BFSI Case Study

By
Laltesh Choudhary
Thai Quang Thinh

Table of Contents

- Problem Statement
- Recommendations for case study
- Steps Followed
- EDA & Feature Selection
- Data Modelling

Problem Statement & Objective of the Study

- As a business analyst for Home Credit, you are supposed to first gather the information and clean it to make it usable.
- The bureau information is at trade level, each individual trade level information is provided. You need to apply 'Feature Engineering' techniques to roll up the information at applicant level, and thereby create manual features for model building.
- Build a classification model to differentiate applicants between approves and rejects.
- As a business analyst, you would want to find answers to the below questions for the bank:
- How to leverage trade level information for Credit Bureaus by aggregating trade level information to applicant level in order to capture their payment behaviour?
- Which application or payment behaviour factors significantly influence borrower's behaviour on any new disbursed loan?
- After identifying these factors, how to leverage them in the form of a model which can be used for decisioning? Once the model is built, how to translate the model output into strategies and business insights for the bank?

Recommendations

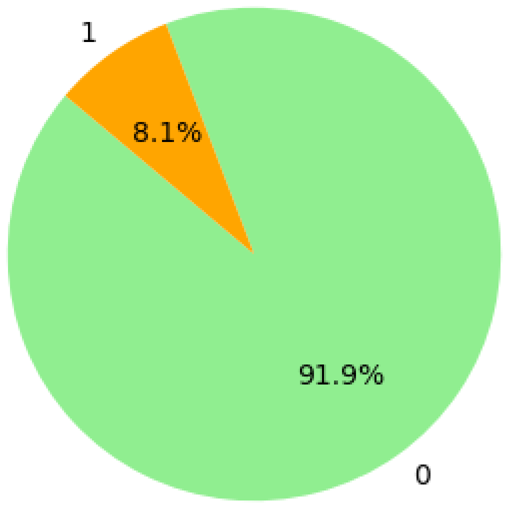
- Top 10 Factors Influencing Borrower Behaviour: 'CREDIT_TERM' 'DEBT_CREDIT_RATIO' 'DAYS_EMPLOYED' 'DAYS_BIRTH' 'TOTAL_DEBT_SUM' 'AMT_GOODS_PRICE' 'AMT_CREDIT' 'DAYS_CREDIT_interval_1-2 years' 'DAYS_CREDIT_interval_3+ years' 'DAYS_CREDIT_interval_<1 year'
- Factors such as 'CREDIT_TERM', 'AMT_CREDIT', 'AMT_GOODS_PRICE', DAYS_EMPLOYED significantly influence borrower behaviour. These features highlight aspects like loan terms, goods price affordability ratios, Employment duration/days ,providing insights into borrowers, age group.
- **The 'FLAG_DOCUMENT' column does not show any significant relationship with the 'TARGET' variable. Therefore, these columns are dropped from analysis.**
- People prefer taking cash loans over revolving loan
- People with an academic degree are more likely to repay their loans compared to those in other categories.
- **All the Students and Businessman are repaying loan.**
- **Widows are more likely to repay the loan when compared to applicants with the other family statuses.**
- **office apartment are more likely to repay the loan when compared to applicants with the other family statuses.**
- **People with high income(>1000000) are likely to repay the loan.**
- As per our analysis **we see that the clients who have difficulty in payment are relatively younger and most of them lie at around 30's.**
- **Individuals employed for less than two years are less likely to repay loans.**

Steps Followed

- Data Exploration of Application Data
- Data Quality Checks
- Univariate / Bivariate Analysis
- Data Exploration of Bureau Data
- Combining Application and Bureau data
- Feature Engineering - Bureau Data
- Feature Selection using F-Score
- Machine Learning Model Building (Train/Test split and Standard Scaler)
- Model Building :
- Recommendation

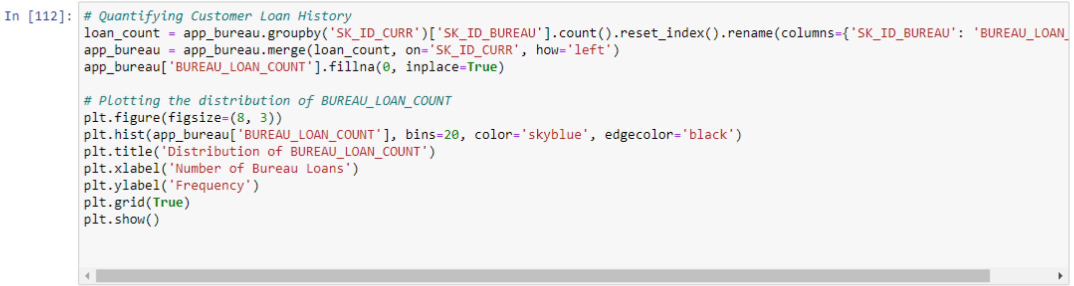
EDA

The data is highly imbalanced, with 91.9% of 'TARGET' variable being 0 and 8.1% being is 1(difficulty in repaying the loan).



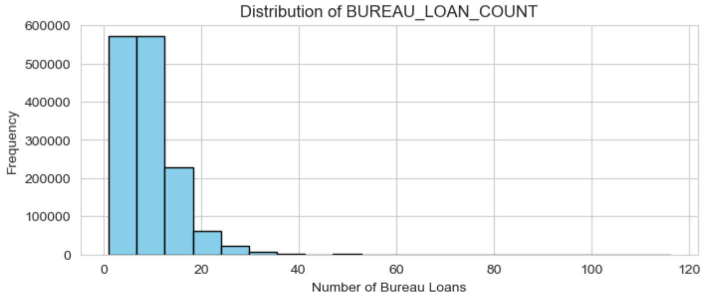
Feature Selection

- High frequency at the lower end of the scale in both the feature indicating individual having no or very low outstanding debt. It will be helpful for understanding the borrowers credit behavior.



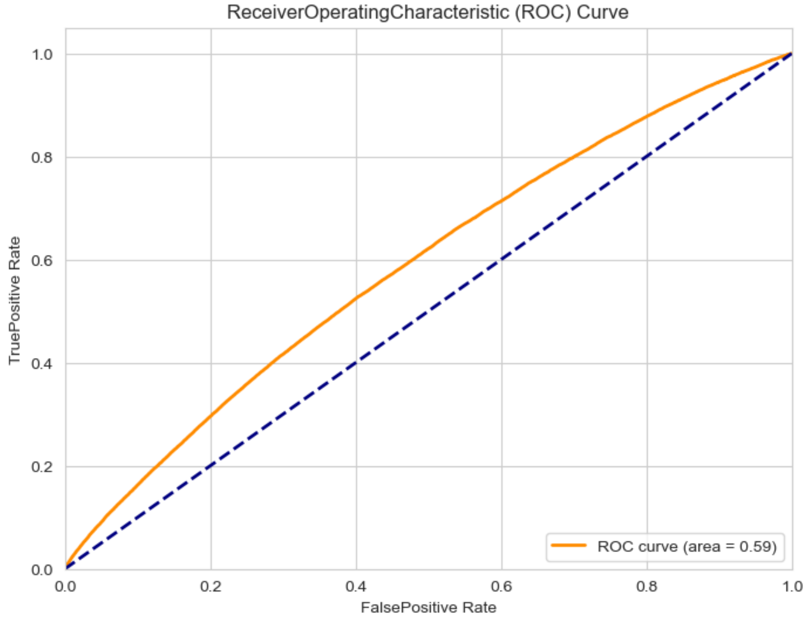
```
In [112]: # Quantifying Customer Loan History
loan_count = app_bureau.groupby('SK_ID_CURR')['SK_ID_BUREAU'].count().reset_index().rename(columns={'SK_ID_BUREAU': 'BUREAU_LOAN_
app_bureau = app_bureau.merge(loan_count, on='SK_ID_CURR', how='left')
app_bureau['BUREAU_LOAN_COUNT'].fillna(0, inplace=True)

# Plotting the distribution of BUREAU_LOAN_COUNT
plt.figure(figsize=(8, 3))
plt.hist(app_bureau['BUREAU_LOAN_COUNT'], bins=20, color='skyblue', edgecolor='black')
plt.title('Distribution of BUREAU_LOAN_COUNT')
plt.xlabel('Number of Bureau Loans')
plt.ylabel('Frequency')
plt.grid(True)
plt.show()
```



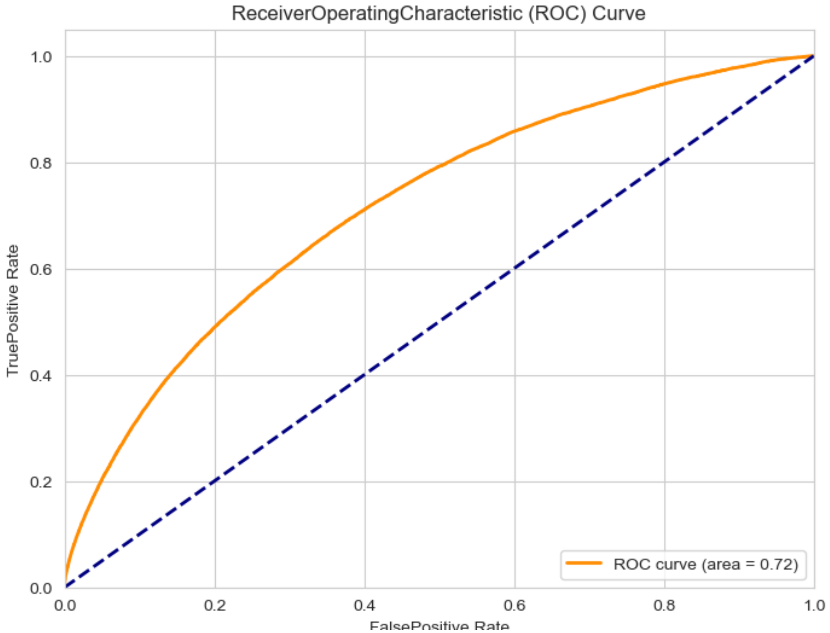
Binary Classification Model with Logistic Regression

- From the ROC curve it is clear that model is skewed towards (Class 0). This model is unable to correctly identify most of the positive outcomes.



Binary Classification Model Random forest along with Hyper parameters

- The ROC curve area of 0.72 demonstrates the model's moderate ability to discriminate between loan approvals and rejections, providing valuable insights for risk assessment.



Thank You