

HOME CREDIT SCORECARD MODEL



TABLE OF CONTENTS

01

PROBLEM
RESEARCH

02

DATA
PREPROCESSING

03

BUSINESS
INSIGHTS

04

CLASS IMBALANCE

05

MACHINE LEARNING
MODEL &
RECOMMENDATION

01

PROBLEM

RESEARCH



CASE STUDY BACKGROUND

The goal of this case study is to help Home Credit (NBFC) to determine which loan applications should be approved or rejected based on applicant behavior and application data. As a business analyst, we are cleaning the dataset and aggregate trade-level bureau data, apply feature engineering, build a classification model, and derive business insights to inform decision-making strategies for the bank.

DATA SOURCE

The data used are **application.csv** and **bureau.csv**. "SK_ID_CURR" is the unique identifier for Application dataset and "SK_ID_BUREAU" is unique identifier for the bureau dataset.

OBJECTIVE

Building a model to differentiate applicants
Loan Application approval /rejection

ACTIONS

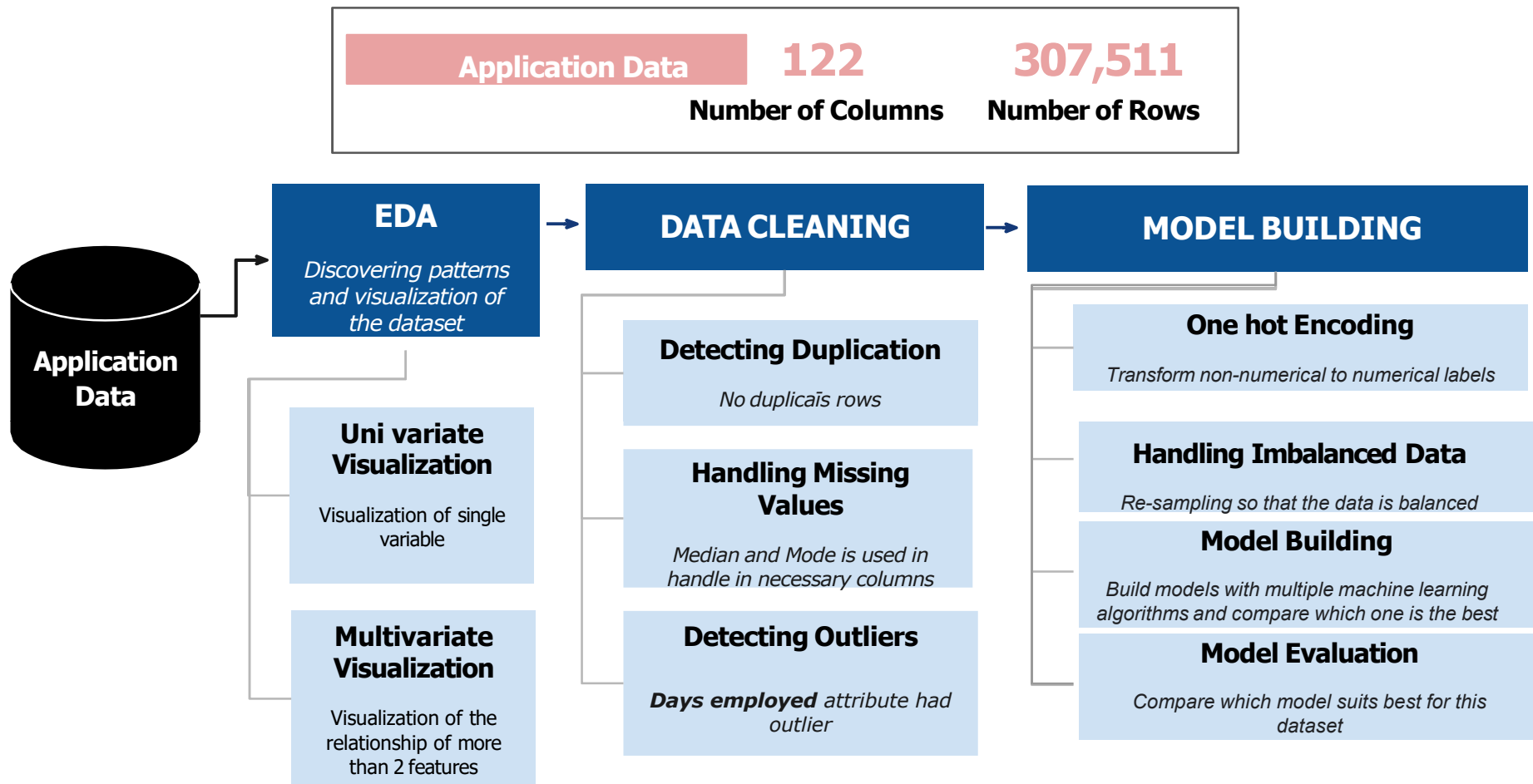
1. Conduct data preparation and visualization to derive business insights.
2. Implement feature engineering techniques.
3. Build and evaluate the model.
4. Offer recommendations to help the company improve the success rate of clients' loan applications.



02

DATA

PREPROCESSING

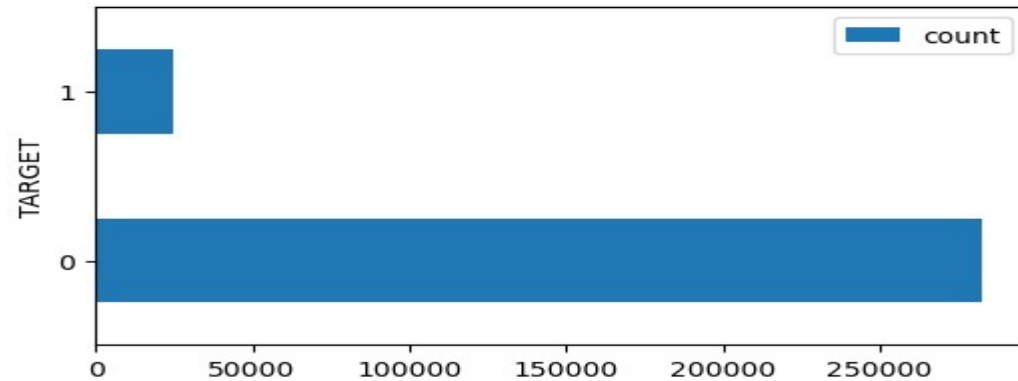


03

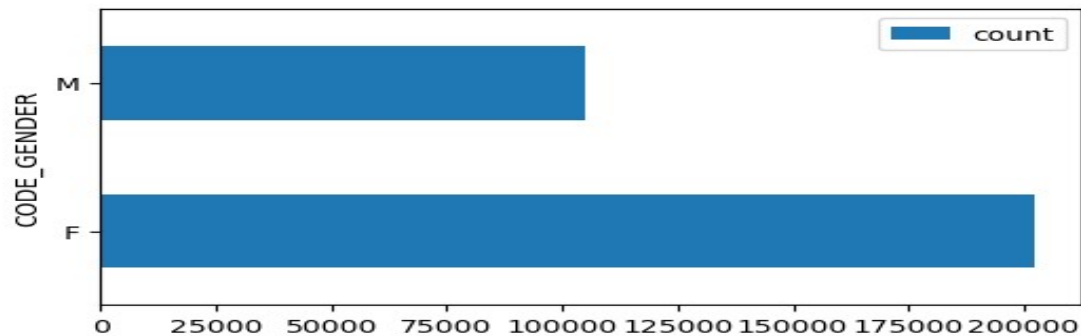
BUSINESS INSIGHTS




```
TARGET
0    91.928194
1     8.071806
Name: count, dtype: float64 2
```



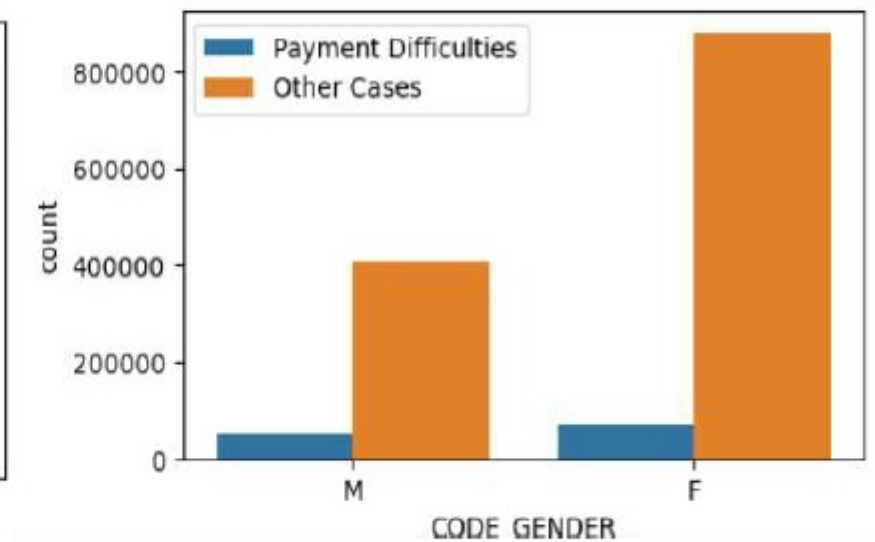
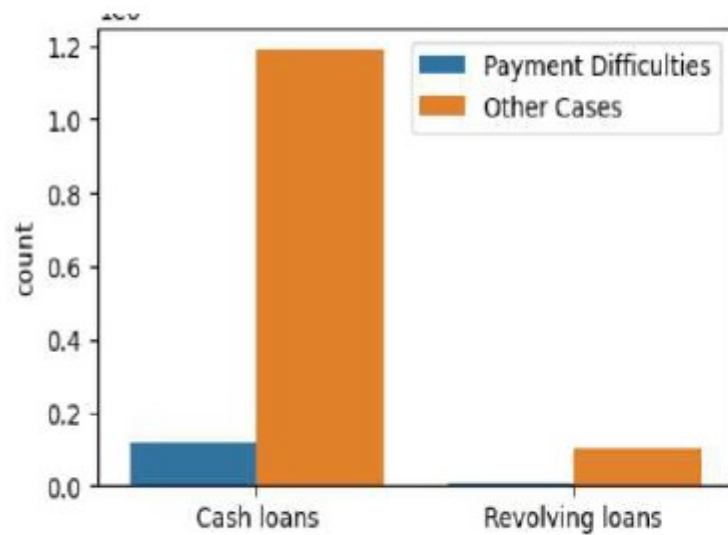
```
CODE_GENDER
F    65.842143
M    34.157857
Name: count, dtype: float64 2
```



- More than 91.92% applicants are not facing any payment difficulties.
- Meanwhile, 8.07% people are facing payment difficulties

- Higher Loan Applications Based on Gender.
- A greater number of females have applied for loans compared to males.
- This suggests a higher level of financial dependency among females.

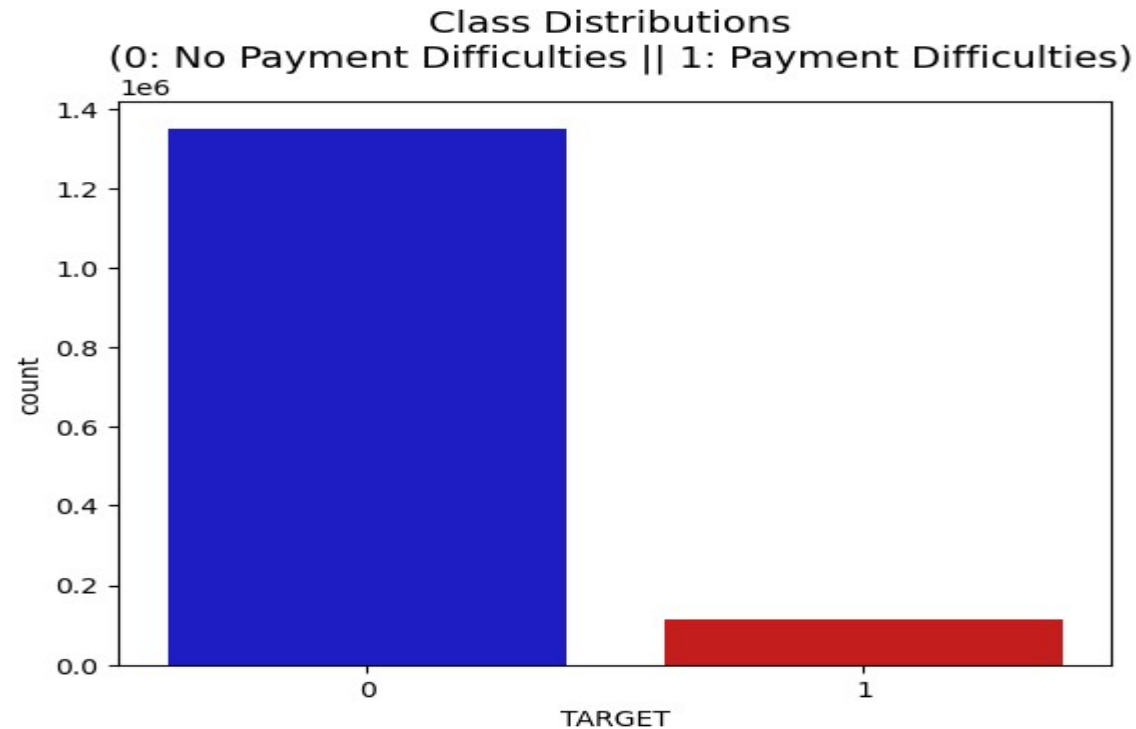
Bi-variate Data Imbalance for TARGET variable:



04

HANDLING CLASS IMBALANCE





We can see that applications with payment difficulties is 7.8%. Therefore this is highly imbalanced. It is important to handle imbalanced classes as challenges related to imbalanced dataset are: 1. Biased predictions 2. Misleading accuracy

Handling Class Imbalance

Class imbalance occurs when the number of observations in one class is significantly higher than in the other. This can lead to biased models that favor the majority class.

Challenges of Class Imbalance

- Model predicts majority class more frequently, ignoring minority class.
- Poor recall (high false negatives) for the minority class.
- Misleading accuracy as the model favors the dominant class.

Techniques to Handle Class Imbalance

1.Resampling Methods

- Oversampling:** Increase minority class samples (e.g., SMOTE – Synthetic Minority Over-sampling Technique).
- Undersampling:** Reduce majority class samples to balance the dataset.

2.Algorithm-Level Approaches

- Use models that handle imbalance well, such as Decision Trees, Random Forests, etc with class weighting.
- Assign higher misclassification costs to the minority class (e.g., class weight in logistic regression, SVM).

3.Evaluation Metrics for Imbalanced Data

- Precision-Recall Curve:** Better than ROC-AUC for imbalanced datasets.
- F1-Score:** Harmonic mean of precision and recall to balance both aspects.
- Balanced Accuracy:** Adjusted to account for imbalance.

4.Anomaly Detection Approaches

- If the minority class represents rare cases (e.g., fraud detection), treat it as an anomaly detection problem.

5.Synthetic Data Generation

- Techniques like **SMOTE**, **ADASYN (Adaptive Synthetic Sampling)** can generate synthetic examples for better balance.

05

MACHINE LEARNING MODEL



MODEL COMPARISON

Algorithm	Accuracy	Precision	Recall	F1 Score
Logistic Reg. Random OS	0.59	0.10	0.53	0.17
Decision Tree Random OS	0.98	0.85	0.86	0.85
Decision Tree ADASYN	0.98	0.85	0.92	0.88
Random Forest OS	0.99	1.00	0.84	0.91
Random Forest ADASYB	0.99	1.00	0.84	0.91

Based on above comparison, we can conclude that Random Forest OS and Random Forest ADASYB model seems fitting for our model.

To evaluate the best-fit model for a BFSI (Banking, Financial Services, and Insurance) case study, we should prioritize the following metrics based on typical industry needs:

Accuracy – Ensures overall correctness.

Precision – Important in BFSI to minimize false positives, especially for fraud detection.

Recall – Important when minimizing false negatives, especially in credit risk scenarios.

F1 Score – Balances precision and recall, useful for imbalanced datasets.

Analysis:

- **Logistic Regression:**

Poor performance across all metrics except the base model's accuracy (0.92). Precision, recall, and F1 scores are very low.

- **Decision Tree:**

Strong recall (up to 0.92) and improved precision when oversampling methods like Random OS and ADASYN are applied. Accuracy is consistently high (0.98) with ADASYN performing the best (F1 score of 0.88).

- **Random Forest:**


Achieves the best overall performance with high accuracy (up to 0.99).

ADASYN and Random OS both deliver strong precision (1.00) and balanced recall (0.84).

The F1 scores for ADASYN and Random OS (both 0.91) are the highest in the table.

Conclusion:

Random Forest with ADASYN or Random OS is the best-performing model for the BFSI case study, providing the highest accuracy, precision, and a strong balance between recall and F1 score. This combination offers robust performance, crucial for identifying potential risks, fraud, or credit defaults in the financial sector.



THANK

YOU