Risk Clients Scorecard Model

Ismi Ana Sulasiyah
December 1st,2025

Home Credit Indonesia PBI FINAL TASK:

Project Journey (PACE Framework)

PLAN

Including study case Background, Goals, and Objective for this project

ANALYZE

Build a dataframe and organize the data for the process of exploratory data analysis

CONSTRUCT

Building Predictive and Machine Learning Model with their model performance

EXECUTE

Choose the best model and conduct recommendation

PACE Workflow: Plan Stage

Project Background

Home Credit saat ini sedang menggunakan berbagai macam metode statistik dan Machine Learning untuk membuat prediksi skor kredit. Sekarang, kami meminta anda untuk membuka potensi maksimal dari data kami. Dengan melakukannya, kita dapat memastikan pelanggan yang mampu melakukan pelunasan tidak ditolak ketika melakukan pengajuan pinjaman, dan pinjaman datap diberikan dengan principal, maturity, dan repayment calendar yang akan memotivasi pelanggan untuk sukses.

Goals

The goal of this project is to create a Scorecard Model that can predict which clients are likely to be late in paying far beyond their due date (payment difficulties).

Objectives

- Conducting EDA as Data Preprocessing process,
- Build Predictive Models and Machine Learning
- Evaluate and Compare which models that perform better
- Implement best model to predict test data
- Give recommendation to reduce potential clients with risky behavior.

PACE Workflow: Plan Stage

1

PACE: Plan

Step 1. Initial EDA of the Dataset

The main dataset in this training stage is application_train.csv. This dataset has 307511 rows and 122 columns.

[3]:

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT
307506	456251	0	Cash loans	M	N	N	0	157500.0	254700.0	
307507	456252	0	Cash loans	F	N	Y	0	72000.0	269550.0	
307508	456253	0	Cash loans	F	N	Y	0	153000.0	677664.0	
307509	456254	1	Cash loans	F	N	Y	0	171000.0	370107.0	
307510	456255	0	Cash loans	F	N	N	0	157500.0	675000.0	

This train dataset is a mix of categorical and numerical value. The TARGET column which is as a differentiate between train and test dataset can be our main parameter for this study case. 'TARGET' variable contains:
1 - client with payment difficulties (he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample)
0 - all other cases

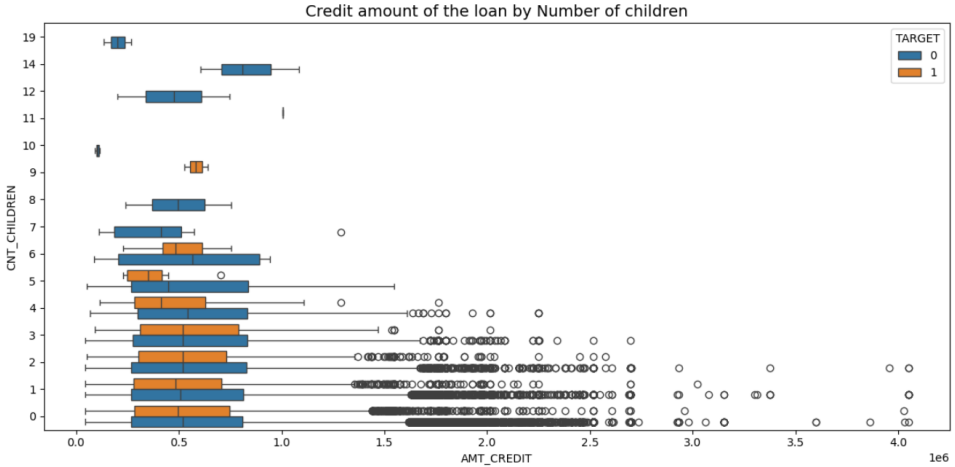
PACE Workflow: Analyze Stage

2

PACE: Analyze

Step 2. EDA Stage and Initial Analyze

This step will be included: Clean the dataset (missing data, redundant data, check outliers) and Initial Analyze with visualization. There are some insights that can be draw from this initial analyze stage.



Insight:

- Clients with 'TARGET' status fall into two general categories: Clients with more children (>6) tend to have less difficulty paying and clients with fewer children (especially those aged 0 - 2) often apply for large loans and are also late in paying.
- The histogram shows that there are relatively few clients with 'Payment Difficulties' along with the increasing of their number of children. It's possible that they're the higher-paid employees or always aware to apply for loans.

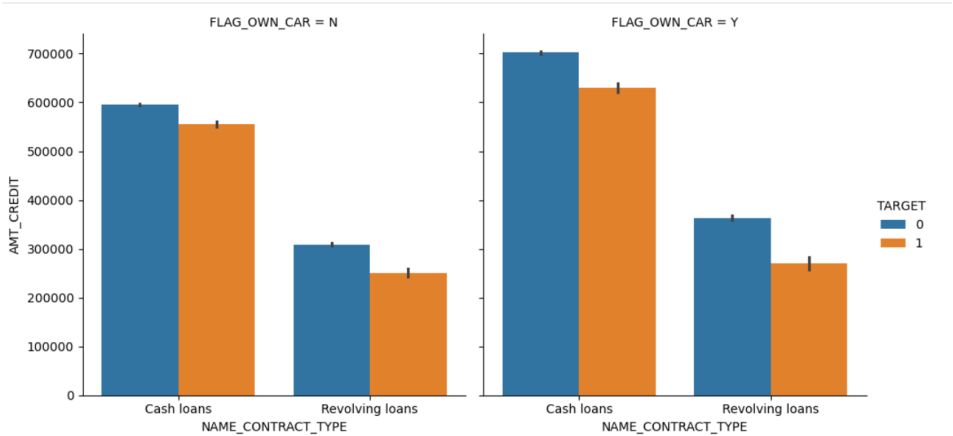
PACE Workflow: Analyze Stage

2

PACE: Analyze

Step 2. EDA Stage and Initial Analyze

Relationship of variable AMT_CREDIT (Credit amount of the loan), Contract Type, & Car Ownership (Split by TARGET).



Insight:

- Clients with higher loan amounts tend to pay better across both car ownership groups (FLAG_OWN_CAR = Y & N). Suggests larger loans are issued to more creditworthy applicants (Stronger financial profile → lower risk of delinquency)
- Car ownership correlates with higher loan amounts but also a higher risk of late payments. Car ownership should be considered as a factor in determining clients' creditworthiness.
- Borrowers with greater financial capacity, evidenced by car ownership and higher credit limits, show lower payment difficulties. Cash loans carry higher exposure but appear to be given more carefully to stable clients.

PACE Workflow: Construct Stage

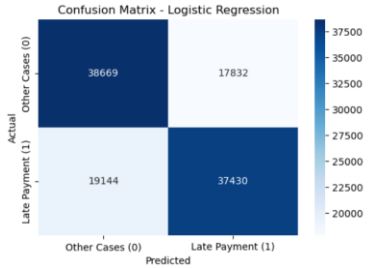
3

PACE: Construct

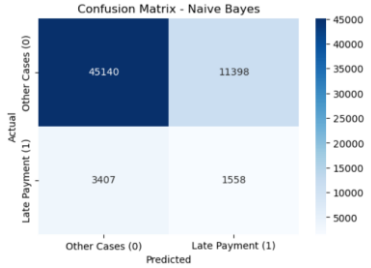
Step 3. Model Building and Performance Evaluation

Since the variables I wanted to predict (Clients with 'Late Payments' and 'Other Cases') were categorical, I tested several Classification algorithms (supervised learning). The Machine Learning algorithms I build are: Logistic Regression, Naive Bayes, Random Forest, and XGBoost, and then determined which model performed best.

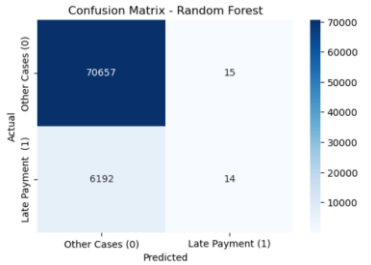
1. Logistic Regression (With Resampling Method)



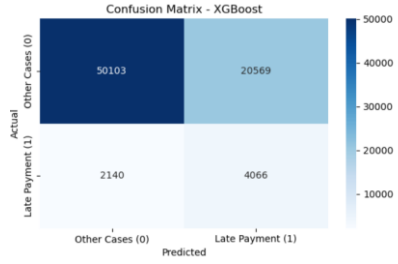
2. Naïve Bayes (With Oversampling Method)



3. Random Forest



4. XGBoost



Build a Multiple Linear Regression Model

4

PACE: Execute

Step 4. Evaluation and Determine which Best Model

	Models	Training Accuracy Score	Testing Accuracy Score	Error	ROC AUC
2	Random Forest	100.00	91.93	-8.07	0.7297
3	XGBoost	71.57	70.46	-1.11	0.7493
0	Logistic Regression	67.16	67.30	0.14	0.7324
1	Gaussian Naive Bayes	60.00	60.28	0.28	0.6514

Evaluation Summary

- Random Forest: Very high overfitting (Train score 100% & 91.9% Testing score), and error −8.07
- XGBoost: Has stable accuracy and best ROC AUC score
- Logistic Regression: Balanced but lower performance
- Naive Bayes: Weakest overall

Based on the machine learning model performance comparison table, **XGBoost** was found to be the best model with high score in AUC score 0.74 and least error with -1.11. Although Random Forest get the highest accuracy, this model has very high overfitting (Train score 100% & 91.9% Testing score), with error −8.07 highest rather than XGBoost has stable accuracy and Best ROC AUC. However, both models perform well and have been shown to work well with imbalanced data without using resampling.

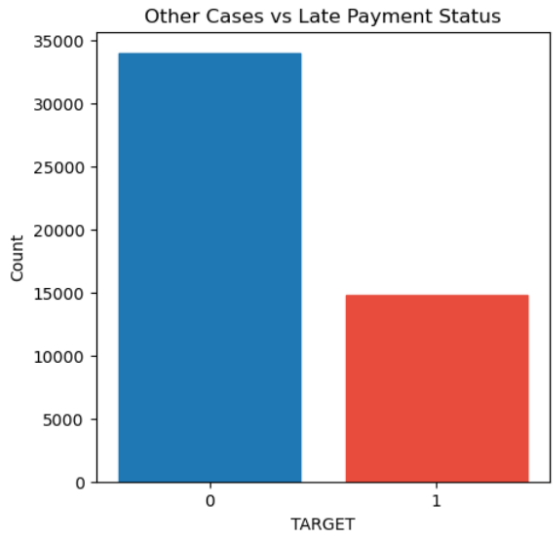
Build a Multiple Linear Regression Model

4

PACE: Execute

Step 4. Test best model (XGBoost) with 'Test' dataset & Recommendation

```
Summary of Payment Difficulties Status
Total Records: 48,744
Late Payment (1): 14,791 (30.34%)
Other Cases (0): 33,953 (69.66%)
```



Recommendation

A 30% late-payment rate is high — nearly one-third of customers struggle to pay on time. The recommendation to Reduce Late Payment Customers (TARGET = 1) as follows:

1. Improve Credit Scoring and Approval Policies. Use ML models (XGBoost, Random Forest) to act as a pre-screening tool.
   - ❑ Predict high-risk clients before approving credit
   - ❑ Set lower credit limits for risky applicants
   - ❑ Require additional verification (e.g. proof of income)
   - ❑ Higher interest for customers flagged as late-payment risk
2. More Flexible Repayment Schedules. People with irregular income often pay late, then I suggests:
   - ✓ Offer smaller, weekly payment options
   - ✓ Give customers payment date choices aligned to salary cycles
3. Early-Warning & Intervention Strategy.
   - ✓ Early SMS/email/push reminders before due date
   - ✓ Suggested minimum payment reminders
   - ✓ Chatbot financial help line

Thank You!

Project

Home Credit : **Predict** Potential Risk Clients
GitHub Project

Date

December 1st, 2025

Rakamin Academy

HOME CREDIT

Project Based Internship Batch November 2025

Author

Ismi Ana Sulasiyah
annaismi17@gmail.com
LinkedIn Ismi