Homecredit Final Project

Mohammad Daffa Aditya

Table of Content

Problem Research

Background

Home Credit Indonesia aims to enhance its credit scoring process by leveraging machine learning and statistical methods to assess loan applications more accurately. The goal is to predict whether customers will encounter repayment difficulties, ensuring that financially capable applicants are not unnecessarily rejected. By analyzing a diverse range of data sources, Home Credit seeks to identify key customer characteristics associated with successful loan repayments.

Data Source

The dataset was provide by Home Credit Indonesia. The data consisting 307,551 rows and 122 column containing customer data.

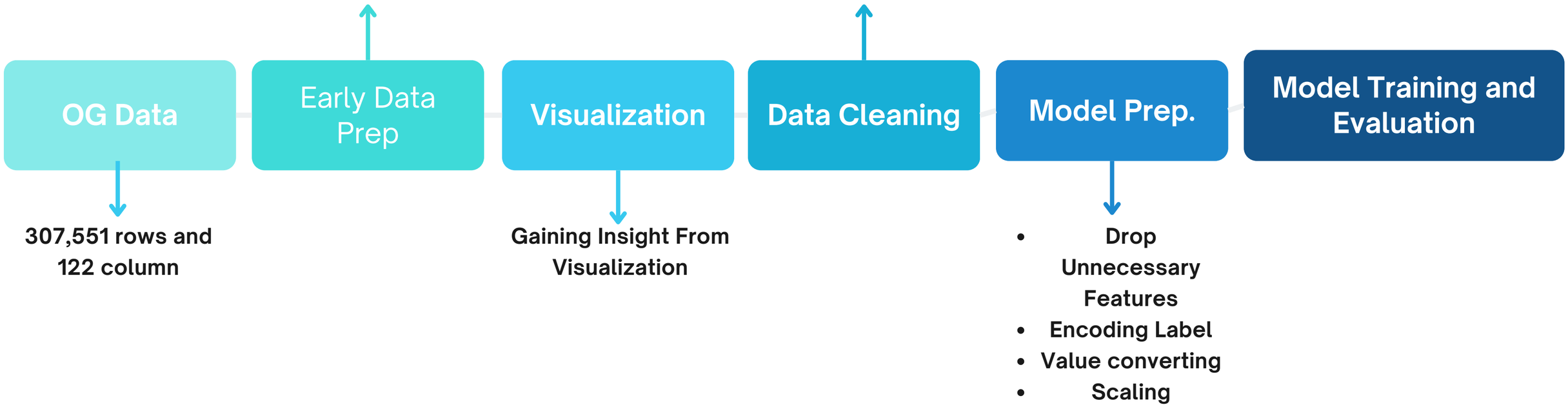
Objective

- Analyse characteristics of possible customers who could be having issues repaying loans.
- Assess customers payback abilities utilizing past data and that are relevant features.

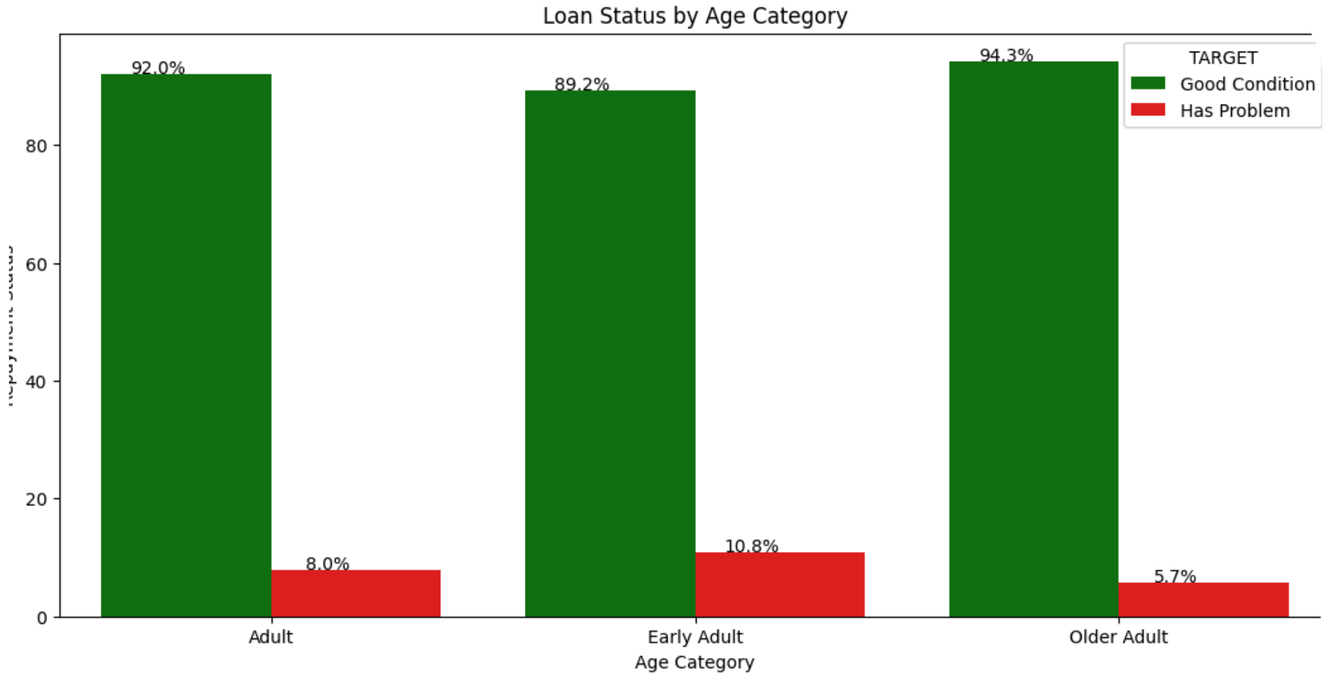
Action

- Perform data preparation and Exploration Data Analysis (EDA).
- Build visualization in order to gaining insights
- Build Machine Learning (ML) models
- Provide business recommendation

Data Preprocessing



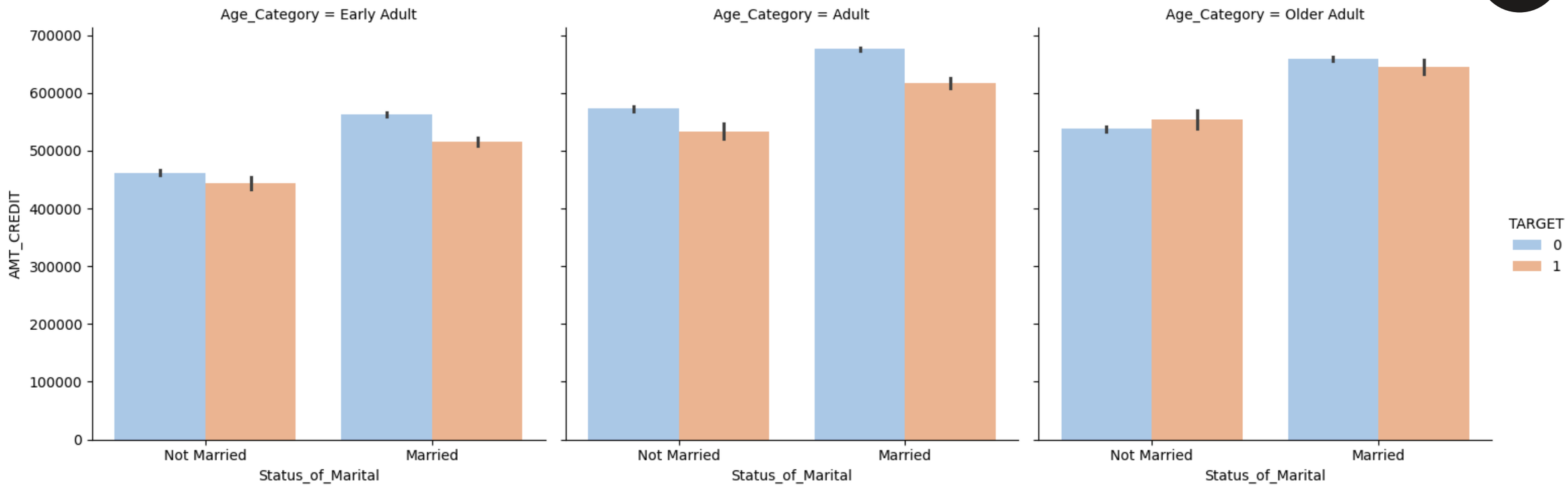
Insights



	Age_Category	TARGET	Customers	Percentage
0	Adult	Good Condition	108677	92.036755
1	Adult	Has Problem	9403	7.963245
2	Early Adult	Good Condition	82286	89.209553
3	Early Adult	Has Problem	9953	10.790447
4	Older Adult	Good Condition	86877	94.250193
5	Older Adult	Has Problem	5300	5.749807

- Most of customer who apply for loan was in age group 35-50
- Older adults have the highest repayment rate (94.3%), while early adults have the lowest (89.2%)
- Which means that the group with higher probability of having repayment problem was early adult <= 35

Insights

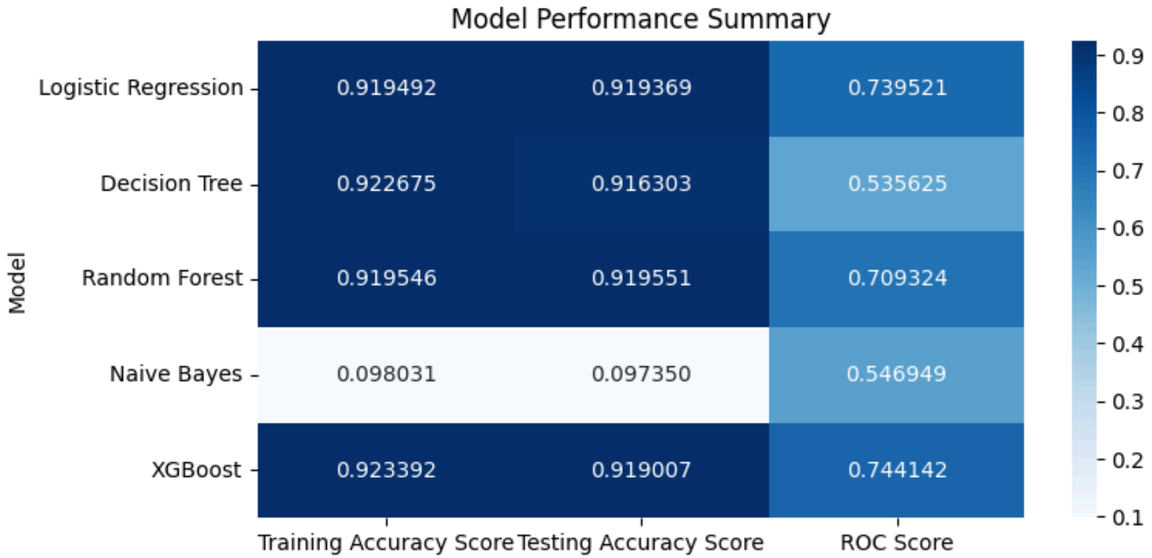


- Adults and Older Adults tend to have higher credit amounts compared to Early Adults, likely due to increased financial stability and borrowing capacity.
- In Early Adults and Adults, Married individuals tend to have higher credit amounts than Not Married individuals.In Older Adults, the credit amount is similar between Married and Not Married.
- Overall, Married individuals tend to receive higher credit amounts compared to non-married individuals.

More Insights:

- Low-skilled labor has the highest tendency for repayment issues.
- Overall loan status considered inhealthy condition (92% good standing, 8% has problem).

ML Models & Evaluation +



- Logistic Regression, Decision Tree, Random Forest, and XGBoost have similar accuracy scores (~91-92%) on both training and testing data.
- XGBoost (0.744) and Logistic Regression (0.739) have the highest ROC scores, making them the best-performing models in distinguishing between classes.
- XGBoost is the best choice due to its highest ROC score (0.744), strong accuracy, and robustness. Conversely, Naïve Bayes completely fails on this dataset, both in accuracy and ROC score.

Business Recommendation

- Personalized Loan Products: Offer younger customers flexible loan terms depending on their financial capabilities.
- Improvement of Scoring and Segmentation: Segment clients according to risk levels (Low, Medium, High) using modeling outcomes, and consider education level and employment type essential considerations in loan approval processes.
- To minimize risk, offer secured loans or lower credit limits to high-risk clients (those with a high likelihood of default); for low-risk customers, offer greater limits on credit and personalized offers to increase retention and revenue.
- Automate risk assessments to increase efficiency and minimize manual intervention.
- Use machine learning to enhance loan decisions and continually develop models in adaptation to evolving market conditions and financial behaviors by implementing a data-driven risk assessment strategy.

*Further Analysis Recommended for Better Results

Thank You!

See the Entire Documentation Here:

Github Repository