**Data Scientist**
VIX Rakamin Academy
@ ID/X Partners


# CREDIT RISK CLASSIFICATION USING MACHINE LEARNING

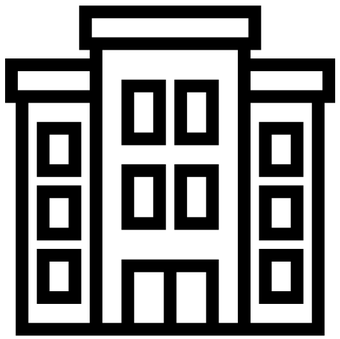
By Jonathan Stanley

# INTRODUCTION

# Background

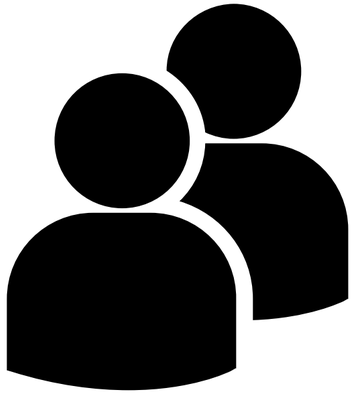
**Lending Company**

A lending company operates by  providing funds to individuals, businesses, or other entities in the form of loans with the agreement to repay it over a specified period with additional interest



**Cash Loan**

**Loan Amount + Interest**



**Lending Company**

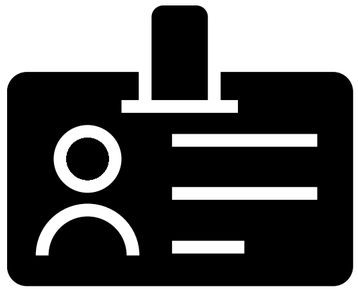
**Customer**

# Background

**Loan Application**

Before approving a loan, a lending company needs to evaluate whether or not a potential customer is eligible for a loan based on the credit risk from the information provided by the customer

## The information for credit assessment include:

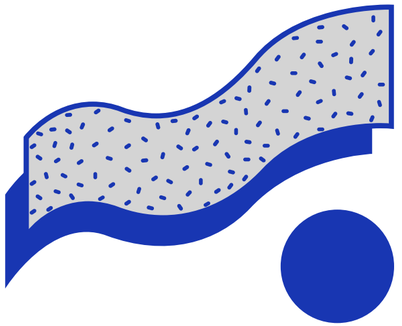


**Personal Information**

**Financial Status**


**Credit History**

# Machine Learning



Most lending companies and financial institutions have begun to incorporate machine learning methods for credit risk assessment. The benefits of using machine learning compared to human resources for this particular task include:

## Analytical Capacity

Machine Learning algorithms excels at handling a **huge amount of data and features**

## Speed and Efficiency

Machine Learning models operate at **high speed** and require **less operational cost** than regular human resources

## Complex Patern Recognition

Machine Learning models allow **a more detailed understanding of credit risk factors** and **relationship among variables**

# Business Undertanding

## Overview

Applying for loan to a lending company is normally a choice for most people who are in need of cash for urgent uses. Whether or not it is suitable to give a certain individual a loan is a problem that every lending company needs to consider.

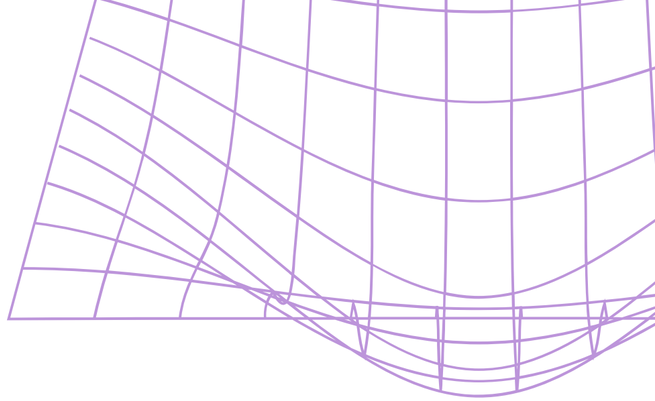
For a lending company, they evaluate the status of every borrower using credit risk. **Credit risk** is the possibility of a loss resulting from a borrower's failure to repay a loan or meet contractual obligations.

## Objective

Create a machine learning model to label whether or not an individual is suitable to apply for loan in the company by analyzing informations from previous loan receivers to minimize failure of loan repayment.

# METHODS

# Machine Learning Models



In this project, there are 3 machine learning models used for credit risk classification. Each model will be built and evaluated separately and the model which provides the best result will be chosen.

01

**Logistic Regression**

02

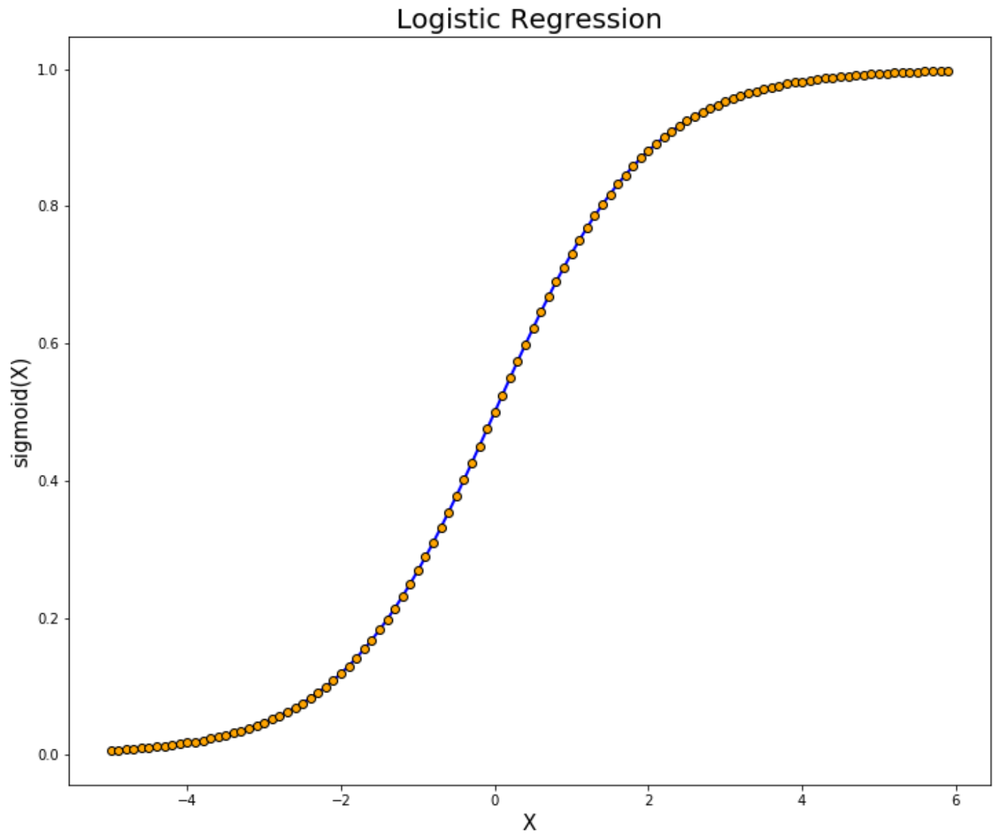
**K Nearest Neighbor Classifier**

03

**Random Forest Classifier**

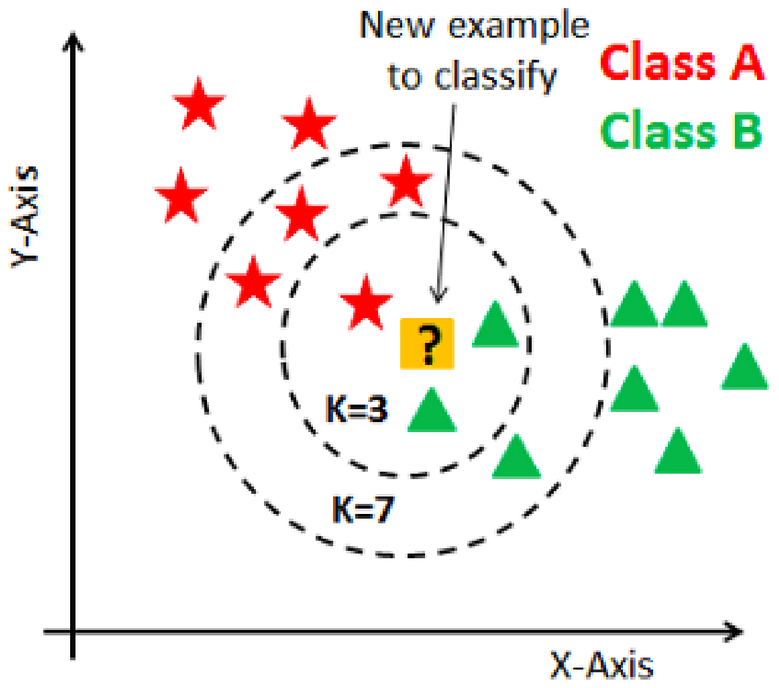
# Logistic Regression

- Logistic Regression is a classification algorithm that estimates the **probability of the dependent variable taking a particular value** based on the value of independent variables.
- The logistic regression model is constructed by applying a linear combination of input features to the **logit function**.



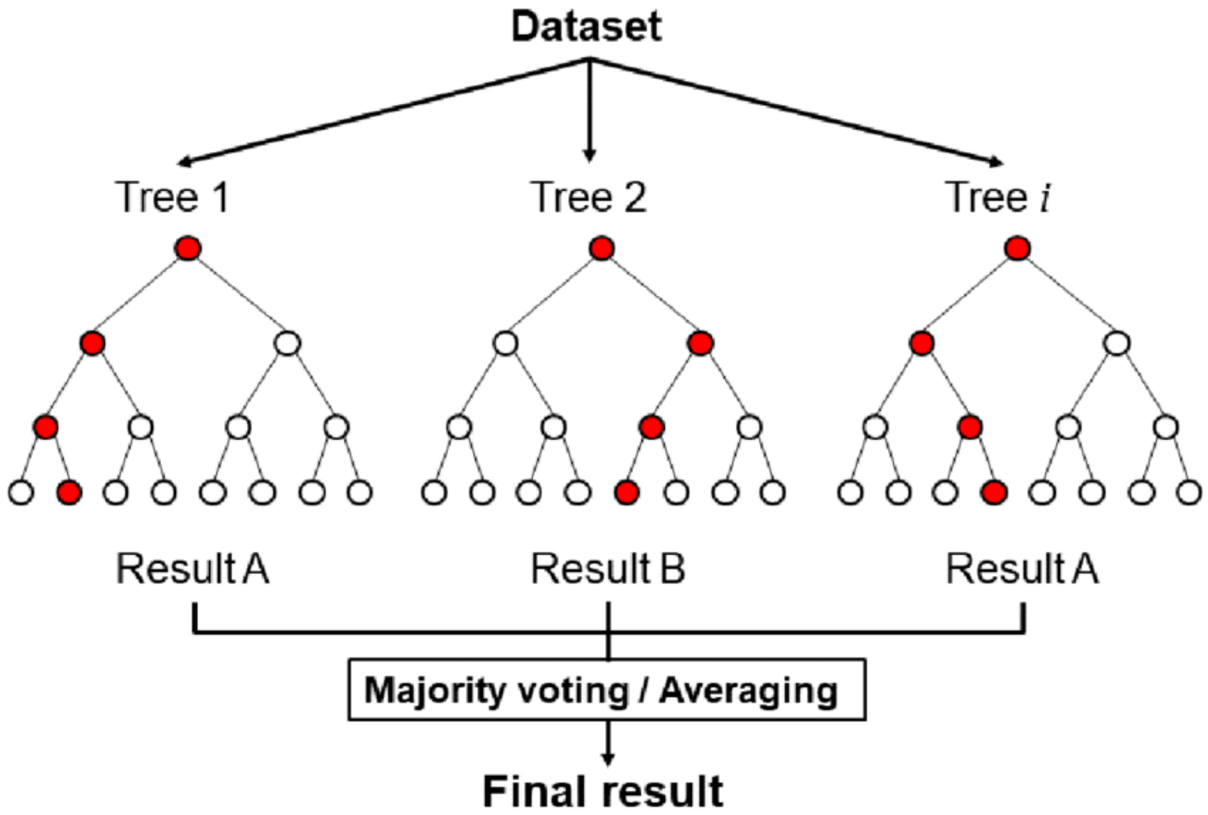
# K Nearest Neighbor

- **K-Nearest Neighbors (KNN)** is a type of **instance-based** learning where the algorithm makes predictions based on the **majority class of the k-nearest data points in the feature space**.
- When a **new data point** is selected, the algorithm calculates its **distance** to the other known points in training or testing set.
- KNN is often used in scenarios where the **decision boundaries are not well-defined** or when the **dataset is relatively small**



# Random Forest

- Random Forest is an ensemble model which is made up of **multiple decision trees** where each tree **is constructed independently and makes a prediction**.
- The final prediction in a Random Forest is determined by **aggregating (voting or averaging)** the predictions of the individual trees.
- Compared to a single decision tree, random forest provides a model which is more **robust and less prone to overfitting**.



# SIMULATION WORFKLOW


# Data Understanding

## Dataset Information

The dataset used in this project is a loan record dataset from the lending company based from previous users and the variables used to determine their loan credit status.

The dataset consists of 74 Columns with 466,285 Rows, with the column "loan_status" as the target variable.

# Data Understanding

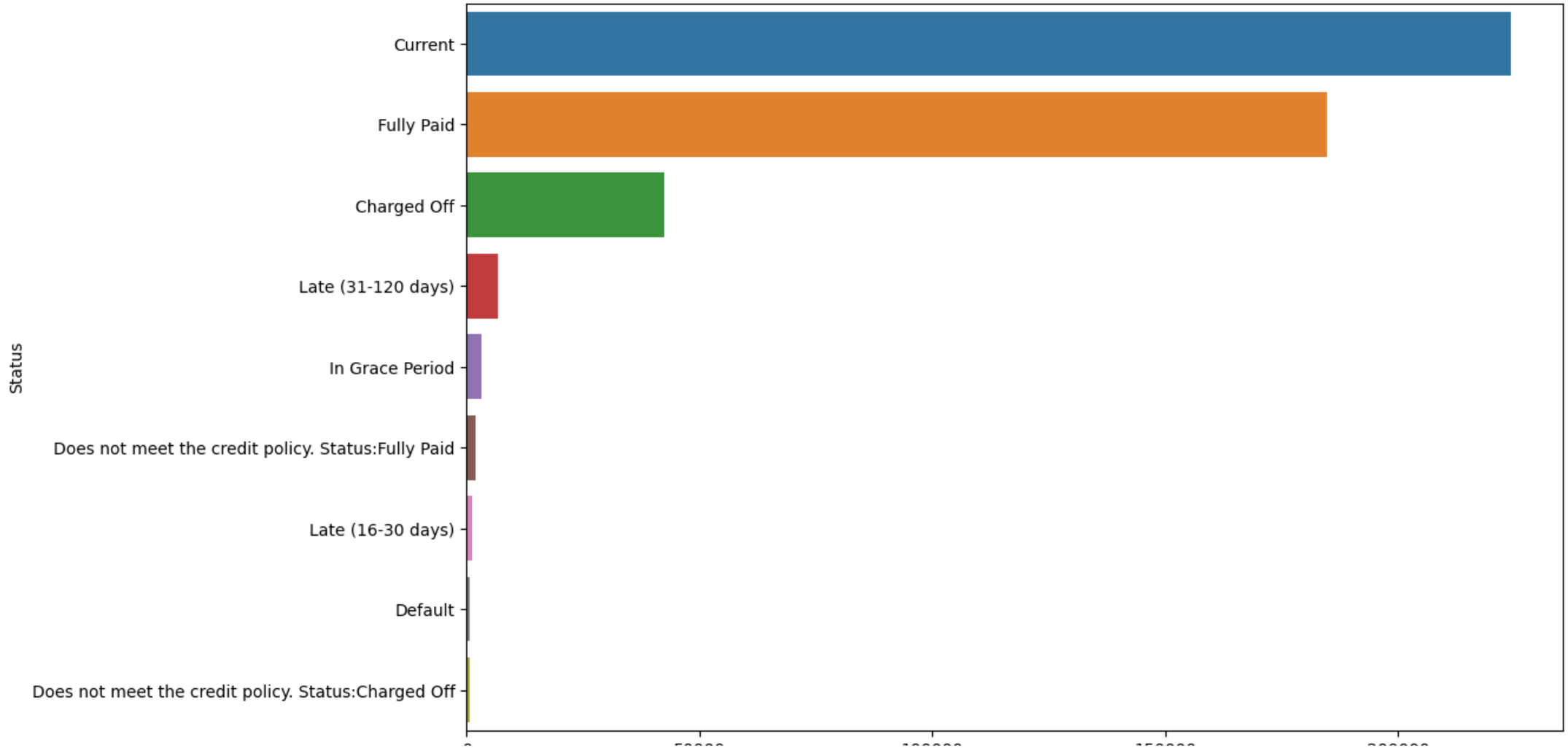
## First 10 Rows of the dataset preview

| | id | member_id | loan_amnt | funded_amnt | funded_amnt_inv | term | int_rate | installment | grade | sub_grade | ... | total_bal_il | il_util | open_rv_12m | open_ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1077501 | 1296599 | 5000 | 5000 | 4975.0 | 36 months | 10.65 | 162.87 | B | B2 | ... | NaN | NaN | NaN | |
| 1 | 1077430 | 1314167 | 2500 | 2500 | 2500.0 | 60 months | 15.27 | 59.83 | C | C4 | ... | NaN | NaN | NaN | |
| 2 | 1077175 | 1313524 | 2400 | 2400 | 2400.0 | 36 months | 15.96 | 84.33 | C | C5 | ... | NaN | NaN | NaN | |
| 3 | 1076863 | 1277178 | 10000 | 10000 | 10000.0 | 36 months | 13.49 | 339.31 | C | C1 | ... | NaN | NaN | NaN | |
| 4 | 1075358 | 1311748 | 3000 | 3000 | 3000.0 | 60 months | 12.69 | 67.79 | B | B5 | ... | NaN | NaN | NaN | |
| 5 | 1075269 | 1311441 | 5000 | 5000 | 5000.0 | 36 months | 7.90 | 156.46 | A | A4 | ... | NaN | NaN | NaN | |
| 6 | 1069639 | 1304742 | 7000 | 7000 | 7000.0 | 60 months | 15.96 | 170.08 | C | C5 | ... | NaN | NaN | NaN | |
| 7 | 1072053 | 1288686 | 3000 | 3000 | 3000.0 | 36 months | 18.64 | 109.43 | E | E1 | ... | NaN | NaN | NaN | |
| 8 | 1071795 | 1306957 | 5600 | 5600 | 5600.0 | 60 months | 21.28 | 152.39 | F | F2 | ... | NaN | NaN | NaN | |
| 9 | 1071570 | 1306721 | 5375 | 5375 | 5350.0 | 60 months | 12.69 | 121.45 | B | B5 | ... | NaN | NaN | NaN | |

10 rows × 74 columns

# Exploratory Data Analysis (EDA)

The **target variable** for the credit risk prediction classification problem is the **Loan Status**. **Loan Status** refers to the current status of the loan of each borrower.



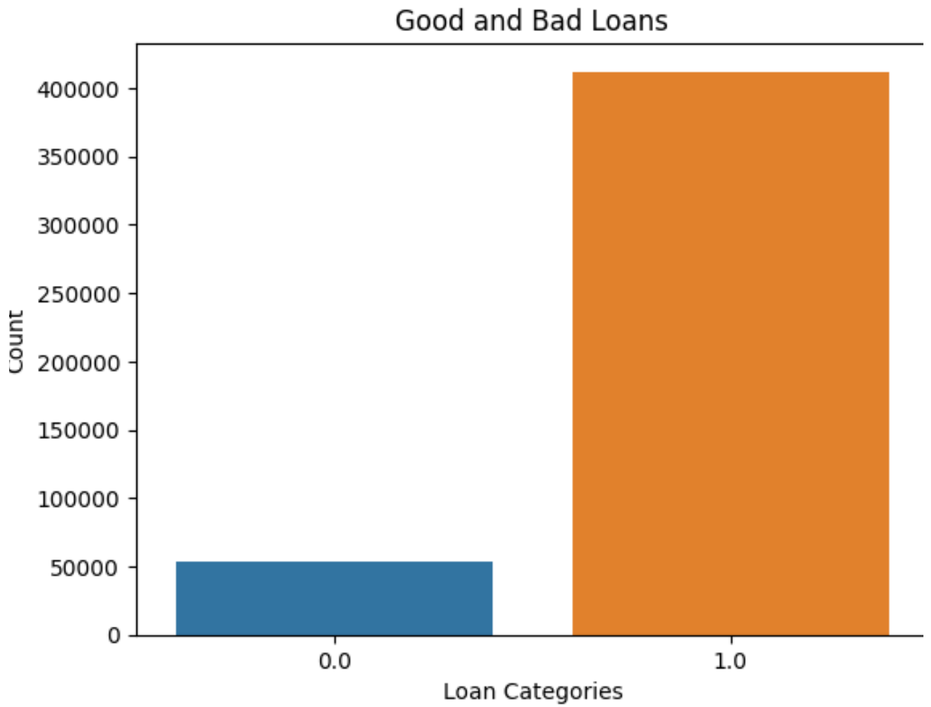
# Exploratory Data Analysis (EDA)

In this dataset, there are 9 different values from the **Loan Status** column. The next step will be dividing the loan status between the good loans and bad loans. In this scenario, Good loans will include "Fully Paid" values, while bad loans will include "Charged Off", "Late", "Does not meet the credit policy", and "Default" values.

```
Current                                            224226
Fully Paid                                         184739
Charged Off                                         42475
Late (31-120 days)                                   6900
In Grace Period                                      3146
Does not meet the credit policy. Status:Fully Paid   1988
Late (16-30 days)                                    1218
Default                                               832
Does not meet the credit policy. Status:Charged Off   761
Name: loan status, dtype: int64
```

# Exploratory Data Analysis (EDA)

For classification purposes, a new feature which is defined as loan_category will be created. Observations with loan status in the good loan category will be represented as ‘1’, while loan status in the bad loan category will be represented as ’0'.




The target variable loan category is not evenly distributed. The number of observations categorized as good loans are significantly larger than the number of bad loans.

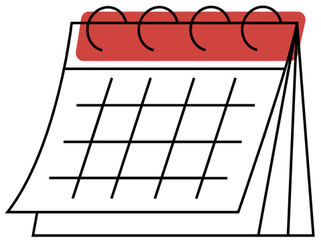

# Data Preparation

Before the modelling process, the dataset used will be processed further to achieve a better performance and result. The processing steps implemented in this project include feature engineering, feature selection, data cleaning, and SMOTE.

# Feature Engineering And Feature Selection

## Converting Datetime Features

- Features with datetime data type will be converted to numerical type by performing mathematic operations.
- The dataset contains data from 2007 to 2016. Mathematical operations which include the end of observed period will use the date ‘2016-12-31’.
- The previous features with datetime data type will be deleted.



## Pearson’s Correlation Feature Selection

- The absolute Pearson’s correlation matrix for numeric columns pairs will be calculated and features with high linear dependencies between them will not be used.
- The correlation threshold in this process is 0.7 (Features with correlation coefficient higher than 0.7 will be dropped).

# Data Preprocessing

## Handling Missing Values

- Dropping features with more than 70% missing values
- Substituting missing values with median for numerical columns, and mode for categorical columns.

## Unique and Irrelevant Features

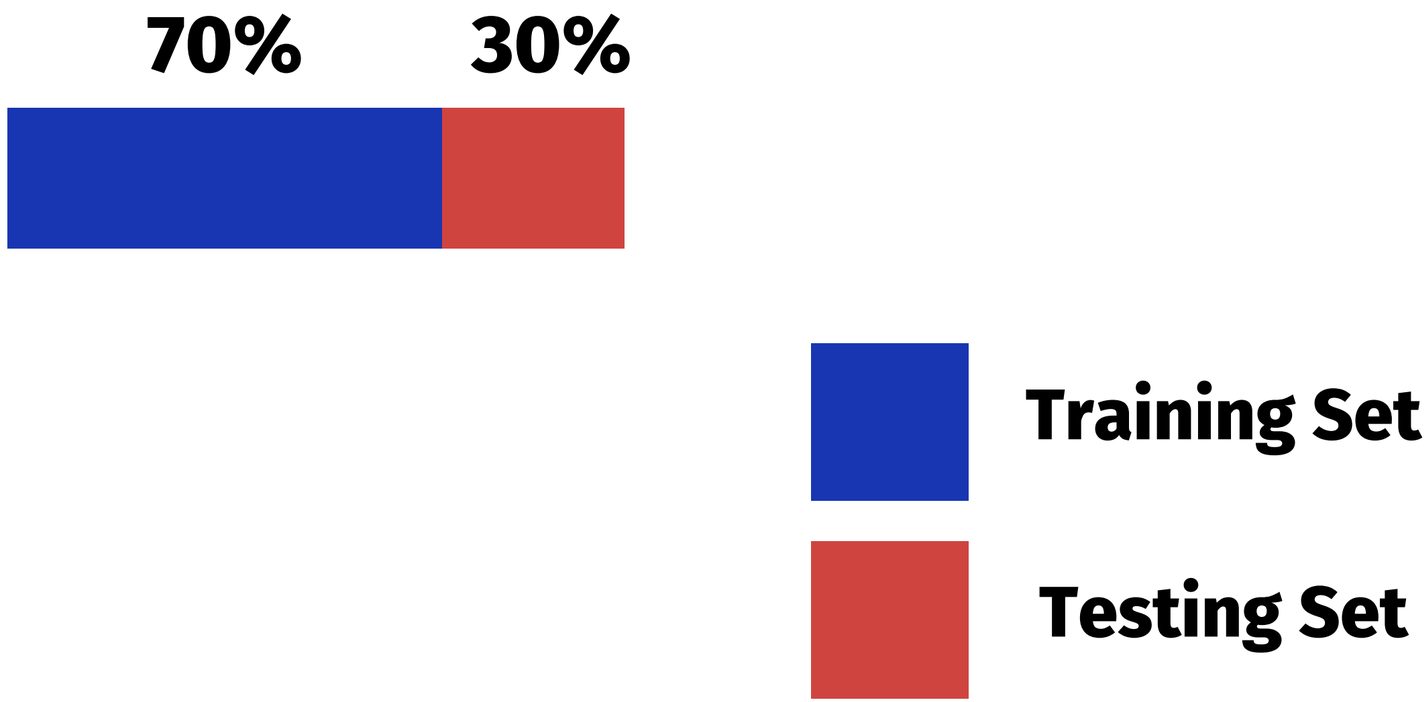
Dropping features where all their values are unique, has only one value, has values that are too specific and features with irrelevant informations.

## SMOTE Resampling

SMOTE generates synthetic samples by creating linear combinations of the feature values for selected instance and its k-nearest neighbors. The synthetic samples are added to the dataset which is used model training.

# Train Test Split

The dataset will be split into two subsets which will be used separately for training and testing. Each subset is labeled as training set and testing set respectively. The training set consists of 70% of the whole dataset while testing sets consists of 30% of the whole dataset.



# Data Modelling

## *Hyperparameter Tuning*

### Bayesian Optimization

For each iteration, 10 combinations of hyperparameters are evaluated

### 4 Fold Cross Validation

Model is trained and evaluated 4 times using different subsets of the data

# RESULT & ANALYSIS

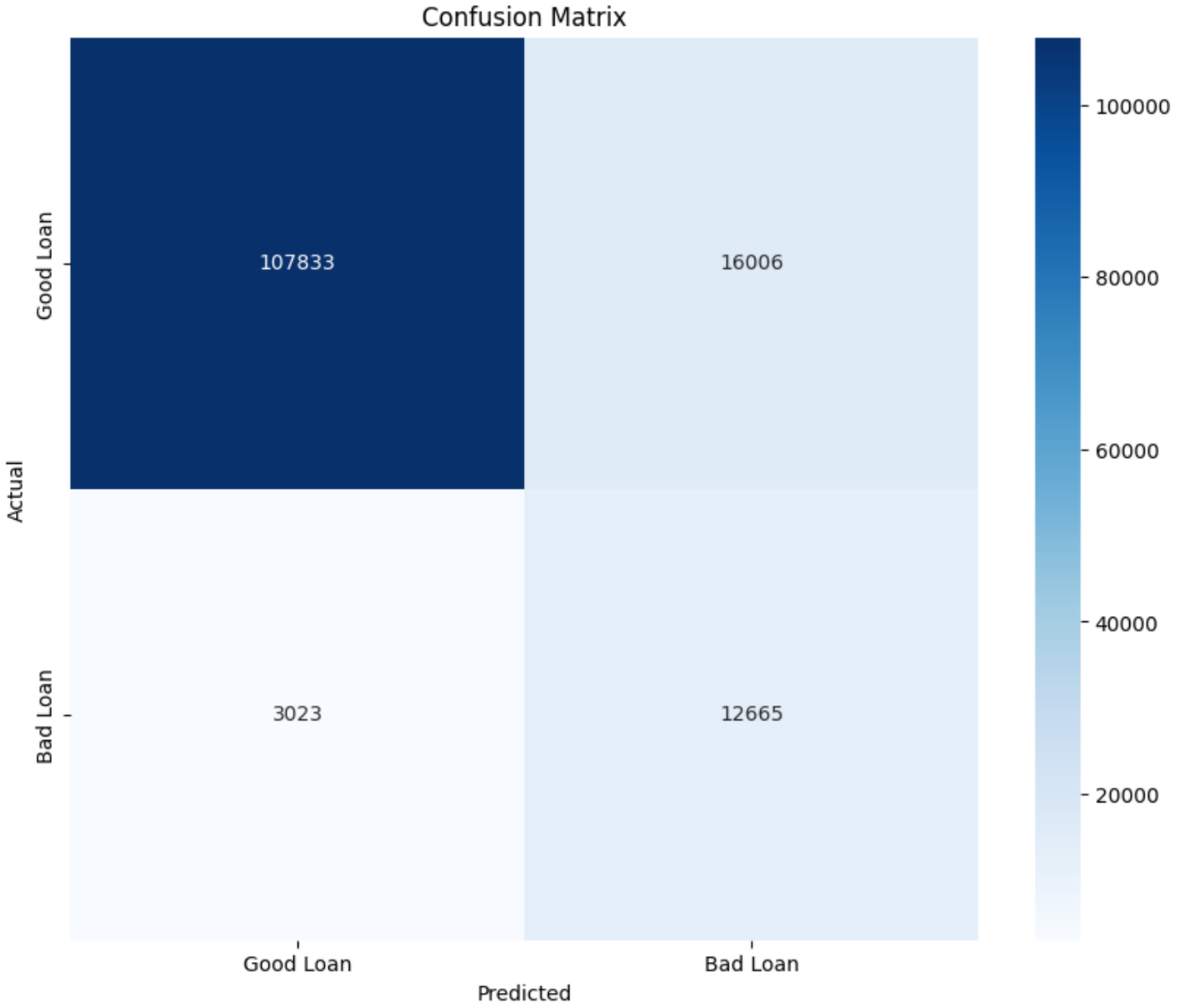
# Evaluation Metric Comparison

| | Sensitivity | Specificity | AUC Score |
|---|---|---|---|
| Logistic Regression | 0.9981 | 0.4627 | 0.8660 |
| KNN Classifier | 0.9490 | 0.4611 | 0.8193 |
| Random Forest Classifier | 0.8708 | 0.8073 | 0.9294 |

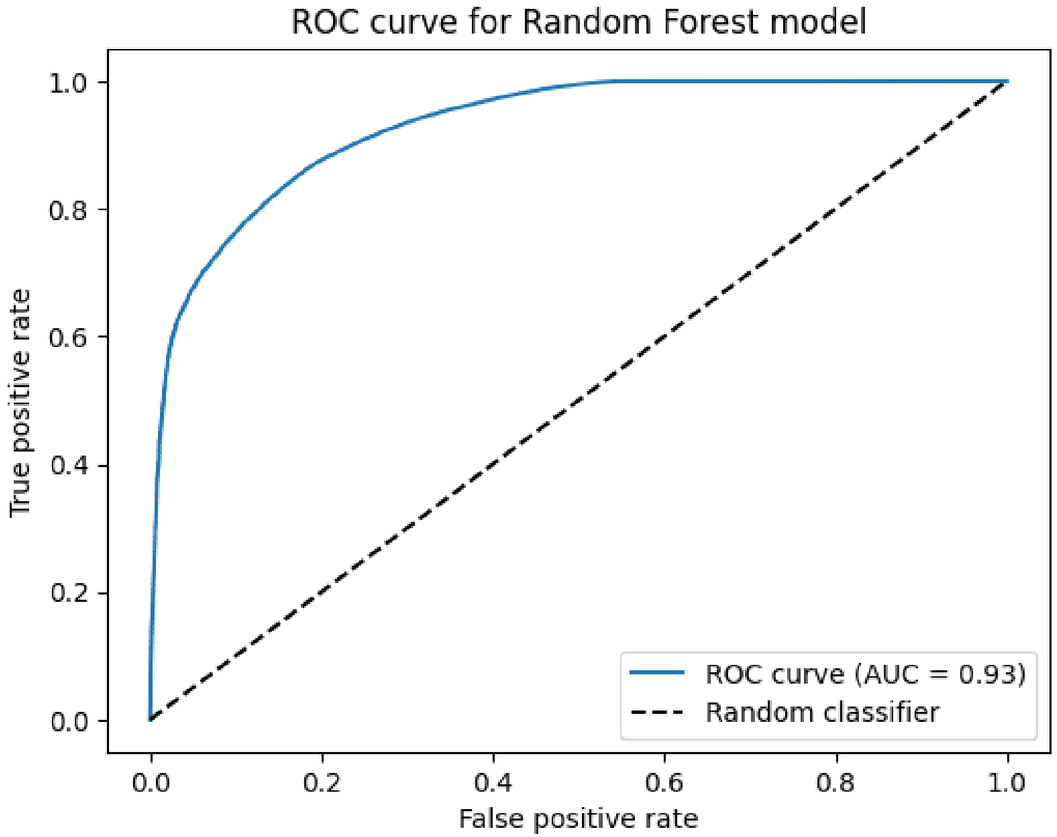
Random Forest Classifier model outperforms the other two model in two out of three evaluation metrics. The random forest model is also better at handling imbalanced data which is shown by a significant better result in specificity compared to the other two models. Considering this evaluation result, the random forest model is selected to be the most appropriate model in this project.

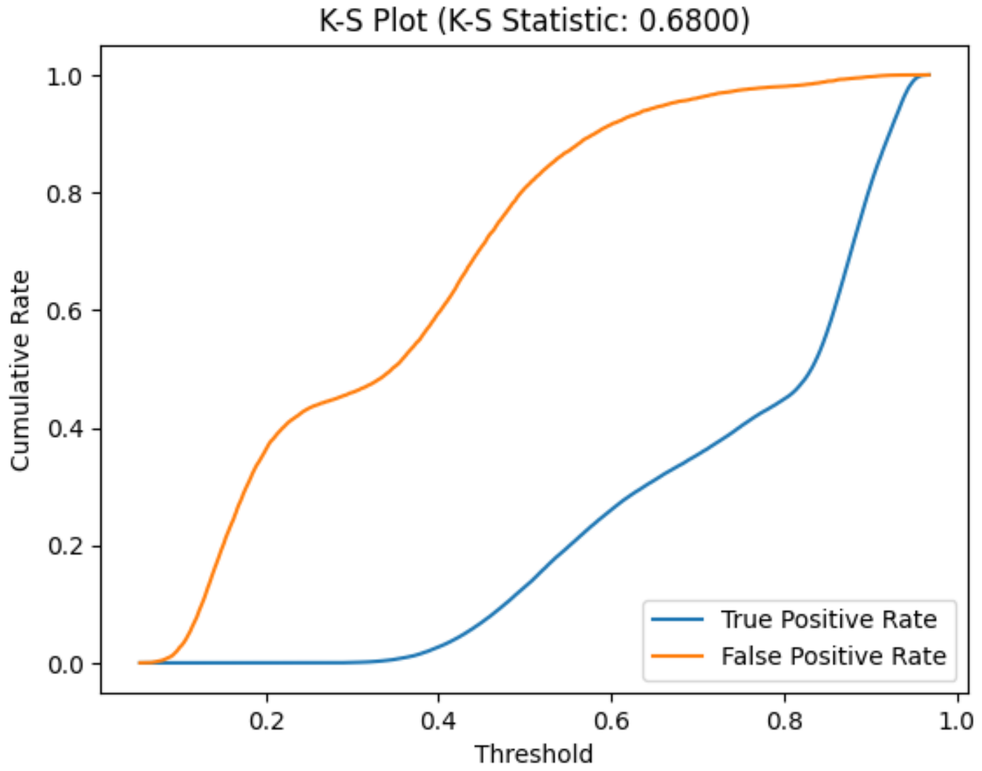
# Confusion Matrix For Random Forest Model

- Out of the **139,077** total loans, the model predicted **110,856** as good loans and **28,671** as bad loans.
- Out of the **110,856** loans predicted as good, **107,833** were correctly predicted as good and **3,023** were actually bad.
- Out of the **28,671** loans predicted as bad, **16,006** were actually good and **12,665** were correctly predicted as bad.



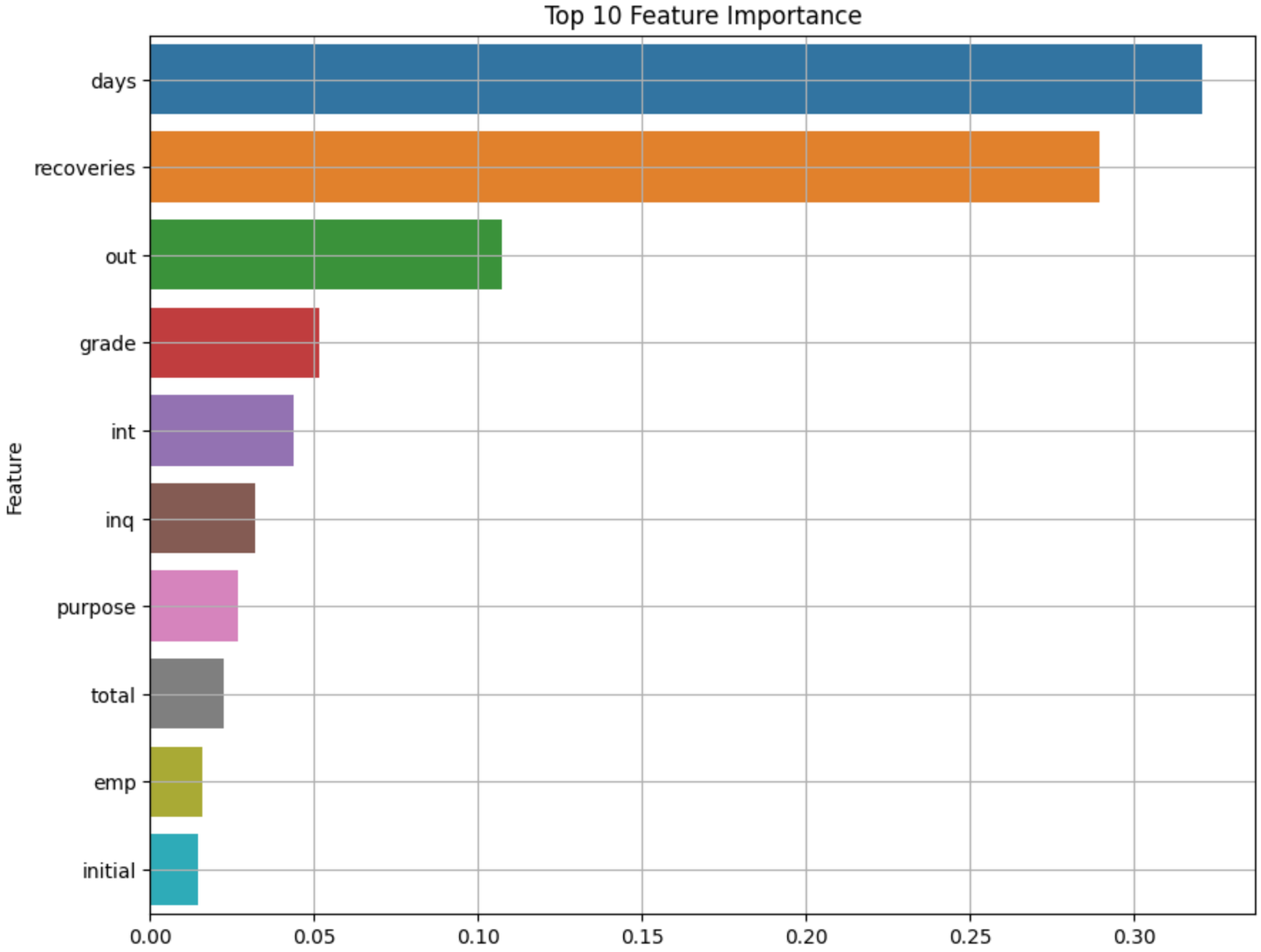
# ROC-AUC Curve and K-S Plot (Random Forest)




In credit risk classification, a prediction result with AUC score higher than **0.7** and K-S statistic higher than **0.3** is acceptable, which indicates that the result from this model is highly sufficient

# Feature Importance




- To interpret the model, feature importance is used to identify which features play a huge role in the prediction process.
- The feature days since last payment is the most important feature, followed by recoveries, and remaining outstanding principal.

# SUMMARY

# Summary

- The Random Forest Classifier model is chosen as the primary model in this project due to its performance which is better than logistic regression and KNN classifier model for credit risk binary classification in **two out of the three** evaluation metrics that are used for comparison.
- Due to the large number of observations and features in the dataset, there are many preprocessing methods used in this project before the modelling begins which includes feature engineering and selection using Pearson's correlation, data cleaning, data imputation, and also SMOTE to handle imbalanced data.
- The Random forest classifier model performed considerably well on the testing set with an AUC score of **0.93** and K-S score of **0.68** which is considered a pretty good result in credit risk classification.
- Feature importance is used to interpret the model based on which features contribute more towards the modelling process. The feature "days_since_last_payment" shows the highest importance for the random forest classifier model.

# Thank You

Feedbacks and Opinions are appreciated

Portofolio Github Link

## Contact

Jonathan Stanley
jonathanstanleyofficial@gmail.com
082112426652