* Title
* Contributors (Team members)
* Background / Motivation
* Overview of the data set (You can include any preprocessing methods here)
* Models (If any was applied)
* Visualisation (either exploration or prediction)
* References

Stutern Graduate Accelerator 07. Data Science Task Group Project

Team Members: Ayooluwa Adedipe

Patience Kadiri

Kenneth Dagogo

Background

This project is meant to apply the theoretical concepts being learnt on a real-life financial data science problem. Data Science and machine learning have become indispensable tools in the financial services sector, especially banking. As the industry becomes more competitive, banks and fintech companies need to be able to handle their increasing number of customers. An important use case is in the allocation of loans.

Loans are critical to the success of any bank as they are the major source of income. However, if handled wrongly, loans are also likely to lead to the demise of the business when customers default. The predictive power of machine learning can be utilized in this use case to predict the likelihood of a customer defaulting on a loan and guide decision making in choosing who to grant a facility. This project has two phases, the first part involves building a model for predicting loan performance based on the data provided. For the second part, a frame work will be created for apportioning credit worthiness scores or rankings to customers to determine how eligible they are for a facility. This ranking system or scoring framework will also be developed based on information extracted from the dataset.

**Overview of The Dataset**

The raw datasets provided include:

* All\_demographics
* All\_investments
* All\_loans
* Bureau\_ scores
* All\_accounts details

These tables were combined in various ways for the two stages of analysis carried out her

**Creation of the Loan Classification and Credit Framework (Accompanying code can be found in R Script):**

For this section, each of the tables that constitute the feature variable were used to create a ranking system that would rate the eligibility of the customer to get a facility.

On the available balance table, for instance, each customer is rated on an increasing scale between 0 and 10 based on the amount of money they have in their account balanced. Based on this same idea, ratings are also generated for cashflow history (based on subtracting total outflow from a customer’s account from total inflow). Customers are also rated based on their bureau scores as well as the investments they hold with the bank. Finally, demographics such as sex and marital status are also incorporated into the ranking system to qualify customers.

A new data set is created from the ranked scores which can be used as feature variables to determine how qualified a customer is.

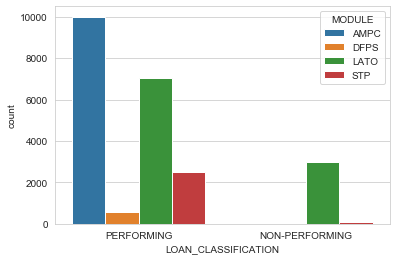
**Possible Practical Application**

For individual customers, a basic API can be developed which takes in necessary details about the customers and returns a credit score. Customer facing staff can therefore let customers know immediately whether or not they qualify for a loan

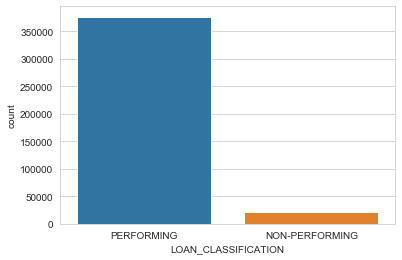
**Predictive Model Building (Accompanying Code is in the Python Jupyter notebook)**

To build the predictive model, information from 3 of the tables in the dataset was combined. The all\_loans application table was the most important for this stage as it contains the target variable (Performing and non-performing) and the CUSTOMER\_UNIQUE\_ID, which is the unique identifier that links the various pieces of information about specific customers across the tables. Information from the all\_loans and all\_demographics tables were also scraped and combined with the previously mentioned all\_loans\_application table to create a master data set containing useful feature variables which will be used to predict whether a loan is likely to perform or go bad.

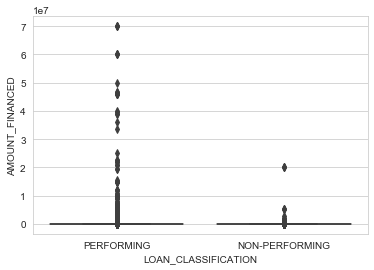
Below are some visualizations from the joined dataset.



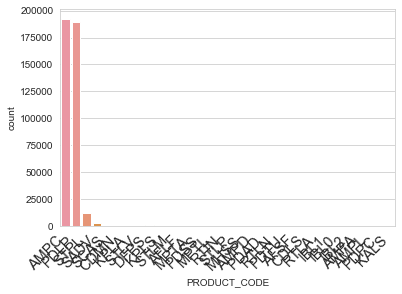
Apparently, the loan modules are not correlated to loan performance, so we can conveniently drop that column.



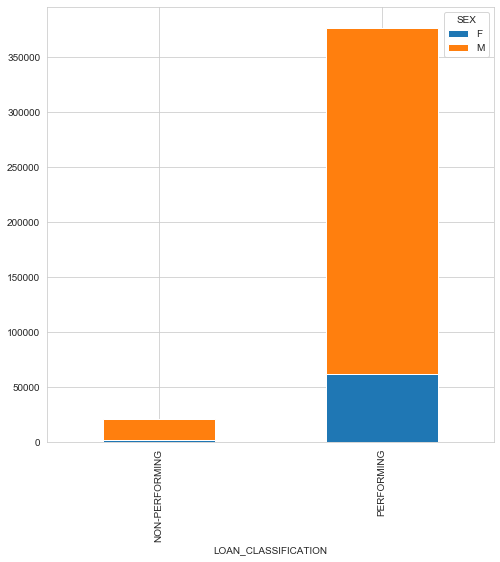
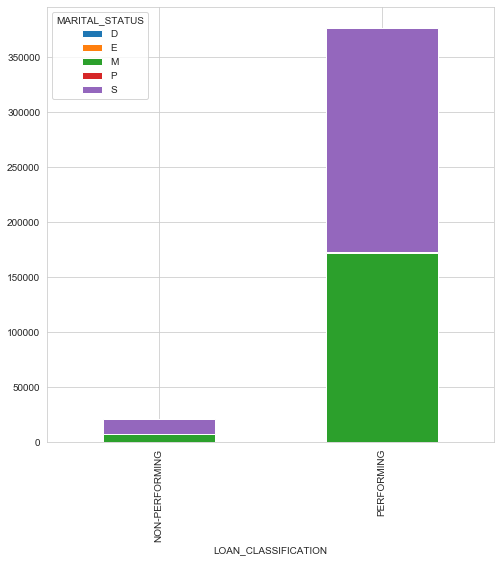
This bar chart compares the performing loans to non-performing and shows that most of the loans are performing as they should.



This boxplot combines the amount financed variable with loan classification and indicates that there are massive outliers in the amount financed, some of which are dropped to prevent skewing the prediction model.



The distribution across product codes shows that most of the loans come from only a few categories.



Also, the plots above show how the loan performance varies across sex and marital status

**Data Preprocessing**

The categorical variables were converted to zeros and ones using dummy variables. Also, before the final prediction was done, feature scaling was used to normalize the feature variables to make prediction more accurate.

**Prediction Models**

Two classification algorithms were implemented to make predictions on the data: A Random Forest Classifier and a Logistic Regression Model.

For the Random Forest Classifier (with number of estimators chosen as 10), a 96.5 % accuracy was gotten on the training data while the test data showed a 95.6% accuracy.

Also, the Logistic regression model showed an accuracy of 95% as well.

**Bias!**

It is important to note that the data and the performance of the model was heavily skewed towards performing loans because most of the data involved performing loans. To develop a more practical system, it is suggested that data involving more non-performing loans be provided to correct this bias.