Home Credit Scorecard

Project Based-Internship at Home Credit Indonesia
Data Scientist

Fadhila Salsabila

Short Brief

Individuals who dont have a credit history often face challenges in obtaining loans, making them suceptible to predatory lenders who exploit their financial vulnerability. Home Credit Indonesia, driven by its commitment to empower communities and provide inclusive and safe lending opportunities. However achieving this objective must be balanced with maintaining a viable business model.

Now, how we can approach individuals lacking credit history? What attributes should we take into account?

Problem Solution

- Predict Home Credit's clients scorecard using statistics and machine learning method.
- Analyze and define the best machine learning model to predict Home Credit scorecard.

Dataset Overview

Dataset

- Our current dataset of this project named by “application.csv” that consist from other mix dataset.
- The dataset has 307,511 rows with unique values by id loan those in “SK_ID_CURR”.
- The “Target” column of this dataset show the label of this problem, 0 = Clients without Difficulties Payment and 1 = Clients with Difficulties Payment

Features

- The dataset has 122 features, consist of categorical values and numerical values.
- The features in the dataset are related to clients demographics such as income, type of housing, type of employee, number of children, status, gender, etc.

Project Workflow (1)

1

Data Preparation

- Import library and load dataset.
- Check data type and data shape.

2

Data Validation

- Correct the incorrect values.
- Drop column if has missing values >50%
- Impute missing values in categorical columns using mode.
- Perform categorization on “AGE” column.

3

Exploratory Data Analysis

- Visualize some features using bar chart and pie chart
- Uncover business insights

Project Workflow (2)

4

Data Preprocessing

- Highlight features that are numerical and categorical.
- Separated input (X) and target (y).
- Performs feature selection.

5

Create ML Model

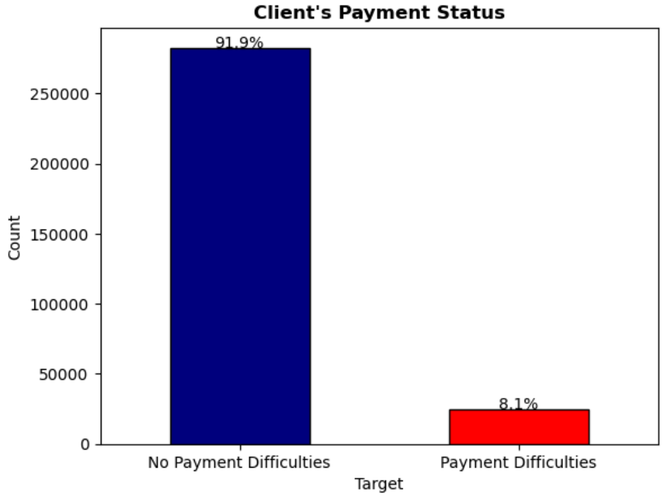
- Handling imbalanced data using oversampling SMOTE.
- Split dataset into training and testing.
- Build Machine Learning model using Logistic Regression, Naive Bayes, KNN, and Neural Network.

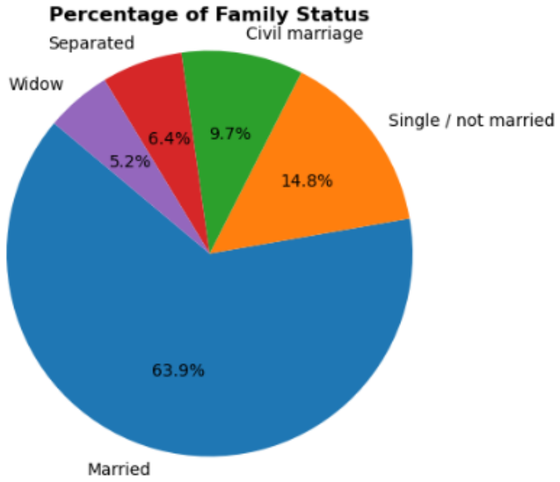
6

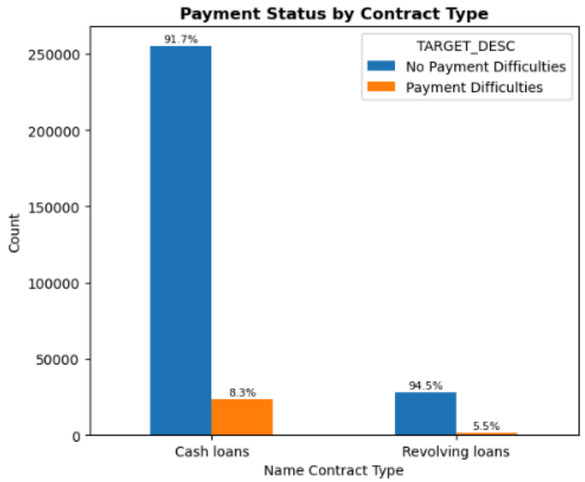
Model Evaluation

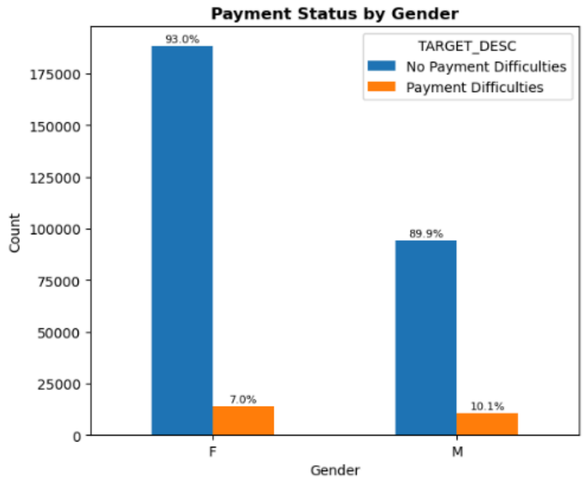
- Evaluation model using accuracy and AUC score.
- Select the best model.

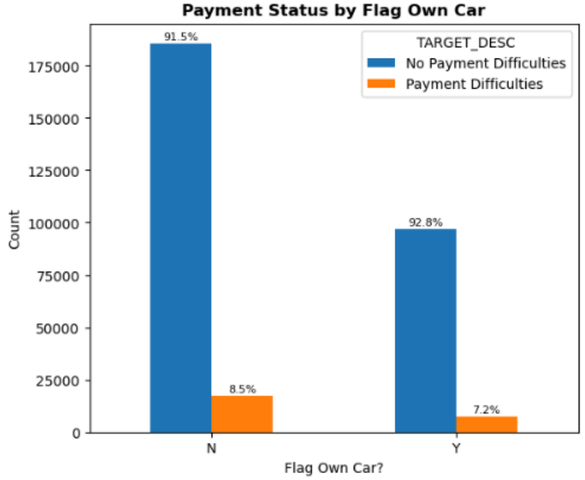
EDA (Exploratory Data Analysis)

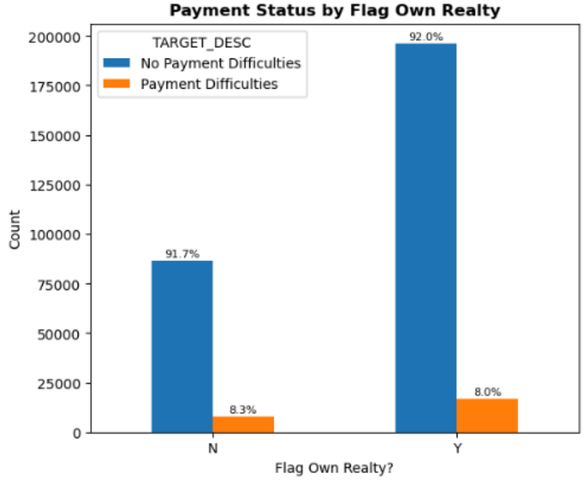












Business Insights (1)

- The percentage of clients who have **payment difficulties of repaying loans is 8.1%** or around **25K (24,908)**
- The **types of contracts** offered to clients are divided into 2 types, such as **cash loans** and **revolving loans** where both types are dominated by female clients, including 65.7% for cash loans and 67.1% for revolving loans
- **Female clients are considered more disciplined** in paying off loans than male clients with a smaller percentage of problems. Clients with payment difficulties for **women is around 7% while for men it is 10%**.
- There are around **9% of clients** who have difficulty **paying cash loans** and **6% of clients** who have difficulty **paying revolving loans**.

Business Insights (2)

- **The majority of clients** who have **payment difficulties dont have homes or private real estate.**
- Based on the **type of suite**, the client category that has **the most payment difficulties** comes from the **'Unaccompanied'** category, followed by **'Family'**
- Based on the **type of highest education** taken by clients, **the majority** are **'secondary/secondary special'** around 8.9%.
- Based on the **employee type**, clients with payment difficulties, **the majority** come from the **'working'** group, followed by the **'commercial associate'**, and **‘pensioner’.**
- Clients with **'housing type' = House/Apartment** have the **highest percentage** that have **payment difficulties** compared to other categories, namely 10% or around (31K) clients

Final Result

Model	Accuracy	AUC Score
Logistic Regression	58%	0,583
Naive Bayes	53%	0,531
K-Nearest Neighbors	**84%**	**0,844**
Neural Network	78%	0,778

Best Model : K-Nearest Neighbors

Business Recommendation

- Evaluate a policy and implement sanction for the clients who can't paying loans.
- Increase the system control in the company, include monitor loans payment, maximum loans limit, etc.
- Educate clients about financial management trough training, webinar, etc.

Thank You!

GitHub https://github.com/dilasal19

Medium https://fadhilaasalsa.medium.com/