**Summary**

Loan default occurs when a borrower fails to pay back a debt based on the initial arrangement. This means that successive payments have been missed over the course of period. Therefore, it is always the threat to any financial institution and should be predicted in advance using various features of the applicant. As I mention on my first delivery report, the objective of this project is to use historical loan application data to predict whether an applicant will be able to repay a loan or not this enables investors to avoid loans likely to default. So, for that we will build some machine learning statistical models, such as Naïve Bayes, random forest, and logistic regression using the lending club data.

**Exploratory Data Analysis**

It is difficult to understand the data set and make conclusions without looking through the entire data set. Understanding and interpreting data from large data sets can be very challenging. Spending more time exploring the data to get a better representation of the data set is important. The very first step in exploratory data analysis is to identify the type of variables in the dataset. Variables are of two types Numerical and Categorical. Once the type of variables is identified, the next step is to identify the futures (input) and Target (output) variables.

In this project, Exploratory Data Analysis is majorly performed using the following methods [1]

* Univariate visualization : to provides summary statistics for each field in the raw data set
* Bivariate visualization : to find the relationship between each variable in the dataset and the target variable of interest
* Dimensionality reduction: to understand the fields in the data that account for the most variance between observations and allow for the processing of a reduced volume of data.

The target variable is loan\_status, and we would like to know whether the loan would be defaulted or not using the selected features from the dataset. Our goal is to become confident that the data set is ready to be used in a machine-learning algorithm.[2]

**Data transformation**

In this final step, we have performed one-hot encoding on the categorical variables and then scale the numerical columns there are many ways to convert categorical values into numerical values. Each approach has its own trade-offs and impact on the feature set. We use one-hot transformation for the categorical variables as well as scale the numerical columns using the StandardScaler module from the sklearn.preprocessing module.and use 80% of the data as train set. We standardized the training set and test set separately afterwards.[2]

**Models**

This project focused on building different machine learning algorisms and evaluating the performance. For our dataset, We use the following Machine learning algorithms:;

* Naïve Bayes
* Logistic Regression
* Random Forest

Naive Bayes

Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes’ theorem with the “naive” assumption of conditional independence between every pair of features given the value of the class variable. Despite its simplicity, the Naive Bayesian classifier often does widely used because it often outperforms more sophisticated classification methods. However, in our case , the performance of the models is not good result shows 54% accuracy and precision of 31% which we will do some research in the final delivery[3]

Logistic Regression

[Logistic regression](http://www.statisticssolutions.com/academic-solutions/membership-resources/member-profile/data-analysis-plan-templates/data-analysis-plan-logistic-regression/) is the appropriate regression analysis to conduct when the dependent variable is binary. It is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. Model performance result shows, the recall score of logistic regression is 92% , which means we can use this model to classify the 92% of the bad loans.[4]

Random Forest

Random forest consists of a large number of individual decision trees that operate as an [ensemble](https://en.wikipedia.org/wiki/Ensemble_learning). Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model’s prediction. Since our data is an imbalanced class We have used oversampling the minority class by Synthetic Minority Over-Sampling Technique (SMOTE) in the training set, as a result, the recall score indicates that the logistic regression can classify 90% of the applicants that would be default will be appropriately identified. based on our model evaluation, we recommend to use logistic regression for default prediction.[5]

**Reference:**

# Exploratory Data Analysis by [Swetha Lakshmanan](https://medium.com/@swethalakshmanan14?source=post_page-----ebdf643a33f6----------------------) 2019 retrieved from <https://medium.com/@swethalakshmanan14/exploratory-data-analysis-in-python-ebdf643a33f6>

# InData Labs, 2017, Exploratory Data Analysis: the Best way to Start a Data Science Project retrieved from

# <https://medium.com/@InDataLabs/why-start-a-data-science-project-with-exploratory-data-analysis-f90c0efcbe49>

# scikit-learn developers, 2007 - 2019, Naïve Bays retrieved from

# <https://scikit-learn.org/stable/modules/naive_bayes.html>

# Statistics Solutions 2020, [Logistic regression](http://www.statisticssolutions.com/academic-solutions/membership-resources/member-profile/data-analysis-plan-templates/data-analysis-plan-logistic-regression/) retrieved from <https://www.statisticssolutions.com/what-is-logistic-regression/>

# [Tony Yiu](https://towardsdatascience.com/@tonester524?source=post_page-----58381e0602d2----------------------), 2019,Understanding Random Forest retrieved from <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>