Predictive Model for Personal Loans

Problem Description –

Universal Bank is a new bank with a rapidly growing customer base. Most of these customers are liability customers (depositors with checking/savings accounts). The customer base of asset customers (borrowers) is quite small, and the bank is interested in expanding this sector to bring in more loan revenue. It wants to explore ways of converting its liability customers to personal loan customers. Last year, Universal Bank ran a marketing campaign where they offered personal loans to select customers. The bank has data from this campaign where it can see which customers accepted/rejected the loan offers.

Universal bank wants to deploy a predictive model capable of identifying customers who are most likely to accept the personal loan in the next marketing campaign.

About the Dataset –

The dataset from that campaign is provided in the *UniversalBank_unprocessed.csv* file, and the data dictionary is provided in *UniversalBank_dictionary.csv* file.

Approach to Analysis –

1. **Explore the Data dictionary**
   This step includes exploring the given data dictionary to understand the definition of different variables in the dataset and differentiate between predictors and response variables.
2. **Data exploration and Preprocessing**
   Exploratory data analysis is done to understand the variables, the relationship between them and to clean the dataset. This comprises of following steps:
   i. Checking number of rows, columns, data types, unique values for each variable, finding outliers etc.
   ii. Determining the relationship between variables
   iii. Predictor selection
   iv. Dealing with null values
   v. Dealing with categorical predictors
   vi. Scaling predictors
3. **Model selection and building**
   Since the response variable is binary (Personal loan – Acceptor or Non acceptor), this is a case of Classification. In this project, the below two classification models have been used:
   i. Logistic Regression
   ii. k-NN
4. **Model performance evaluation and comparison**

Tools used – Python (Jupyter Notebook)

Conclusion –

F1 scores are the harmonic mean between Precision and Recall. It ranges between 0 to 1, higher the score, better the predictive performance of the model.

Both the models are just fit as the F1 scores for the training and test dataset is almost same. So there is no case of underfitting or overfitting.

F1 scores from both the models for the test data were significantly high – **Logistic Regression (86.1%)** and **k-NN (91.9%).**

k-NN model should be used for further testing and deployment for predicting which customers are more likely to be personal loan acceptors.