**Data Source:** [**https://www.kaggle.com/datasets/sakshigoyal7/credit-card-customers**](https://www.kaggle.com/datasets/sakshigoyal7/credit-card-customers)

**Introduction**

Data Science has been a very useful tool in various industries including Finance and Business. Data Science can help businesses make diagnostic and predictive analyses that will help improve the productivity of the business in terms of customer growth and satisfaction. In finance, data science can help with fraud detection, customer satisfaction, etc. Data Science is a great tool that can help the finance industry look at past events and policies, the response of customers to those events and policies, and help find solutions to whatever problems might have been faced in history.

**Aims and Objectives**

Aim

The aim of this project is to build a model to implement customer churn prediction using machine learning, which will train a machine learning model on the available data and predict with a high accuracy which customers are about to churn, which in turn will help the managers in making useful marketing decisions.

Objectives

Churn prediction is the process of identifying which customers are most likely to stop using a service or to cancel their subscription. Because getting new customers frequently costs more than keeping existing ones, it is an important prediction for firms. The following steps are carried out;

* Understanding a problem and final goal
* Data Collection
* Data Wrangling
* Exploratory Data Analysis
* Modeling and testing
* Model deployment and monitoring

**Flow Process**

Data Gathering

This involved downloading the dataset used for this project from the Kaggle database.

Data Cleaning

This involved the use of the Pandas library to clean the data in order to make it fit for analysis.

Exploratory Analysis

This involved the use of the Pandas, Numpy, and Matplotlib libraries to investigate the data for insights, trends and patterns that might help with predictive analysis.

Model Training

This involved the use of the machine learning library to create a model that will be trained according to the dataset provided in order to become familiar with the insights, trends and patterns that were found during exploratory analysis.

Model Evaluation and Validation

This involved the evaluation of the model created for accuracy and precision in order to prevent or minimize underfitting, overfitting and other problems of a sub-optimal model.

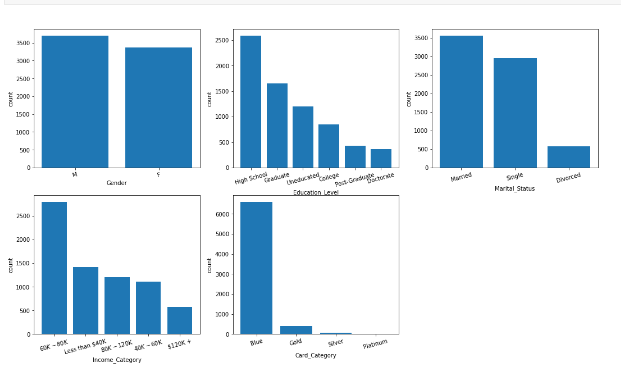
Model Deployment

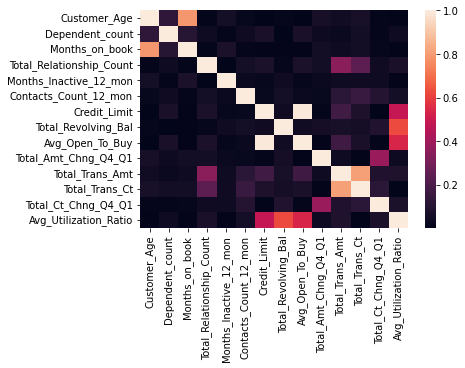
This involves making the model available for real world predictions.

Data Cleaning and Exploratory Analysis

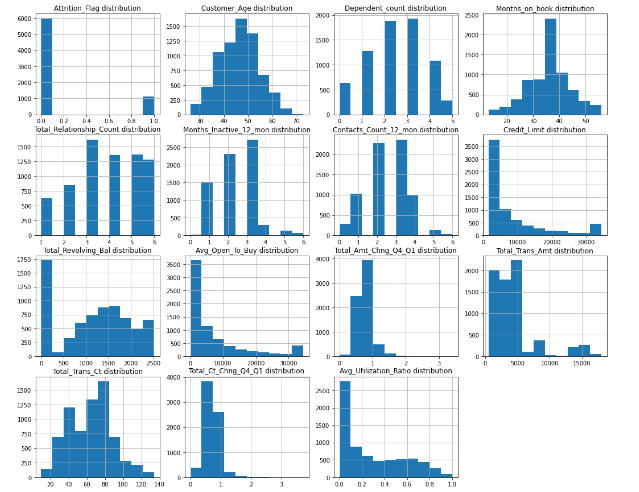
There are 10,127 entries in the dataset with 23 columns. Unnecessary columns were dropped from the data frame. The attrition flag column was transformed into a numerical column by replacing existing customers with zero (0) and attrited customers with one (1). Unknown entries were dropped to avoid uncertainty. After data cleaning, 7081 entries were left in the data frame with 20 columns.

A total of 1,113 customers are attrited customers, representing about 16% of all entries analyzed. In the image below, there is almost equal distribution between male and females. Most customers earn between 60 and 80K US dollars. Most customers are on the blue card category while the platinum card category has the least number of customers. A lot of the customers are graduates or married.



From the heatmap below, there is a strong correlation between two variables, *Credit Limit and Average Open to Buy*. 

From the image below, the modal age is seen to be between 45 and 50 years with a symmetrical distribution. Most of the customers have spent between 35 and 40 months on book. The average utilization ratio for most customers is less than 0.2.



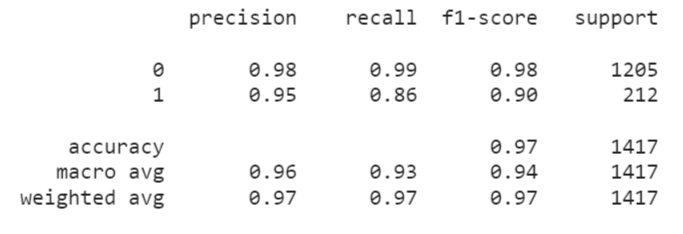
**Model Training, Evaluation and Validation**

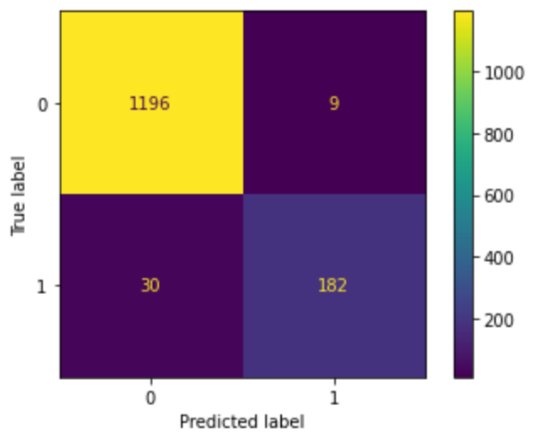
We analyze and used different baseline models such as SVC, LogisticRegression, KNeighborsClassifier, RandomForestClassifier, ExtraTreeClassifier, LGBMClassifier, and XGBClassifier before dive into the full comparison of the model score.

 Based on the model score results, LGBMClassifier had the highest accuracy of 97.2477 %, being over 0.2824 % higher and faster than the next algorithm while being the best in terms of test accuracy. Its implementation is actually so fast that it can train the entire ensemble of trees faster than others.

After data cleaning, 7081 entries were left in the data frame with 20 columns. The train and test data were split out of these 7081 entries, having 5664 train and 1417 test dataset. The train dataset was used in the different baseline model listed above.

After the model evaluation, hyperparameter tuning was done to improve on the complexity of the model which had a recall score of 0.86 for LGBMClassifier model.

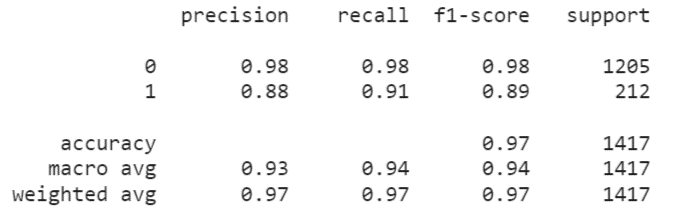


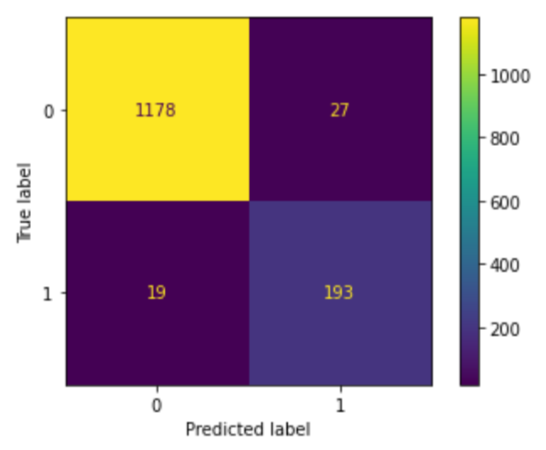


Confusion Matrix Display for **untuned** model

For this project, according to the aim, a recall score is more important than a precision score. This is because we are more interested in capturing all the customers that are more likely to churn/attrite, even though we incorrectly include those that are less likely to churn.

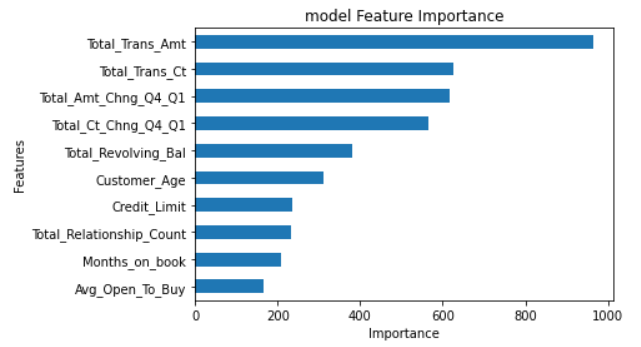
After tuning the hyperparameters, the recall score for our model (for the target/label) increased from 86% to 91% even though the precision score dropped from 95% to 88%





Confusion Matrix Display for **tuned** model

The bar chart below shows a list of the 10 most important features (in decending order from top to bottom).



**Model Deployment**

Sometimes, communication means sharing visualizations (as seen above). Other times, it means sharing the actual model.

The model was saved to a Pickle binary file (model-team-azure.pkl) that can be deployed in production

**Active Team Members**

1. Adedamola Adedoyin
2. Kenneth Mambo
3. Elijah Fagbohun
4. Bakare Olajumoke
5. Gyang Paul Pam
6. Ramesh Vhanamane