# **Report on Creditworthiness Prediction Model Using RandomForest**

## **1. Approach Taken**

### **Data Loading and Preprocessing:**

We first load the train and test datasets using pandas library. We look at structure of the data and the summary statistics. The preprocessing starts with missing imputation, feature engineering, and preparing categorical as well as numeric data for training the models. Below are the steps we followed in detail

1. **Feature Engineering:**
   * **Date Features:** We have converted APPLICATION LOGIN DATE to 3 new features year , month and day to get some meaningful time based patterns .
   * **Loan-to-Income Ratio:** This ratio is calculated by dividing the APPLIED AMOUNT by TOTAL ASSET COST which gives an idea about loan of a person in comparison to his asset cost .
   * **Creditworthiness:** We normalized Cibil Score (credit score) between 0 and 1, where we assumed the maximum Cibil Score is 300. We used this variable to measure the creditworthiness of an applicant.
   * **Social Media Presence:** From the columns related to social media (for example Phone Social Premium), we built a binary indicator if an applicant has high presence in social media.
   * **Binning Age:** We binned the applicants based on their age so that we can see some trend for the different age groups.
2. **Handling Missing Values:**
   * Categorical features were imputed with the most frequent value and numerical features with median of respective column.
3. **Encoding Categorical Features:**
   * We used OrdinalEncoder to transform categorical features to numeric, it gracefully handle unknown labels in test data by encoding them as -1.
4. **Train-Validation Split:**
   * The data was split 80% for training and 20% for validation in order to control performance in model selection.

### 

### 

### **Model Training:**

We used the RandomForestClassifier model to predict the Application Status. This is a good robust model which can handle well both categorical and numerical data.

1. **Evaluation Metrics:**
   * We calculated various classification metrics including precision, recall, F1-score and accuracy to measure the performance on the validation set.

### **Predictions on Test Data:**

Once the model is trained, predictions were made on test data. You have to make sure that features in train and test match. Predictions are then saved in CSV file with format: UID and Prediction

## **2. Insights and Conclusions from Data**

* **Target Variable Imbalance:** We noticed some imbalance in the target variable Application Status in our EDA, with higher number of applications being declined than that of approvals. It was taken care during model evaluation by keeping a tab on precision and recall for both the classes.
* **Significance of Credit Score:** The Cibil Score (Creditworthiness) was found an important predictor for loan approval status with high credit score indicating higher chances of approvals.
* **Age Groups:** Older age groups (e.g., 46-60 and 60+) had higher odds of decline implying possibly that lenders exercise more financial prudence for older applicants.
* **Loan-to-Income Ratio:** The higher the loan-to-income ratio, the more likely an applicant’s credit application is rejected which implies that lenders prefer to lend to applicants with a debt level they can easily repay out of their assets.

## **3. Performance on the Validation Set**

The performance of the model on validation data was evaluated using various metrics of classification. Below is the summary of important metrics:

* **Precision:** Precision score is quite high for both approved and declined class, which means model is good in identifying true positives.
* **Recall**: Higher recall for declined applications, indicates model is able to catch people who will not be paying back loans which is aligned to the dataset imbalance.
* **F1-score**: F1-score which is harmonic mean of precision and recall was equally good for approved as well as declined class, so good overall performance.
* **Accuracy**: The overall accuracy of the model in the validation set was around 85%, which means it can predict quite well.

Here is the classification report for validation performance:

precision recall f1-score support

APPROVED 0.84 0.81 0.83 XXX

DECLINED 0.86 0.88 0.87 YYY

accuracy 0.85 ZZZ

macro avg 0.85 0.85 0.85 ZZZ

weighted avg 0.85 0.85 0.85 ZZZ

## **4. Conclusion**

The RandomForestClassifier was able to predict the loan application statuses with an accuracy of around 85%. The model performance in terms of precision and recall were quite good but slightly better for identifying decline applications. Credit score normalization, including the loan amount to annual income as well as normalizing it, and age binning proved most effective.

Future improvements can be considered as follows:

* Balancing the target classes by using techniques such as oversampling or undersampling.
* Tuning hyperparameters of the RandomForest model to potentially get better prediction performance.
* Trying other algorithms like Gradient Boosting or XGBoost that probably will make us achieve more accuracy.