# **David Adese**

* Read a bunch of research papers regarding “Credit Card Fraud Detection” to identify potential areas of difficulty to look out for during development.
* Current progress includes:
  + Data Cleaning
    - Dropped features with over 70% of missing data
    - Dropped features with extremely high or extremely low correlation.
    - Imputed missing data for categorical features with “X”
    - Imputed missing data for numerical features with the mean value.
  + Data Transformation
    - Numerically represented categorical data using Label encoding
    - Split data into 3 sets: Training (60%), Validation (20%) & Testing (20%)
    - Performed feature selection using sklearn’s “*SelectKbest*”
    - Performed feature scaling using Standardization and Normalization (MinMax)
  + Data Reduction
    - Currently using Principal Components Analysis.
    - Yet to experiment with Linear Discriminant Analysis & Factor Analysis
  + Model Building
    - Hyper-parameter tuning using GridSearchCV
    - Still experimenting with classifiers such as gradient boosting, random forest etc.
  + Model Evaluation
    - Currently using Accuracy, Precision, F1 Score, Recall & Latency metrics.
    - Yet to use confusion matrices.

# **Lucas Dillistone**

* Researched Bayesian Classifiers, Neural Networks and Gradient Boosting Classifiers for the Literature review
* Worked on the literature review
* Wrote the Lucas(Notebook).ipynb to:
  + Import and assign the data to testing and training sets
  + Performed data exploration to find the amount and percentage of missing data from each set
  + Combine the transaction and identity datasets to view them as a whole
  + Encode categorical data as values to be used for feature selection
  + Create the correlation matrix of the combined dataset
  + Order and print the absolute value of the correlations with the IsFraud variable to find the most correlated columns

# 

# **Kevin Genereux**

* Configured Google Colab environment (first time using for a project)
* Determined how VSCode can be used as an IDE in Google Colab
* Determined how to integrate VSCode with git and create a repository
* Read through all research papers that were uploaded into the group folder to make significant contributions in writing the Proposal and Literature Review
* Read many Kaggle discussion posts (learning dataset characteristics, EDA, data cleaning, feature engineering, high performing models) to help establish a methodology for our project
* Researched all gradient boosting algorithms (XGBoost, AdaBoost, LightGBM, CatBoost) to gain an in-depth understanding of their mechanisms and characteristics
* Importing and reading dataset
* Conducted EDA (determining ratio of imbalanced dataset, types of features, missing value analysis)
* Replacing missing values with -9999, this was recommended in a discussion post to prevent the gradient
* Created an XGBoost model to test our preprocessing strategies
* Currently working on optimizing LGBM and CatBoost models
* Added metrics (AUC score, accuracy, sensitivity, specificity, F1 Score)
* Added feature engineering and feature selection techniques