Loan Approval Prediction

This repo contains the Loan Approval Prediction project as part of my data science portfolio. This project is completed as part of the online hackathon organized by Analytics Vidhya. Evaluation metric of the hackathon is accuracy i.e. percentage of loan approval that is correctly predicted. After trying and testing 4 different algorithms, the best accuracy on the public leaderboard is achieved by Logistic Regression (0.7847), followed by Decision Tree (0.7778) while Decision Tree performed the worst (0.6458).

This project covers the whole process from problem statement to model development and evaluation:

1. Problem Statement
2. Hypothesis Generation
3. Data Collection
4. Exploratory Data Analysis (EDA)
5. Data Pre-processing
6. Model Development and Evaluation
7. Conclusion

Problem Statement

Business Problem

"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."

Loan prediction is a very common real-life problem that every retail bank faces in their lending operations. If the loan approval process is automated, it can save a lot of man hours and improve the speed of service to the customers. The increase in customer satisfaction and savings in operational costs are significant. However, the benefits can only be reaped if the bank has a robust model to accurately predict which customer's loan it should approve and which to reject, in order to minimize the risk of loan default.

Translate Business Problem into Data Science / Machine Learning problem

This is a classification problem where we have to predict whether a loan will be approved or not. Specifically, it is a binary classification problem where we have to predict either one of the two classes given i.e. approved (Y) or not approved (N). Another way to frame the problem is to predict whether the loan will likely to default or not, if it is likely to default, then the loan would not be approved, and vice versa. The dependent variable or target variable is the Loan Status, while the rest are independent variable or features. We need to develop a model using the features to predict the target variable.

Logistic Regression and Decision Trees were the two ML models deployed in this project

	Recall Score	F1 score
Logistic Regression	87.5	93.33
Decision Trees	87.5	93.33

Decision Trees: The basic algorithm of decision tree requires all attributes or features should be discretized. Feature selection is based on greatest information gain of features. The knowledge depicted in decision tree can represented in the form of IF-THEN rules.

IV. CONCLUSION: Both Models predicted the outcomes in an equal proportion.