, datasets, exported models)

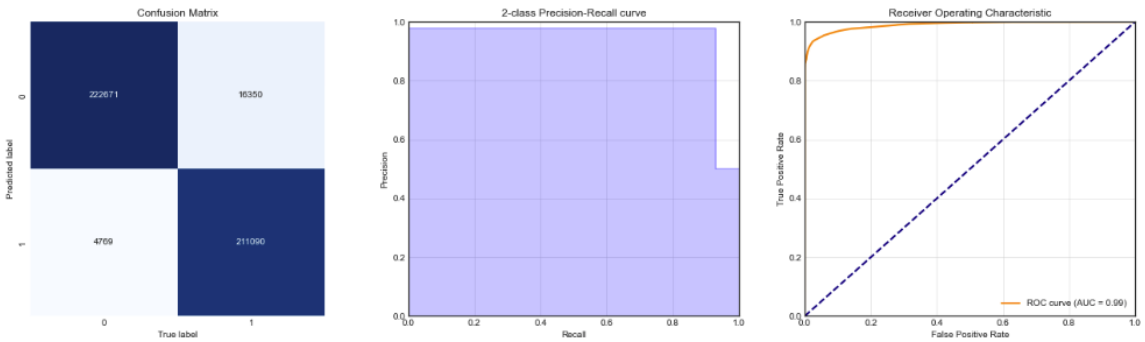
```
]: 1 pd.set_option('display.max_columns', None)
   2 data =pd.read_csv('/Users/maxkim/Desktop/DS Camp/Mini Project/Capstone Project/creditcard.csv')
   3 data.head()
```

]:

| | Time | V1 | V2 | V3 | V4 | V5 | V6 | V7 | V8 | V9 | V10 | V11 | V12 | V13 | V14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0.0 | -1.359807 | -0.072781 | 2.536347 | 1.378155 | -0.338321 | 0.462388 | 0.239599 | 0.098698 | 0.363787 | 0.090794 | -0.551600 | -0.617801 | -0.991390 | -0.311169 |
| **1** | 0.0 | 1.191857 | 0.266151 | 0.166480 | 0.448154 | 0.060018 | -0.082361 | -0.078803 | 0.085102 | -0.255425 | -0.166974 | 1.612727 | 1.065235 | 0.489095 | -0.143772 |
| **2** | 1.0 | -1.358354 | -1.340163 | 1.773209 | 0.379780 | -0.503198 | 1.800499 | 0.791461 | 0.247676 | -1.514654 | 0.207643 | 0.624501 | 0.066084 | 0.717293 | -0.165946 |
| **3** | 1.0 | -0.966272 | -0.185226 | 1.792993 | -0.863291 | -0.010309 | 1.247203 | 0.237609 | 0.377436 | -1.387024 | -0.054952 | -0.226487 | 0.178228 | 0.507757 | -0.287924 |
| **4** | 2.0 | -1.158233 | 0.877737 | 1.548718 | 0.403034 | -0.407193 | 0.095921 | 0.592941 | -0.270533 | 0.817739 | 0.753074 | -0.822843 | 0.538196 | 1.345852 | -1.119670 |

```
]: 1 lgr_y_train_pred =lgrGS.predict(X_train)
   2 lgr_y_train_prob =lgrGS.predict_proba(X_train)
   3
   4 show_summary_report(y_train, lgr_y_train_pred, lgr_y_train_prob, 'Logistic Regression (C =10)')
```

```
Accuracy : 0.9536 [TP / N] Proportion of predicted labels that match the true labels. Best: 1, Worst: 0
Precision: 0.9779 [TP / (TP + FP)] Not to label a negative sample as positive.       Best: 1, Worst: 0
Recall   : 0.9281 [TP / (TP + FN)] Find all the positive samples.                     Best: 1, Worst: 0
ROC AUC  : 0.9894                                                                    Best: 1, Worst: < 0.5
---------------------------------------------------------------------------------------------------------
TP: True Positives, FP: False Positives, TN: True Negatives, FN: False Negatives, N: Number of samples
```



- What are the resources used in the project? (libraries, algorithms, etc)
    - Numpy, pandas and matplotlib libraries were used mainly when working on Exploratory data analysis.
    - Various algorithms such as logisticregression, train_test_split, sklearn.ensemble and standardscaler when running the algorithms.
    - Sklearn.metrics is used to check and compare the performance of the model.