HOME CREDIT SCORECARD MODEL

PROJECT-BASED VIRTUAL INTERN: DATA SCIENTIST - HOME CREDIT INDONESIA X RAKAMIN ACADEMY

Johanes Purek
March 31st, 2025

TABLE OF CONTENTS

PROBLEM RESEARCH

- Credit Access Challenges: Many individuals struggle to get loans due to insufficient or non-existent credit histories.
- Limitations of Traditional Scoring: Standard methods fail to assess repayment abilities for those without formal financial records.
- ML-Based Solution: Leverage statistical techniques and Machine Learning to predict and assess customers' loan repayment capabilities.
- Ensuring Fair Lending: Help Home Credit Indonesia approve creditworthy clients while structuring loans for their success, with additional background checks when needed.

Goals & Objective

1. Identify key characteristics of clients likely to face repayment difficulties.

2. Develop predictive models to assess clients' repayment abilities.

3. Ensure fair and data-driven loan approval decisions.

4. Provide actionable insights to improve loan structuring and customer experience.

Actions & Metrics

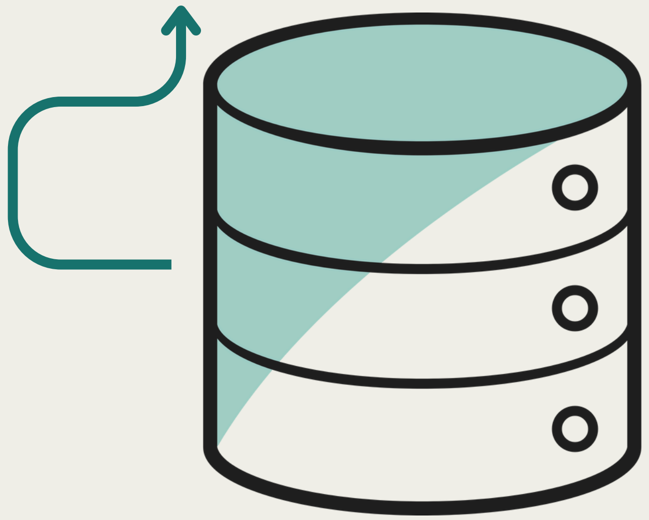
1. Data Exploration & Preprocessing: Clean and visualize data to extract meaningful business insights.
Metric: Completeness and accuracy of preprocessed data.

2. Model Development & Evaluation: Build at least two machine learning models, including Logistic Regression, to predict repayment risk. Compare their performance to identify the best approach.
Metric: Accuracy, ROC-AUC score, and Confusion Matrix analysis.

3. Business Recommendations: Provide insights on how Home Credit can improve loan approvals while minimizing risk.
Metric: Feasibility and impact of proposed strategies.

DATASET OVERVIEW

contains from other mix dataset



Application Train Dataset (application_train.csv)

122 *columns*

consist of many numerical and categorical values

column ...	column ...	column 'TARGET'
row ...	row ...	row ...
row ...	row ...	row ...

307.511 *rows*

TARGET

Unique values:
0 : Clients without Payment Difficulties
1: Clients with Payment Difficulties

Application Test Dataset (application_test.csv)

121 *columns*

column ...	column ...
row ...	row ...
row ...	row ...

48.744 *rows*

NOT HAVING TARGET

PROJECT WORKFLOW

Import Library, Load Dataset Train & Test

- Check Data Type and Data Shape

Data Preparation & Cleaning

- Dataset Train Summary Statistics
- Handling Missing Values by drop column with >50% missing values & impute using median and mode
- Check data duplication
- Detecting outliers & data distribution which resulted by drop some column which having lower correlation with TARGET
- Preparing data by perform categorization on AGE

EDA & Business Insights

- Check correlation matrix using heatmap
- Univariate Visualization
- Bivariate Visualization
- Multivariate Visualization
- Uncover Business Insights

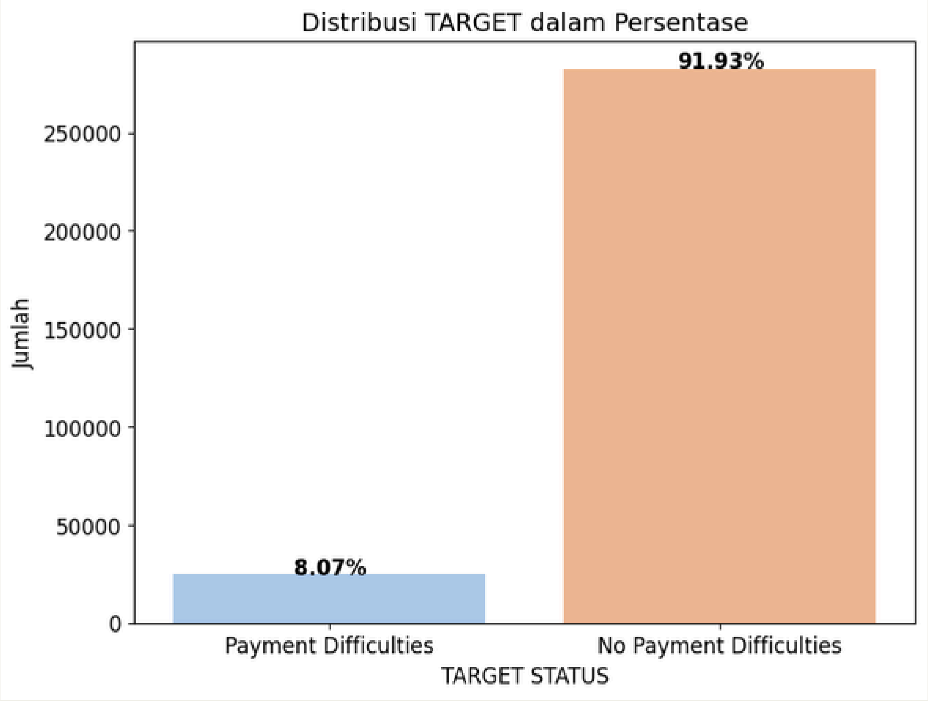
Model Building

- Data Preprocessing by separated categorical-numerical and X, y
- Label Encoding to transform non-num to num labels
- Feature Selection with 20 best features
- Handling Target Imbalanced Data by resampling data
- Scaling Data using StandardScaler
- Machine Learning Model with its hyperparameter
- Model Evaluation & Selection

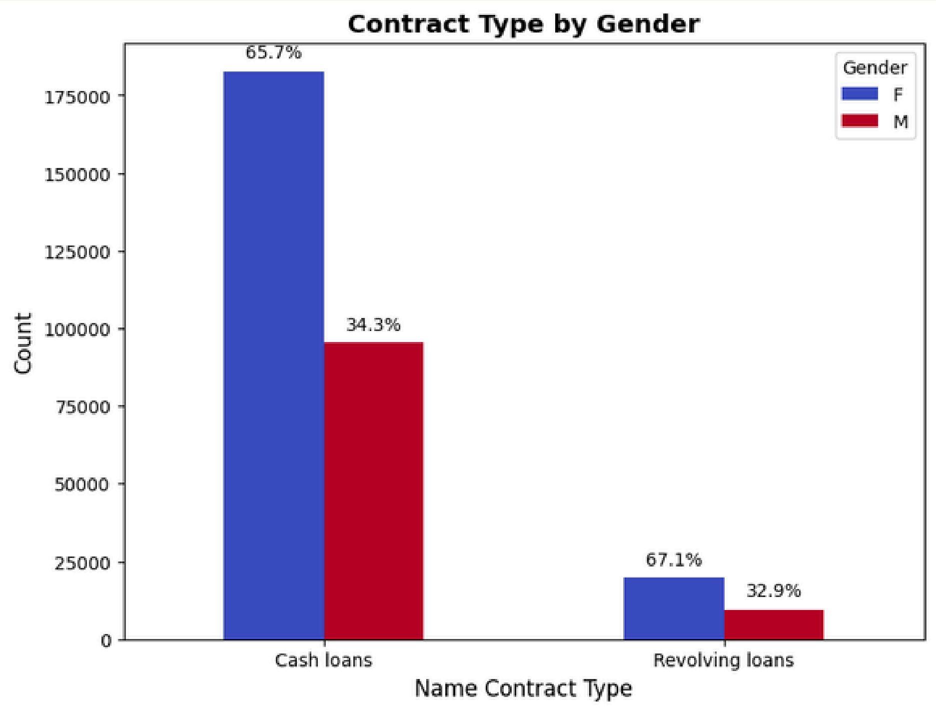
Dataset Test Prediction

- Handling Missing Values
- Check data duplication
- Preparing data
- Data Preprocessing
- Label Encoding
- Predict clients repayment abilities using best ML model
- Business Recommendation

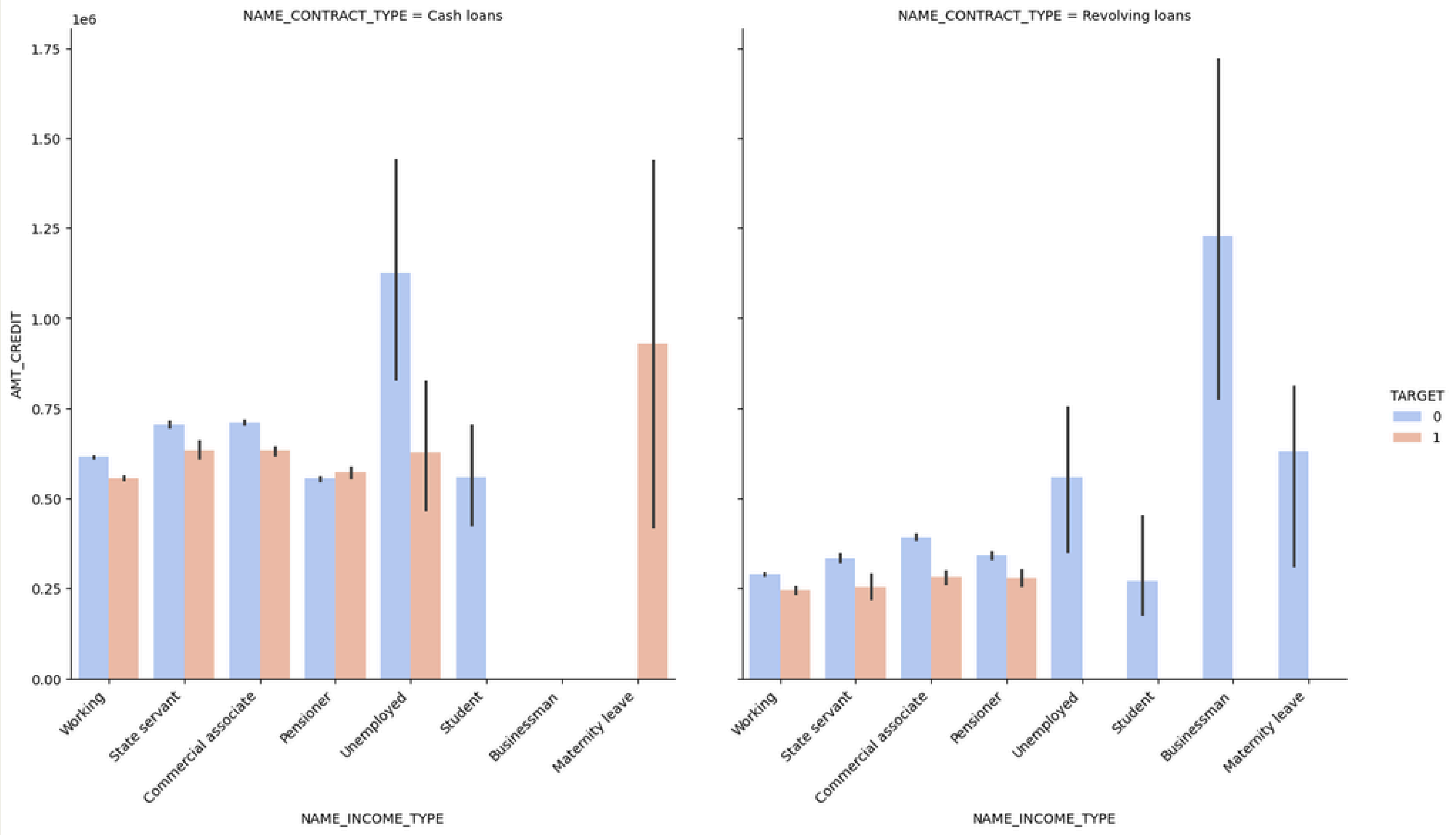
EDA & BUSINESS INSIGHTS (1)



1. Around **8.1%** of total borrowers, or approximately 25K clients, **face challenges in repaying their loans.**



2. The two types of loan contracts offered to clients are cash loans and revolving loans, with **both being predominantly held by female clients** — 65.7% for cash loans and 67.1% for revolving loans.

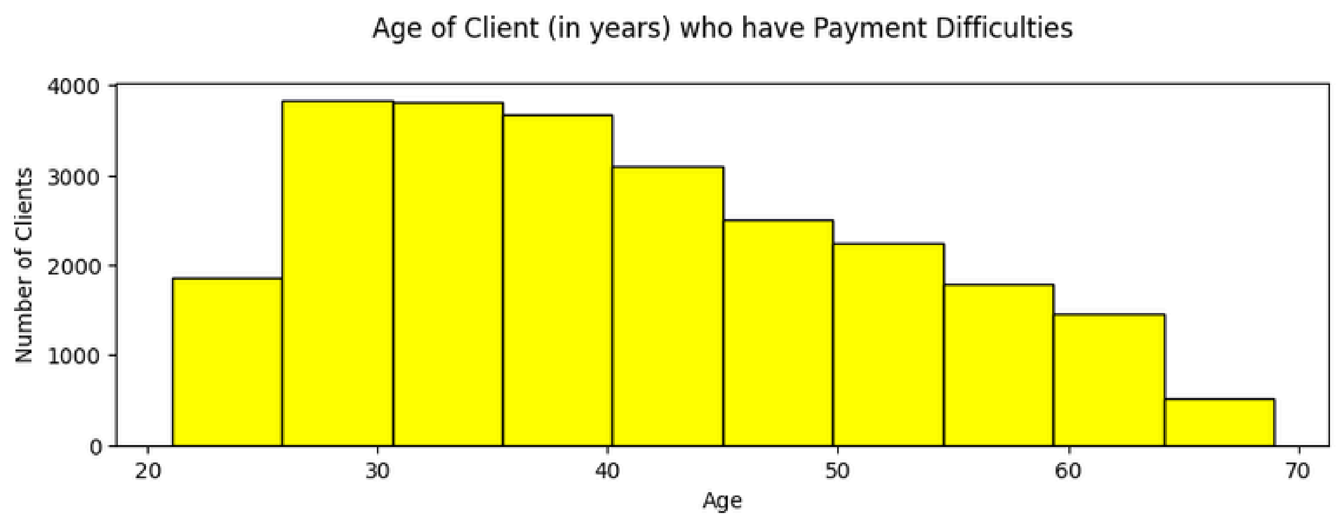


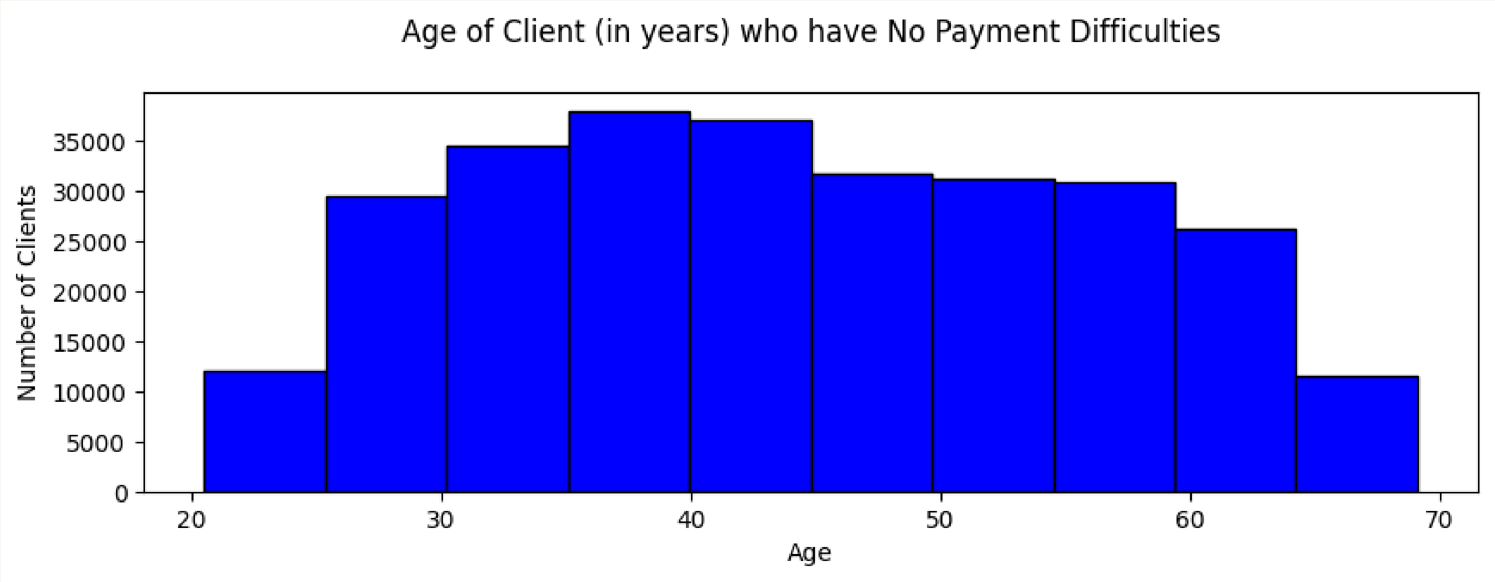
3.

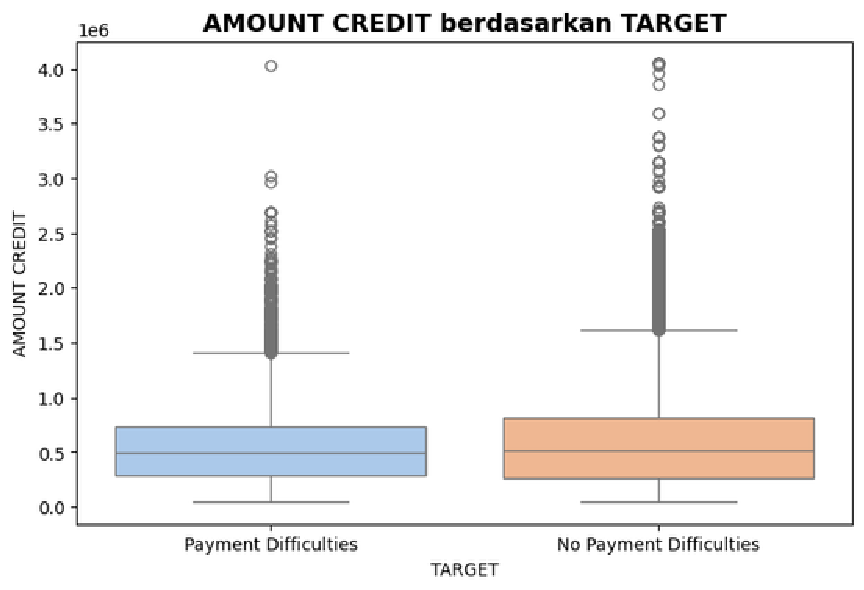
- Among **clients on maternity leave with cash loans, all face repayment difficulties** for medium credit amounts, whereas those with **revolving loans have no such issues.**
- More than **half of unemployed clients** with **cash loans struggle** with **repayment on medium credit amounts**, while **all unemployed clients** with **revolving loans** repay **without problems**.
- All **student** clients **manage to repay** their loans, whether **cash or revolving**, for low to medium credit amounts.

EDA & BUSINESS INSIGHTS (2)

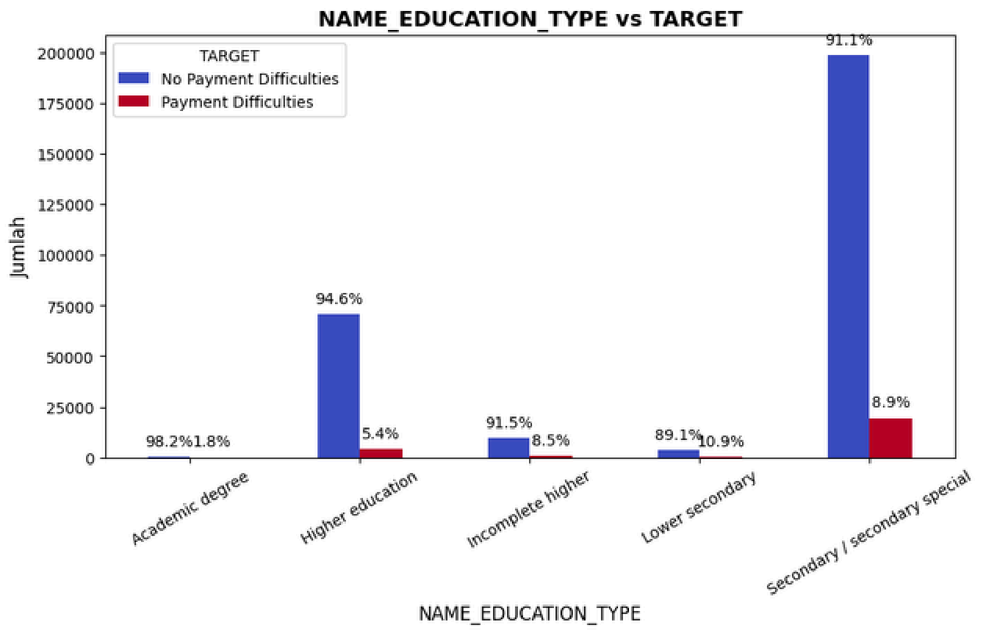
4. Most **loan applicants** are **between 35 and 40** years old, followed by those aged 40 to 45. Conversely, very few applications come from clients younger than 25 or older than 65. Clients aged **35-45 generally do not struggle with repayments**, whereas those aged **25-35 are more likely to face difficulties.**







5. The **median loan amount** for clients **without repayment issues** is **slightly higher** than that of clients with difficulties, suggesting that those **with larger loan amounts** have a **slightly better ability** to repay compared to those with smaller loans.



6. Clients with **lower secondary education** have the **highest repayment difficulties** (11%), despite **most applicants** having a **secondary education** (218K).

MACHINE LEARNING IMPLEMENTATION

Using the **top 20 column** selected based on the **score and p-value**, for the training process.
5 most important features: EXT_SOURCE_2, EXT_SOURCE_3, DAYS_BIRTH, DAYS_ID_PUBLISH, & DAYS_REGISTRATION

ML Model	Accuracy Score		Error Margin	ROC-AUC Score
	Training	Testing		
Logistic Regression	**73.19%**	**73.23%**	**0.04%**	**67.28%**
Random Forest	**97.89%**	**95.55%**	**2.34%**	**95.55%**
Gradient Boosting	**71.46%**	**70.1%**	**1.36%**	**70.1%**
K-Nearest Neighbor	**91.62%**	**88.13%**	**3.49%**	**88.12%**

The **Random Forest** model was identified as the best-performing model. The 2.34% error margin indicates a well-balanced performance, with **no signs of underfitting or overfitting**. Moving forward, this model will be used to predict the clients repayment abilities using application_test dataset.

BUSINESS RECOMMENDATION

1. Since younger clients (25-35 years old) and those with lower secondary education have higher loan repayment difficulties, Home Credit should implement **stricter risk assessment measures** for these groups. This could include offering **smaller initial loan amounts**, requiring **financial literacy training**, or providing **flexible repayment plans** to improve repayment success rates.

2. Given that women hold the majority of loans and have a lower default rate than men, Home Credit could consider **expanding loan programs** specifically tailored for **female borrowers**. Offering incentives like **lower interest rates** or **financial education programs** could further enhance repayment rates and financial inclusion for women.

3. Since repayment success improves with higher income, Home Credit should consider **income-sensitive loan products**. For example, **unemployed** clients and those on **maternity leave** should have **customized repayment plans**, such as **deferred payments** or **income-adjusted installments**, to reduce the likelihood of default while still allowing access to financial services.

Thank you!

LinkedIn
Johanes Purek

Email Address
johanesbernard26@gmail.com

GitHub
github.com/johanbernardd

Final Task Notebook:
https://bit.ly/finaltasknotebook-johanes