BFSI Analysis : A Domain Oriented Case Study

SUBMITTED BY,

Josemon Joy, Kavitha, Kanika

Problem Statement

Develop an internal credit risk scoring model for Home Credit by leveraging applicant-level information from loan applications and aggregated trade-level data from credit bureaus.

The goal is to classify loan applicants into approvals and rejections based on their past payment behavior and application details, enabling Home Credit to make informed lending decisions while balancing risk and customer acquisition.

Key challenges :

1.Aggregating trade-level data to the applicant level to capture credit behavior.

2.Identifying key factors influencing loan repayment behavior.

3.Building a classification model to distinguish between approved and rejected applicants.

4.Translating model outputs into business strategies for risk management and loan decisioning.

Objectives

- Data Collection & cleaning
- Feature Engineering
- Exploratory Data Analysis (EDA)
- Model Development
- Identification of Key Factors
- Model Interpretation & Business Strategy
- Implementation & Decision making

Data Understanding and Preparation

- Two Data Sets – application information and Bureau level data with shape of (307511, 122) and (1716428, 17)
- Inner joined both data set with 'SK_ID_CURR

Handling Missing Values

Since the data set is so huge, deleted columns having high null values above 30%. And imputed rest missing values with median and mode imputation. Used Mode Imputation, because of the presence of Outliers for the continuous variable columns

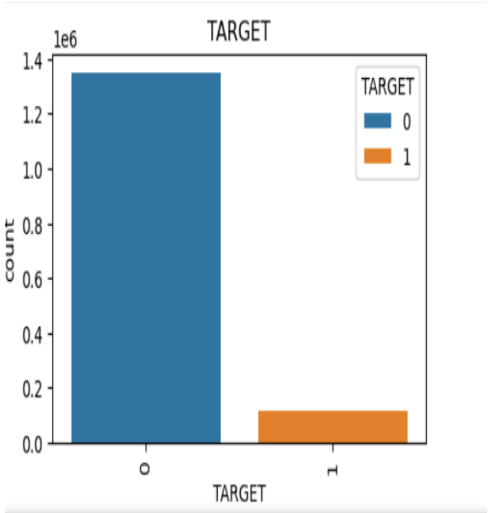
FEATURE ENGINEERING

- ❑ Added New Features: 'NUM_TRADES_REPORTED', 'NUM_CLOSED_TRADES', 'MAX_CREDITDAY_OVERDUE', 'CREDIT_ASSET_RATIO', 'DOCUMENT_COUNT', 'MOB_NO_COUNT', 'CREDIT_OVERDUE_COUNT', 'INCOME_CAT', 'AGE_CATEGORY',
- ❑ Dropped high dimensional categorical columns above 20
- ❑ Label Encoding – 'One-Hot-Encoding' - For Categorical Variables

EDA

- 'NAME_CONTRACT_TYPE' - Persons who take cash loans have more number of payment difficulties
- 'CODE_GENDER' - Feamales have more loans and more difficulties in number but Males are more proportionate payment difficulties
- 'NAME_FAMILY_STATUS' - Persons who married were got married gets more loans and more number of payment difficulties.
- 'NAME_HOUSING_TYPE', - Persons who lives in House/aparments have more number of payment difficulties
- 'CREDIT_ACTIVE', credit_active customers have more shows more proportionate payment difficulties...
- 'CREDIT_TYPE', Those who use consumer credit have more number of payment difficulties
- 'CREDIT_OVERDUE_COUNT' - Those who have no overdues shows up more number of payment difficulties, but those who have one overdue have more proportionate payment difficulties
- 'INCOME_CAT' - Middle_Income category have more number of payment difficulties but Lower Income have more proportionate payment difficultiy
- 'AGE_CATEGORY', 30-40 and 40-50 age groups shows more number of payment difficulties, but 20-30 age group have more proportionate payment difficulty

- TARGET column, the persons payment difficulties is unbalanced-
- For the 'FLAG_PHONE' feature those who have no phone have higher number of payment difficulties- T
- hose who not provided email have higher number of payment difficulties-
- for the REGION_RATING_CLIENT column 2 rated regions have higher payemnt difficulties and 3 rated columns have higher rate of payment difficulties-
- REGION_RATING_CLIENT_W_CITY 2 rated city have more payment difficult persons and 3 rated city have more proportionate payment difficult persons-
- Those who live in REG_CITYNOT_WORK_CITY have higher proportionate payment dificulties-
- Those who have DOCUMENT_COUNT only 1 have higher number of payment difficulties-
- Those who have MOB_NO_COUNT 2 or more have higher payment difficulties

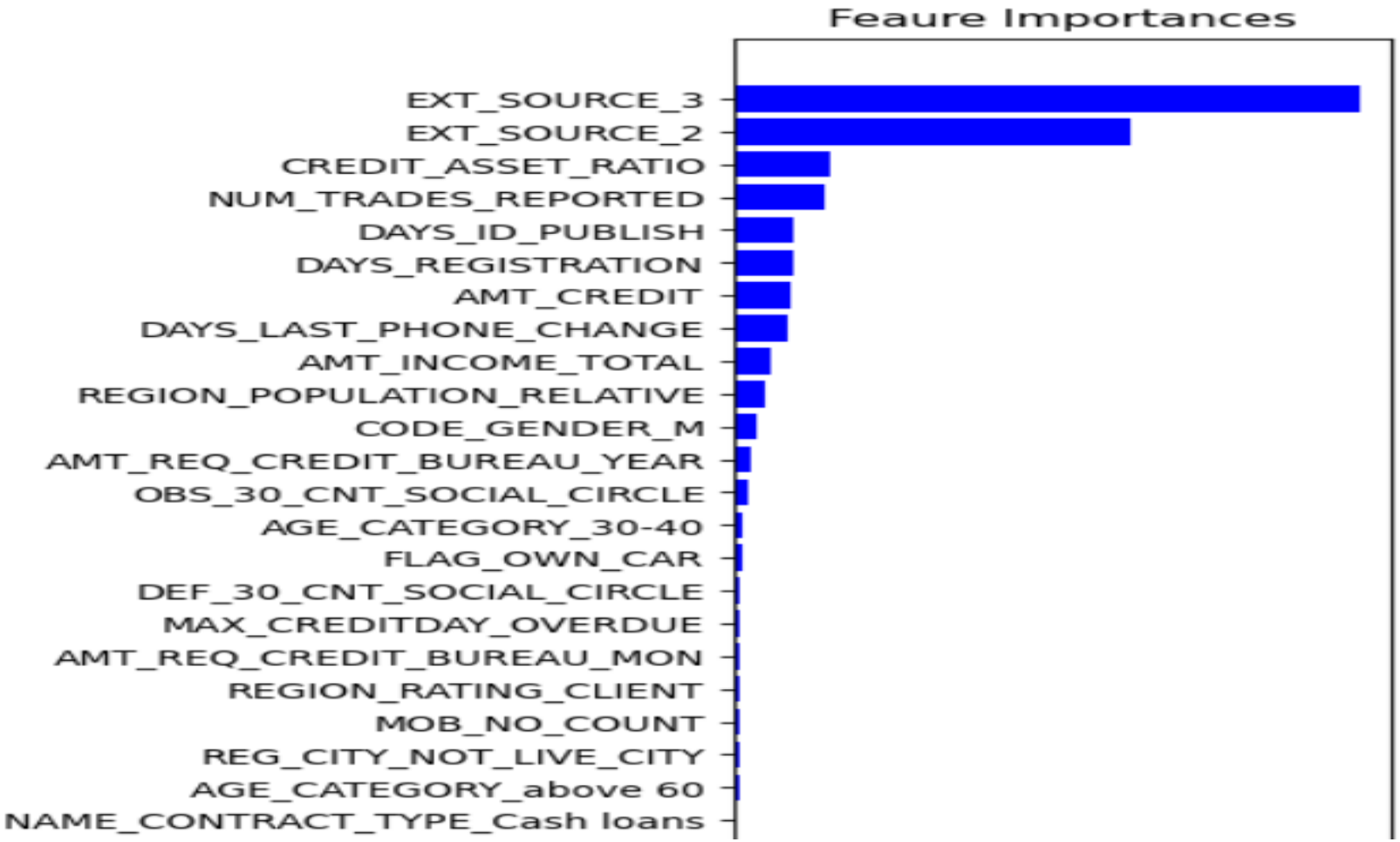


MODEL APPROACH

1. Data Preparation
2. Model Selection
3. Model Training & Prediction
4. Model Evaluation

- ✓ ***Handled Missing Value***: Columns eliminated having Greater than 30% missing values and used Median and Mode Imputation for the rest.
- ✓ ***FEATURE SELECTION & ENGINEERING***: Created 8 Newly Derived features. Used Label Encoding Technique. Dropped Highly Correlated Columns.
- ✓ ***Handling Outliers***: Since we use Random Forest Classifier technique, Not Handled Outliers.
- ✓ ***Model Selection*** : Used Train-Test Split with train size of 0.7, and Model used is Random Forest Classifier
- ✓ ***Handling Unbalanced Data***: There is inbuilt option to handle unbalanced data set in Random Forest Classifier.
- ✓ ***Model Optimization & Fine Tuning:*** Model Optimisation Using GridSearchCV
- ✓ ***Model Evaluation:*** Train &Test Performance using Precision & Recall

Feature Importance



MODEL EVALUATION

Train Data:

- **Precision:** 0.1669 (16.69%)
- **Recall: 0.6743 (67.43%)** (high recall, which is good for credit risk)
- **Accuracy:** 71.13%
- **F1-score:** 27% (due to class imbalance)

Test Data:

- **Precision:** 0.1651 (16.51%)
- **Recall: 0.6684 (66.84%)**
- **Accuracy:** 71.04%
- **F1-score:** 26%

- **High Recall (67%)**: The model correctly identifies most defaulters, reducing **false negatives**, which is the main goal.
- **Low Precision (~16%)**: Many non-defaulters are flagged as defaulters (**false positives**), but this is acceptable in credit risk since avoiding defaulters is the priority.
- **Accuracy (~71%)**: Acceptable given class imbalance, but not the best metric for imbalanced datasets.

Why Recall ?

- ***Banks prioritize identifying defaulters*** over wrongly rejecting some good applicants.
- ***Higher recall (67%)*** ensures fewer defaulters slip through and reduces financial risk.
- ***Precision trade-off is acceptable*** because wrongly rejected applicants can reapply with additional verification.

KEY VARIABLES & BUSINESS INSIGHTS

- **Key Variables Identified from Feature Importance:**

1. **AMT_CREDIT** – Total loan amount requested.
2. **DAYS_BIRTH** – Age of the applicant in days (negative values).
3. **EXT_SOURCE_1, EXT_SOURCE_2** – External credit bureau risk scores.
4. **DAYS_EMPLOYED** – Employment duration (negative values indicate errors).
5. **CREDIT_OVERDUE_COUNT** – Number of overdue payments.
6. **NAME_CONTRACT_TYPE** – Loan type (cash loans show higher defaults).
7. **CODE_GENDER** – Males have proportionately higher default risk.
8. **INCOME_CAT** – Middle-income category shows more defaults.
9. **AGE_CATEGORY** – 20-30 years group has higher default proportion.
10. **CREDIT_ACTIVE** – Active credit accounts increase risk.

Business Insights & Recommendations

Business Recommendations for Immediate Implementation:

1. Strengthen Decision-Making Using Top Features

- Implement **rule-based filters** for loan approval using top predictors:
  - Set **minimum cutoffs** for **income levels and employment duration**.
  - Reject applications with **negative DAYS_EMPLOYED** (indicates data issues).

2. Address High-Risk Segments

- Customers with **high AMT_CREDIT but low AMT_INCOME_TOTAL** need **stricter credit checks**.
- Borrowers with a **history of missed payments** (CREDIT_OVERDUE_COUNT >1) should be flagged.
- Apply **higher scrutiny for younger borrowers (20-30 age group)** due to higher default rates.

3. Optimize Credit Approval Strategy

- **Adjust probability cutoffs** for classification to balance risk vs. approval rates.
- Implement **preliminary risk scores**:
  - **Low-risk applicants** → Auto-approve.
  - **Moderate risk** → Manual review.
  - **High risk** → Reject or require additional verification.

4. Improve Data Quality & Feature Engineering

- **Transform DAYS_EMPLOYED** (negative values should be categorized correctly).
- **Handle correlated variables** (drop redundant ones).
- **Standardize missing value handling** for **EXT_SOURCE_3**.

5. Immediate Model Refinements

- **Drop low-importance features** to improve model efficiency.
- **Retrain the model** using only the most predictive variables.
- Consider **ensemble methods** (e.g., boosting) to improve recall