"E-Bike Loan Default Prediction: Mitigating Credit Risk"

PRESENTER - SUYASH PANDEY

Problem Statement:

- A bank is suffering from increasing incidence of non-payment in their E-bike financing division.
- The default is causing big impact in banks probability as due to high competition in the market the IRR has already very less and bank is facing challenges to sustain in this condition

“Problem Objective:”

- They want to build a credit model based on their customer data predict default so that they can control the risk.
- An effective model will definitely help bank to overcome the challenges of default and maintain profitability.

Data Description:

- The dataset contains total 39 features divided further in 3 categories:
  - ID – Identifier for each unique Customer
  - Default – Target feature (Whether customer is defaulter or not)
  - Independent variables – Example – Salary , Residence , Job status , Childrens , previous loans , Automobile Possession etc.

Data Pre-processing Steps and Inspiration:

- Exploratory Data Analysis – EDA .
  - Identification and removal of Null Values Based on Domain knowledge, Correlation with the target
  - Replacing the Null values with mean, Median and mode based on distribution and Category
  - Identification of Outliers (Not removed as the impact was huge and removal didn't suit the model)
  - Visualizing the correlation between each Feature
  - Feature selection Techniques used IV Analysis for categorical features and VIF for Numerical features to avoid multi collinearity

Plotting correlation of dataset:

Complexity of the dataset :

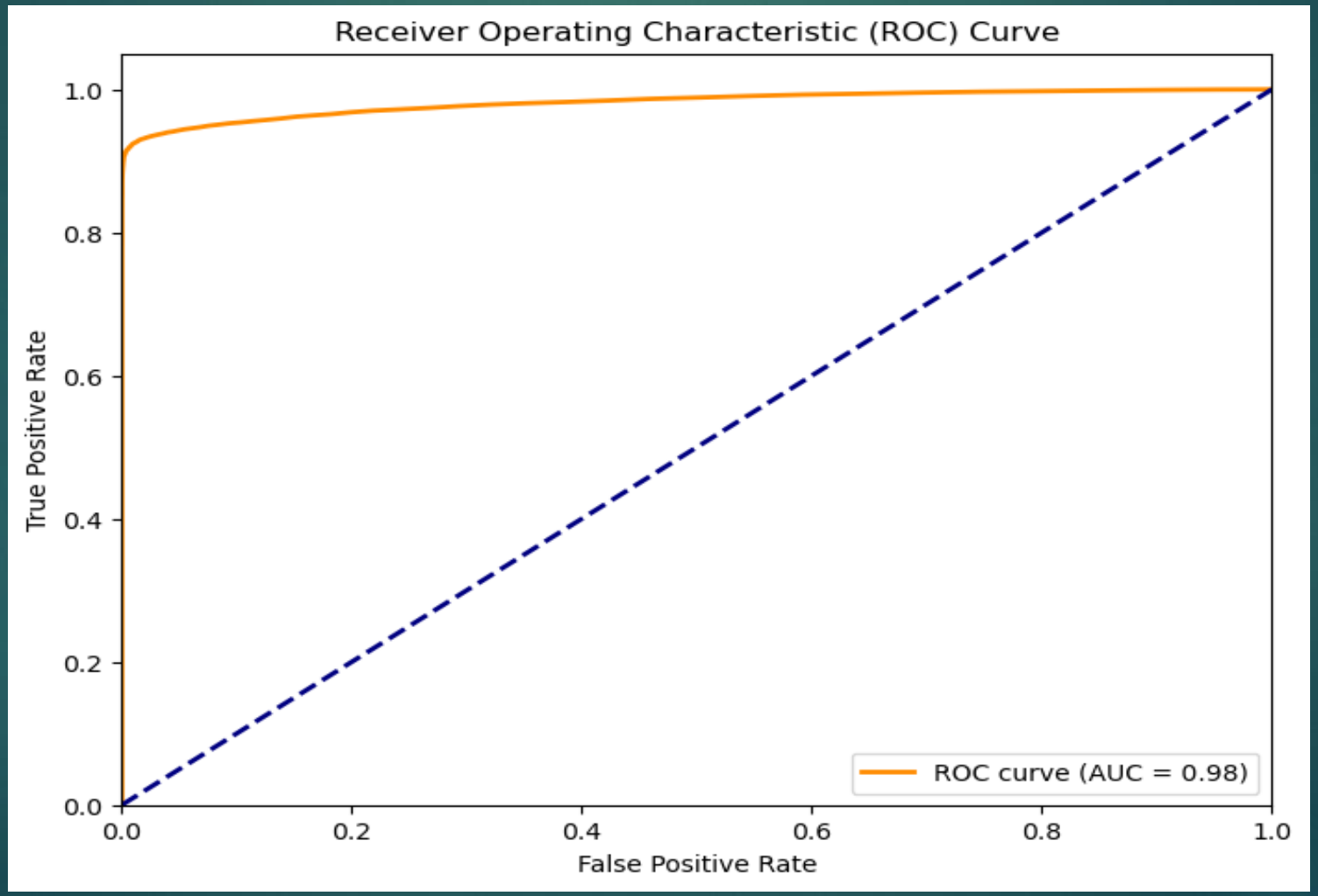
- The dataset was highly imbalanced dataset as the data was divided in the ratio of 92:8 so below techniques were used to overcome Imbalance nature of the dataset :
  - Under sampling - Majority class was reduced to the size of minority class.
  - Over sampling - Minority class was Oversized to the size of Majority class.
  - SMOTE – Synthetic samples were produced for minority to match majority.

  Conclusion – SMOTE gave the best accuracy among all so it was finalized for final model building.

Choosing the Algorithm for the Project:

- Since this was a classification problem , Logistic Regression , Decision trees and Random forest were the best choices.
- Based on the complex nature of the dataset and highest accuracy Random forest proved to be the choice for the model.

ROC curve Plot for Random forest model:



Model Evaluation and Techniques:

- Being an Imbalance dataset accuracy score couldn't be the correct choice so we vouched for Classification report with focus on F1 score to be approximately same for both classes and we ended up with an 96% accuracy for both the classes in Random forest algorithm.

References:

- I have Referred to one of my previous projects for classification problems.
- YT lecture by Code basics on how to deal with Imbalanced dataset.

Thank You