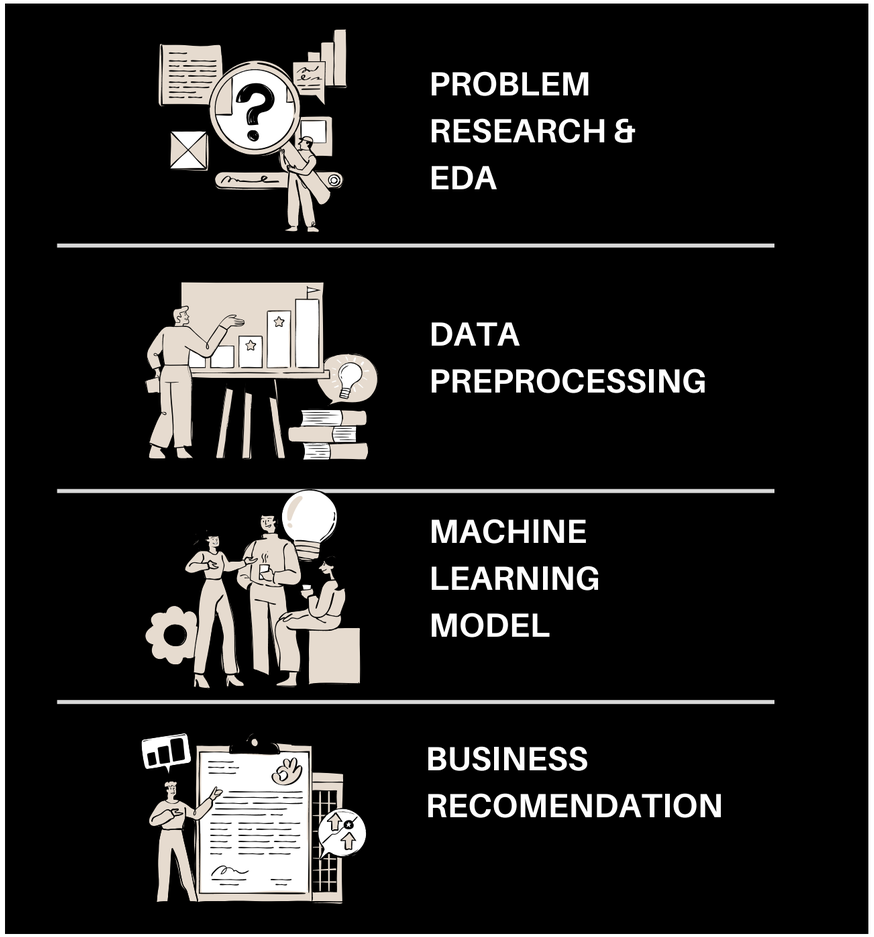
HOME CREDIT SCORECARD MODEL PREDICTION

CONTENT

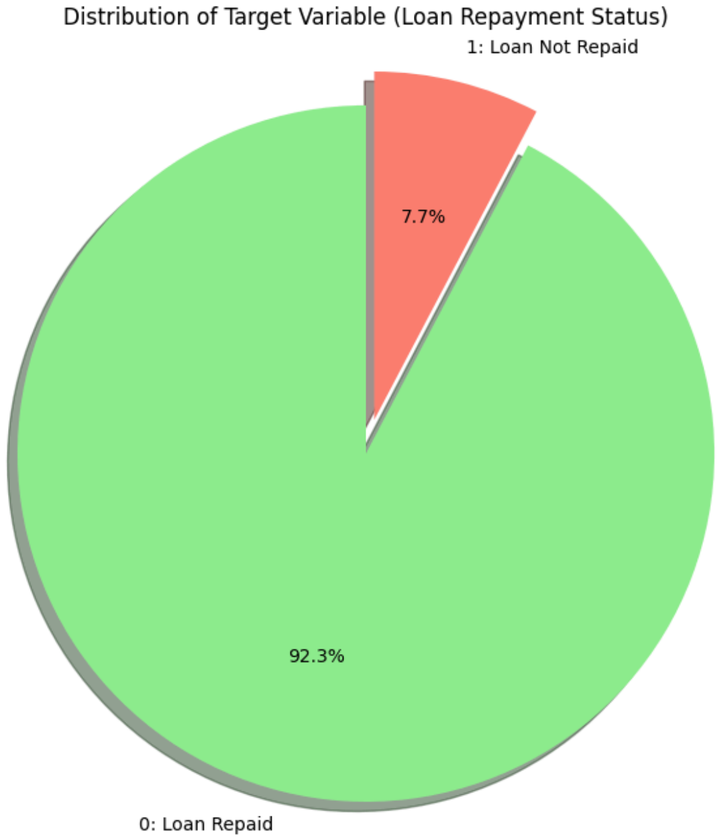

PROBLEM
RESEARCH &
EDA
DATA
PREPROCESSING
MACHINE
LEARNING
MODEL
BUSINESS
RECOMENDATION

PROBLEM RESEARCH

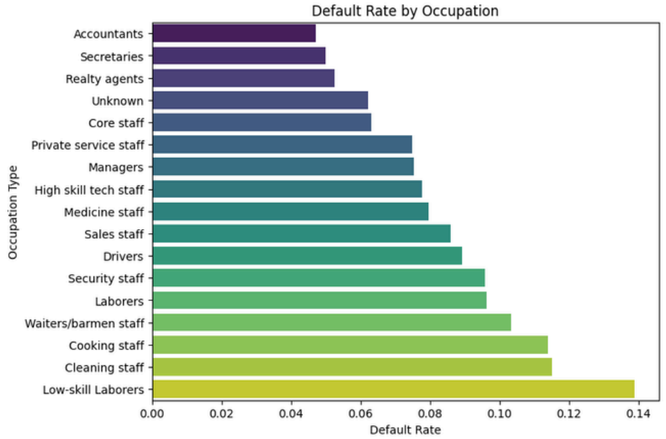
Home Credit faces a high level of loan repayment failure, which can lead to significant financial losses.

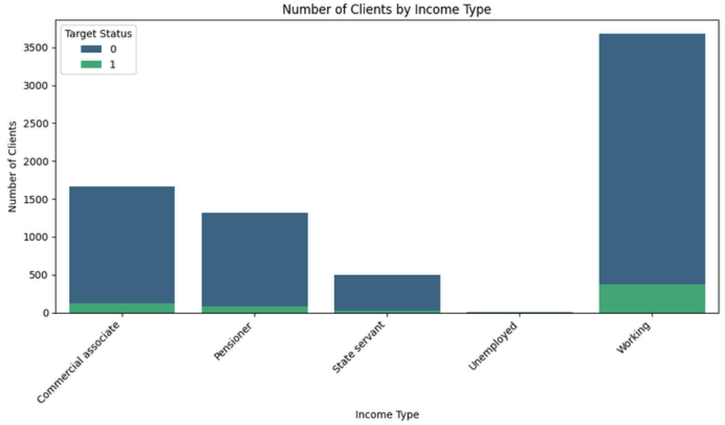
- Approximately 7.7% of customers experience repayment failure or late payments
- Despite the small proportion, these cases contribute significantly to financial losses

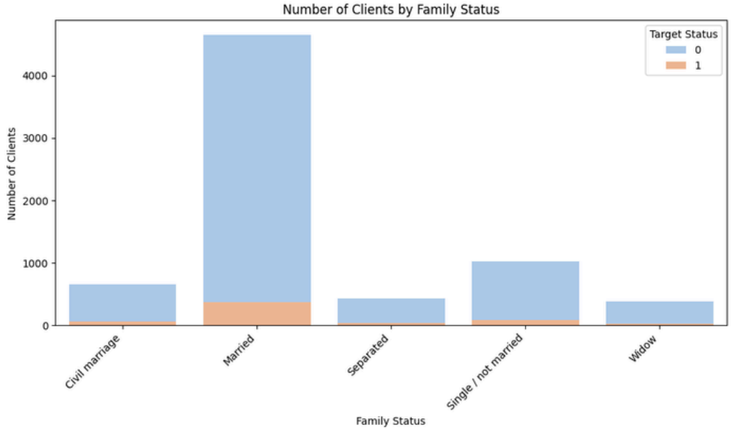
The company needs a more accurate risk assessment approach to reduce repayment failure, while still maintaining access to credit for customers who are financially eligible.

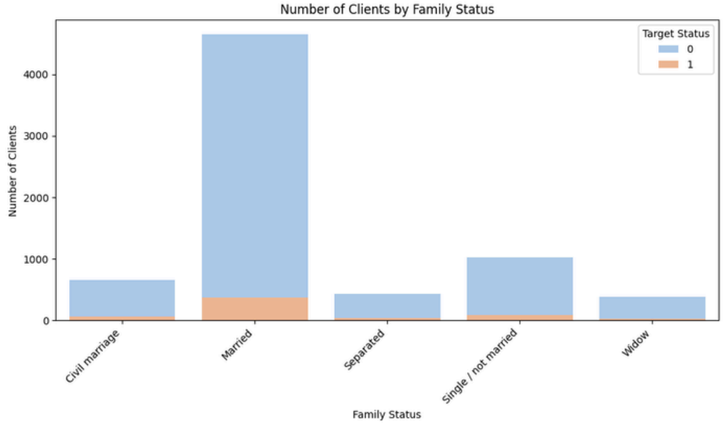


EDA


Default Rate by Occupation
Accountants
Secretaries
Realty agents
Unknown
Core staff
Private service staff
Managers
High skill tech staff
Medicine staff
Sales staff
Drivers
Security staff
Laborers
Waiters/barmen staff
Cooking staff
Cleaning staff
Low-skill Laborers
Occupation Type
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Default Rate


Number of Clients by Income Type
Target Status
0
1
Number of Clients
0
500
1000
1500
2000
2500
3000
3500
Commercial associate
Pensioner
State servant
Unemployed
Working
Income Type


Number of Clients by Family Status
Target Status
0
1
Number of Clients
0
1000
2000
3000
4000
Civil marriage
Married
Separated
Single / not married
Widow
Family Status


Number of Clients by Family Status
Target Status
0
1
Number of Clients
0
1000
2000
3000
4000
Civil marriage
Married
Separated
Single / not married
Widow
Family Status

PREPROCESSING

Handling Missing Value

- Remove features with more than 30% missing values
- A total of 49 columns were removed due to excessive missing data
- Impute missing values in numerical features using median
- Impute missing values in categorical features using most frequent value

Encoding Categorical Variables

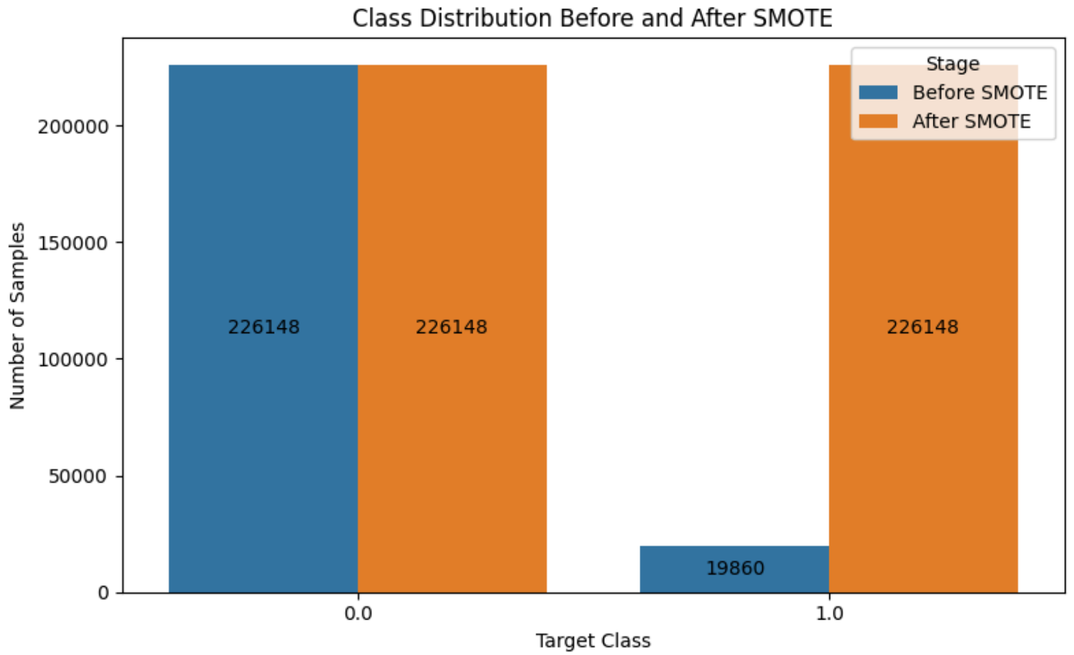
- Identify all categorical features with object data type
- Apply Label Encoding to convert categories into numerical values
- Encode each categorical column separately
- Store label encoders for each feature to ensure consistent transformation

PREPROCESSING

Data Splitting

- Separate features and target variable
- Remove identifier column (SK_ID_CURR) from the feature set
- Split data into training (80%) and testing (20%) sets
- Apply stratified splitting to preserve the original class distribution

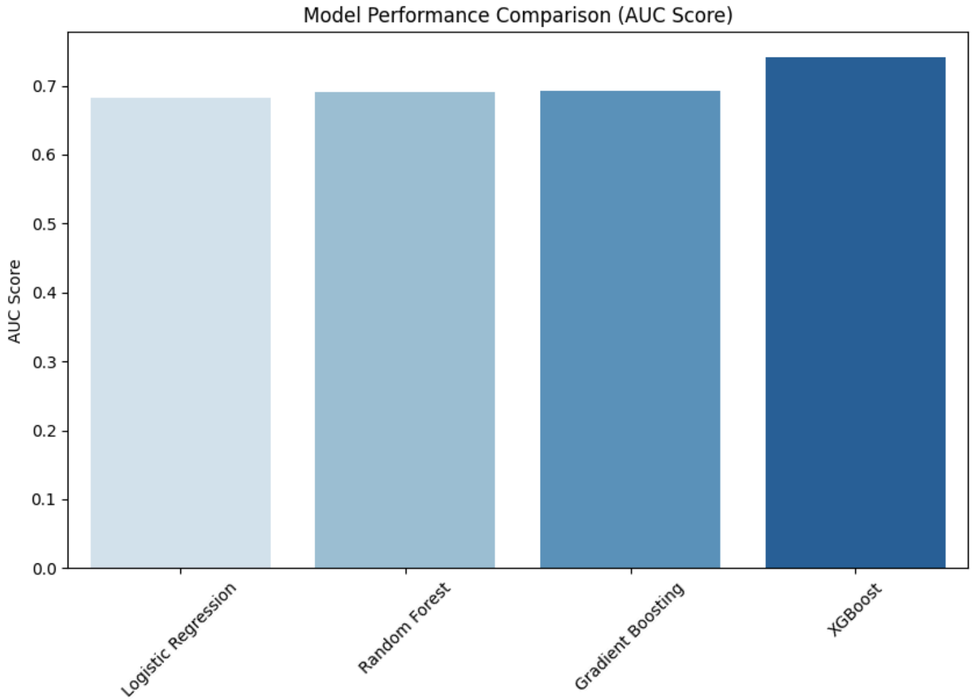
Handling Imbalaced Data with SMOTE

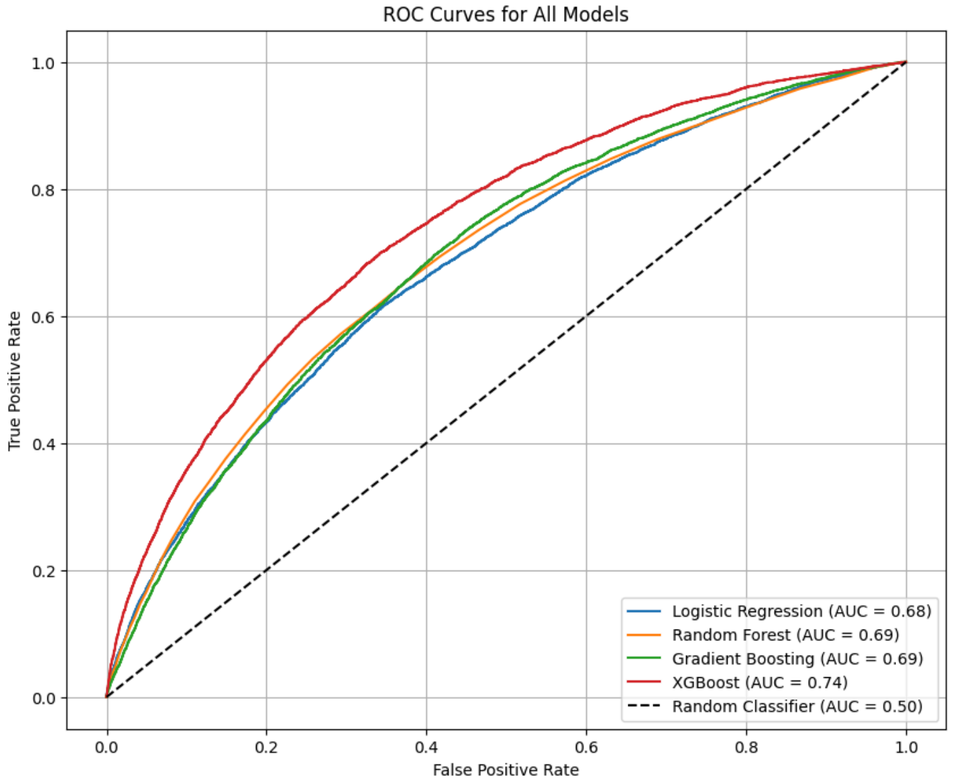


SMOTE is applied to balance the dataset by increasing minority class samples, reducing model bias, and improving classification performance.

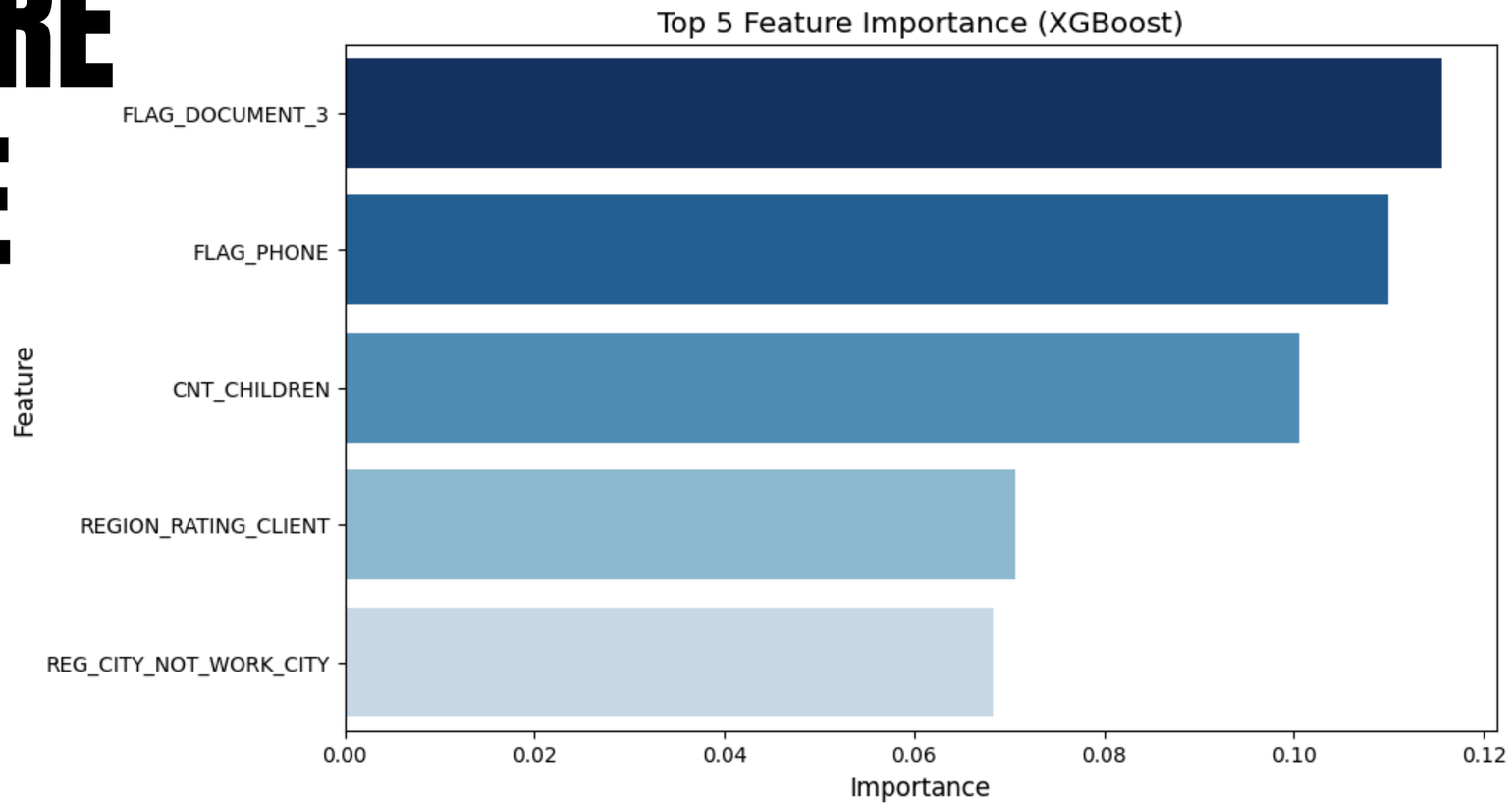
MODELLING

Based on both AUC scores and ROC curves, XGBoost was selected as the best-performing model.





TOP 5 FEATURE IMPORTANCE



BUSINESS RECOMENDATION

1. Strengthen Credit Risk Assessment

- Prioritize and validate key predictive features identified by the model
- Focus on accurate phone information, address consistency, document completeness, regional risk, and social exposure indicators
- Use these variables as critical inputs in the loan application process

2. Implement Risk-Based Loan Product Strategy

- Design tailored loan terms for higher-risk occupation groups
- Offer more competitive terms to lower-risk occupations and demographics
- Incorporate income type, education level, family status, and housing type into risk profiling

3. Leverage the Best-Performing Model

- Integrate XGBoost as the primary credit scoring model due to its strongest predictive performance
- Prioritize evaluation metrics that focus on identifying high-risk customers
- Continuously monitor and retrain the model to adapt to changing conditions

4. Improve Data Quality and Collection

- Enhance data collection processes to reduce missing values
- Focus on improving completeness and accuracy of high-impact features
- Minimize reliance on data imputation to reduce potential bias

THANK YOU

GITHUB

LINKEDIN

E-MAIL

TANJUNG DUREN, JAKARTA BARAT