LOAN PREDICTION PROJECT

Objective of the Course

This course is designed for people who want to solve binary classification problems using Python. By the end of this course, you will have the necessary skills and techniques required to solve such problems.

We will take following actions -

1. Introduction to the problem
2. Exploratory Data Analysis (EDA) and PreProcessing
3. Model building and Feature engineering

Let’s look at the steps that we will follow in this course.

1. Problem Statement
2. Hypothesis Generation
3. Getting the system ready and loading the data
4. Understanding the data
5. Exploratory Data Analysis (EDA)
    - Univariate Analysis
    - Bivariate Analysis
6. Missing value and outlier treatment
7. Evaluation Metrics for classification problems
8. Model Building : Part I
9. Logistic Regression using stratified k-folds cross validation
10. Feature Engineering
11. Model Building : Part II
    - Logistic Regression
    - Decision tree
    - Random Forest
    - XGBoost

Problem Statement :

Dream Housing Finance company deals in all home loans. They have presence across all urban, semi urban and rural areas. Customer first apply for home loan after that company validates the customer eligibility for loan. Company wants to automate the loan eligibility process (real time) based on customer detail provided while filling online application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others. To automate this process, they have given a problem to identify the customers segments, those are eligible for loan amount so that they can specifically target these customers.

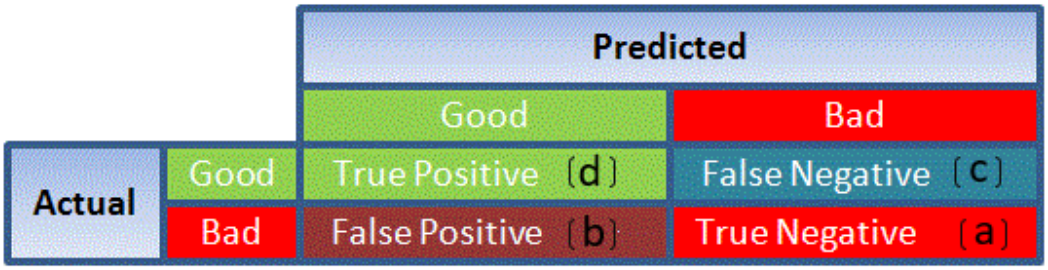
It is a classification problem where we have to predict whether a loan would be approved or not.

Hypothesis :

Lets recall some of the hypotheses that we generated earlier:

- Applicants with high income should have more chances of loan approval.
- Applicants who have repaid their previous debts should have higher chances of loan approval.
- Loan approval should also depend on the loan amount. If the loan amount is less, chances of loan approval should be high.
- Lesser the amount to be paid monthly to repay the loan, higher the chances of loan approval.

Accuracy :



- **Precision**: It is a measure of correctness achieved in true prediction i.e. of observations labeled as true, how many are actually labeled true.

Precision = TP / (TP + FP)

- **Recall(Sensitivity) -** It is a measure of actual observations which are predicted correctly i.e. how many observations of true class are labeled correctly. It is also known as 'Sensitivity'.

Recall = TP / (TP + FN)

- **Specificity -** It is a measure of how many observations of false class are labeled correctly.

Specificity = TN / (TN + FP)

ROC curve:

- Receiver Operating Characteristic(ROC) summarizes the model's performance by evaluating the trade offs between true positive rate (sensitivity) and false positive rate(1- specificity).
- The area under curve (AUC), referred to as index of accuracy(A) or concordance index, is a perfect performance metric for ROC curve. Higher the area under curve, better the prediction power of the model.

***The area of this curve measures the ability of the model to correctly classify true positives and true negatives. We want our model to predict the true classes as true and false classes as false.**