HOME CREDIT SCORECARD MODEL

PRESENTED BY NICHOLAS MARCO WEINANDRA

Workflow

For this final task, as a Data Scientist Intern at Home Credit Indonesia, I was assigned to perform analysis and make predictions on consumer credit applications at Home Credit. The objective of this project is to analyze loan data and build predictive Machine Learning models to generate strategic insights.

Problem Research

Mengidentifikasi masalah apa yang ingin diselesaikan

Data Preprocessing

Melakukan Data Cleaning, Feature Engineering, dan Feature Selection

Data Visualization & Business Insight

Memvisualisasikan Data dan memberikan rekomendasi keputusan bisnis

Machine Learning Implementation and Evaluation

Melakukan training model ML dan mengevaluasi hasilnya

Business Recommendation

Memberikan rekomendasi bisnis berdasarkan hasil prediksi

Problem Research

Problem Statement

Home Credit Indonesia aims to enhance its **credit risk assessment** process by leveraging data-driven insights. As part of this initiative, the objective is to analyze historical consumer loan data and **develop a predictive model** to determine **the ability of a customer defaulting on a loan**. This will support the company in making more accurate, efficient, and strategic credit approval decisions, ultimately minimizing financial risk while maximizing loan approval opportunities for trustworthy customers.

Data Source

- **application_train** : Main application data for training. Contains Demographic and financial information about the applicant such as income, age, employment status, family status, and the loan amount requested. The target feature is 'TARGET' that represents **value 1** (**Clients with payment difficulties**) and **value 0** (**Clients with all other cases**)

- **application_test** : Main application data for testing. Contains the same feature as the train dataset but without the 'TARGET' feature.

Objectives

- **Analyze Consumer Loan Applications**

Explore and understand the patterns, trends, and characteristics in the historical loan application data from Home Credit Indonesia.

- **Identify Key Risk Indicators**

Determine the most relevant features and customer behaviors that influence loan repayment performance and default risk.

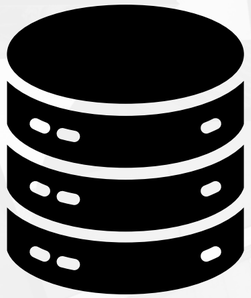
- **Build a Predictive Machine Learning Model**

Develop and evaluate a classification model to predict whether a loan applicant is likely to repay or default on the loan.

- **Support Strategic Decision-Making**

Provide actionable insights to assist the credit risk and underwriting teams in making more accurate, data-driven loan approval decisions.

Data Preprocessing



application_train

307,511	122
Rows	Columns

EDA

Understand the structure, quality, and patterns within the data

Data Visualization

Identify trends, discover new insights between multiple features

DATA CLEANING

Handle Missing Values

Use imputation method and drop columns with excessively high percentage of missing values

Check Duplicate Data

There are no duplicate data found

Check Outliers

Outliers were found but it was decided not to be handled

FEATURE ENGINEERING

Encoding

Encode all categorical columns to be trained by the model

Feature Selection

Normalize numerical features to a standard range

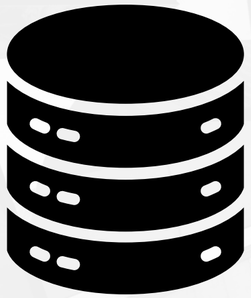
Scaling / Normalizing

Normalize numerical features to a standard range

Handling Class Imbalance

Apply oversampling to class minority

Data Preprocessing



application_test

48,744	121
Rows	Columns

DATA CLEANING

Handle Missing Values

Use imputation method and drop columns with excessively high percentage of missing values

Check Duplicate Data

There are no duplicate data found

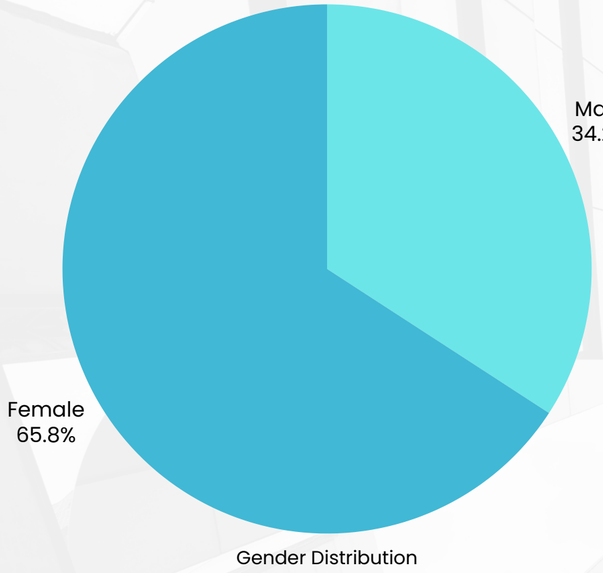
Encoding

Encode all categorical columns to be trained by the model

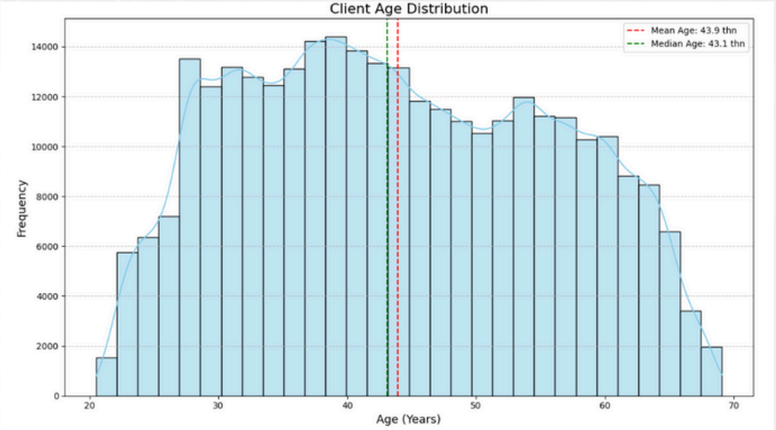
PREDICTION

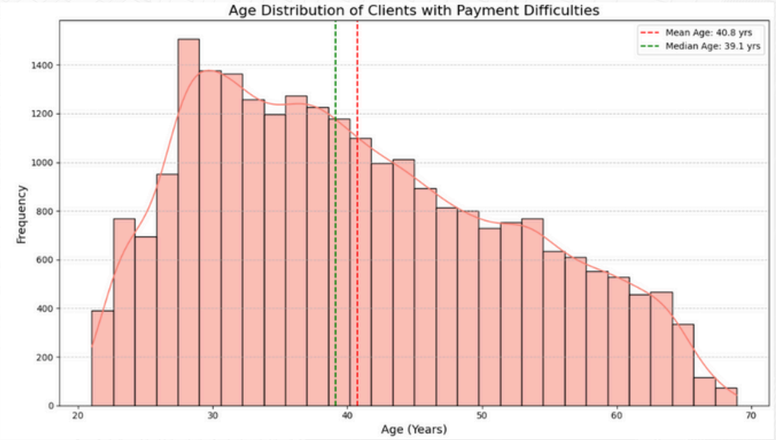
Predict the ability of a customer defaulting on a loan

Business Insights



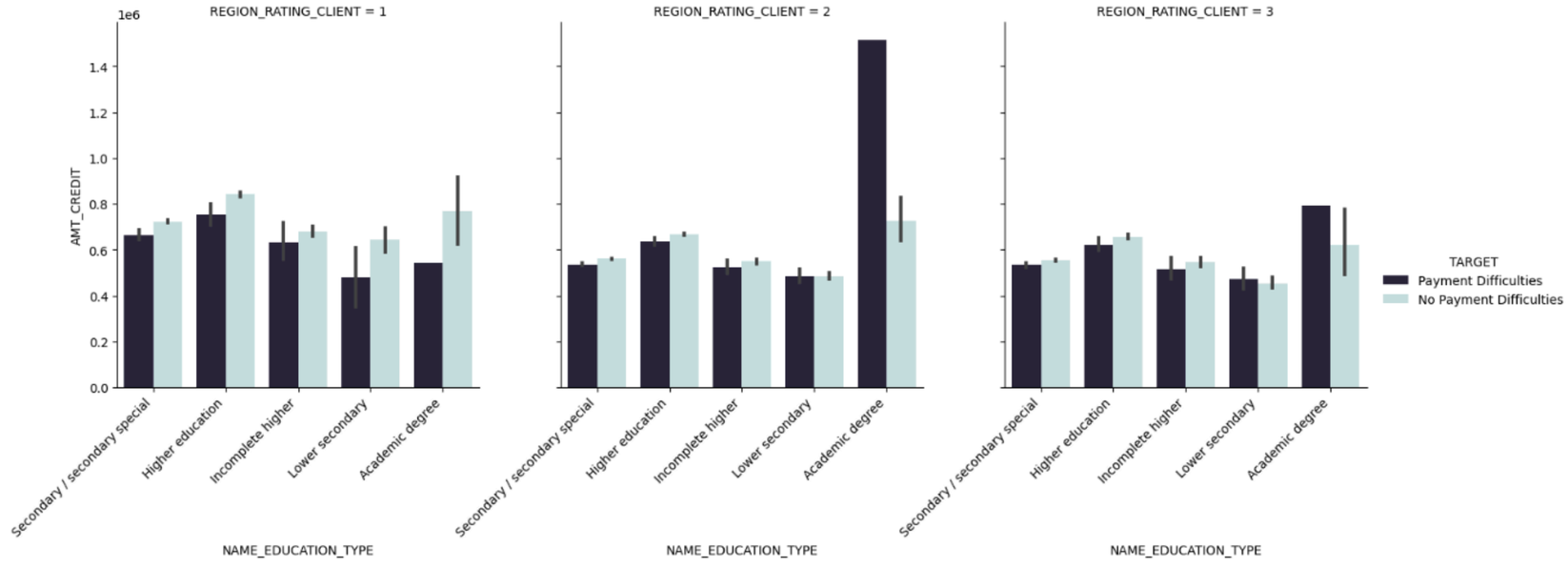
Female applicants make up **65.8%**, while Male applicants account for 34.2% of total loan applications. This indicates that women are the majority of credit applicants in this dataset, suggesting a potentially more **active participation of female customers** in seeking consumer financing with Home Credit Indonesia.





- Most number of clients who apply for loans are in the range of **35-40 years old.**
- Client's **mean age is 43.9 years old** with the **median age is 43.1 years** old
- Meanwhile, the number of applicants for clients aged <25 or age >65 is very low.
- Clients with **no payment difficulties** are mostly consisted of those in the range of **35-40 years old**. These customer segmentation should be your priority.
- While clients with **payment difficulties** are younger clients that might be only starting their career, those in the range of **25-35 years old**,.

Business Insights



Applicants with higher education or an academic degree generally request larger credit amounts compared to those with lower education levels (like "Secondary" or "Lower secondary").

This is expected, as higher education often correlates with better job prospects and income, enabling higher credit eligibility.

Interestingly, for clients with an academic degree in Region Rating 2, those who have payment difficulties had a significantly higher average credit amount than those who did not.

This could indicate:

- **Riskier lending behavior to highly educated individuals in mid-rated regions.**
- **Or possibly overestimation of creditworthiness based on education alone.**

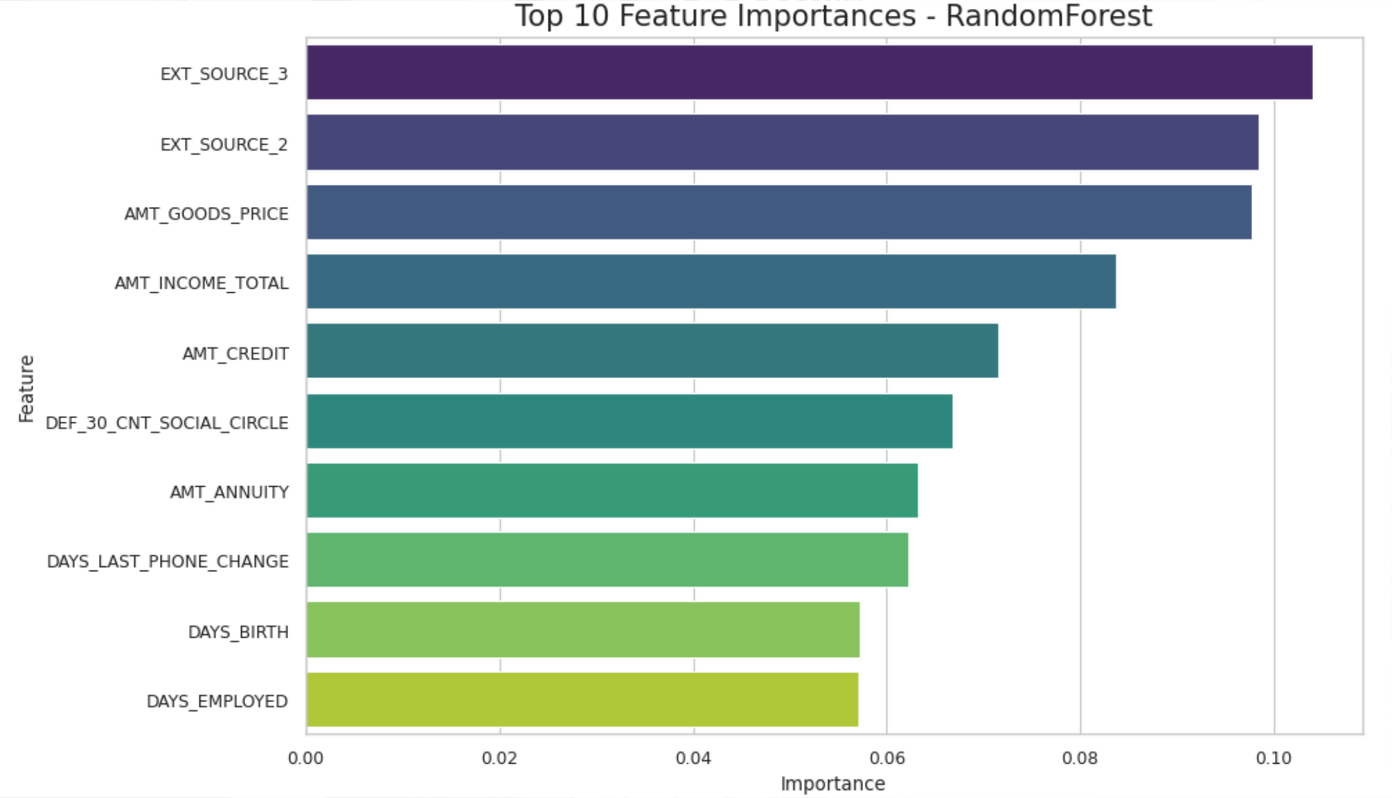
ML Implementation

Algorithm	Accuracy	Macro Average F1 Score	ROC AUC Score
Logistic Regression	0.58	0.58	0.50
Random Forest Classifier	0.93	0.93	0.98
Gaussian Naive Bayes	0.63	0.63	0.69
Decision Tree	0.83	0.83	0.83
K - Nearest Neighbors	0.85	0.85	0.94

From the evaluation table, it is clear that **Random Forest Classifier** significantly outperforms the other models across all three metrics:

This indicates that Random Forest not only predicts the majority class **accurately** but also **handles class imbalance effectively**, maintaining high precision and recall for both classes — particularly crucial for detecting cases with payment difficulties.

ML Implementation



Algorithm → **Random Forest Classifier**

Performance
- Random Forest model gives **93% accuracy**
- **98%** ROC-AUC Score

Feature Importance
- Score from external data source 3
- Score from external data source 2
- The price of the goods for which the loan is given

Business Recommendation

- **Maternity Leave Segment**: Current practices effectively bar clients on maternity leave from cash loans due to a 100% rejection rate, presumably due to repayment concerns. Revolving loans are the only accessible credit product for them.

- **Unemployed Segment**: Cash loans are problematic for unemployed clients, with a majority experiencing repayment issues. Revolving loans, on the other hand, perform well with this segment, showing full repayment capability. This suggests a strong preference for offering revolving loans to unemployed individuals.

Github Link: https://github.com/nicholasmarco27/Rakamin-Home-Credit-Scorecard-Model