CREDIT RISK SCORECARD MODELING

Author: Bintang Phylosophie | Data Scientist Intern

Project Overview

PT Home Credit Indonesia, commonly known as Home Credit, is a multinational company specializing in multipurpose financing. The company offers in-store financing services—non-cash financing provided directly at the point of sale—for consumers purchasing items such as household appliances, electronic devices, mobile phones, and furniture. In addition, Home Credit has developed technology-based financing solutions. Established in Jakarta in 2013, the company now operates at over 19,000 distribution points across 144 cities in Indonesia. As of March 2019, Home Credit has served 3.4 million customers both online and offline.

Data and Business Understanding

Dataset Information:
There are 7 datasets given for this project:

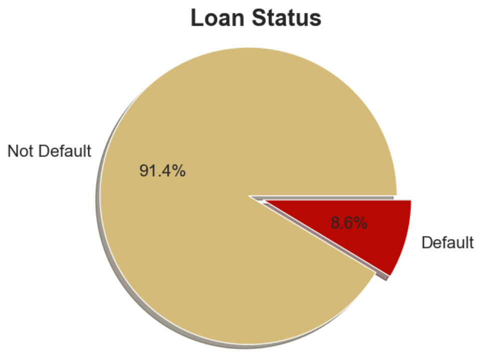
- **application_train.csv** (the main table)
- **bureau_balance.csv** (monthly balances of previous credits)
- **bureau.csv** (client's previous credits)
- **credit_card_balance.csv** (monthly balance snapshots)
- **installments_payment.csv** (repayment history for the previously disbursed credits)
- **pos_cash_balance.csv** (monthly balance snapshots of previous POS (point of sales) and cash loans)
- **previous_app.csv** (all previous applications)

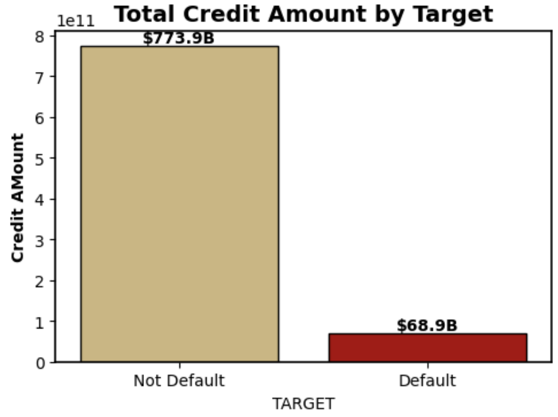
This dataset is a comprehensive collection of applicant information used for credit risk assessment in the context of loan applications. It contains **307.511 row**, each row represents an individual applicant. The dataset includes 122 columns, a mix of categorical (16 columns) and numerical (106 columns) data.. These features cover a wide range of data, including personal demographics, income and employment details, loan and credit information, housing characteristics, social and regional indicators, and external risk scores.

Attribute Information:

- **Identifier:**
  `SK_ID_CURR` is unique customer ID.
- **Clients Attributes:**
  These features cover a wide range of data, including personal demographics, income and employment details, loan and credit information, housing characteristics, social and regional indicators, and external risk scores.
- **Target:**
  `TARGET` variable indicates whether the applicant defaulted on their loan (1 for default, 0 for no default

What Happened?



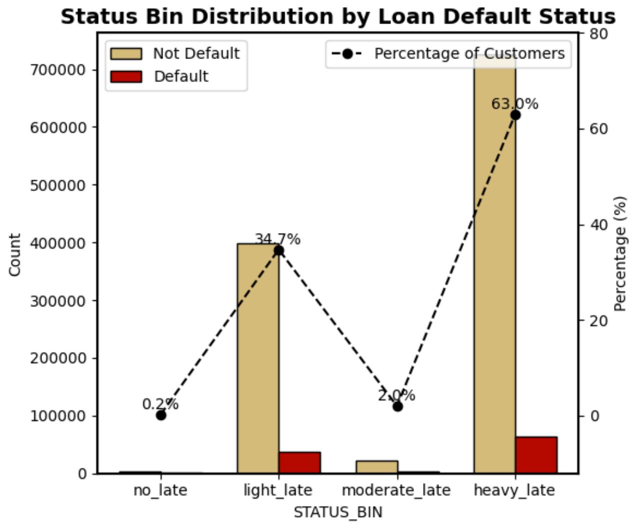


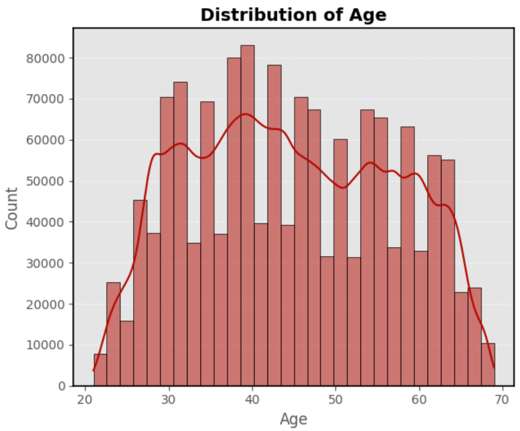
● **Default** total **24.825**
● **No Default** total **282.686**

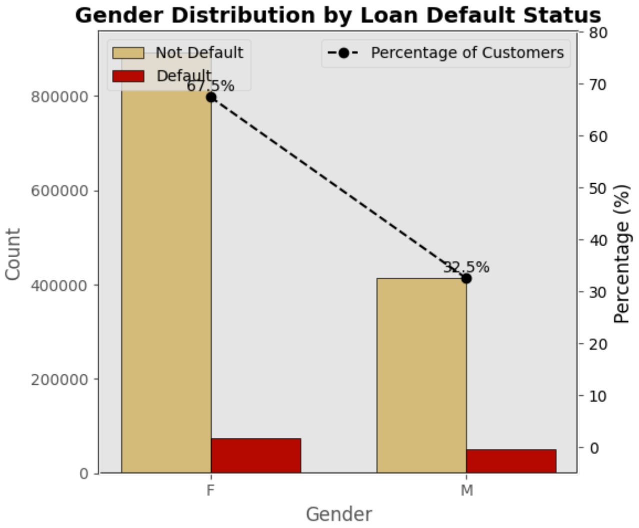
Problem Statement

- **Problems:**
  A review of customer credit data reveals that **8.6% of customers defaulted**, accounting for **$68.9 billion** in defaulted loan amounts. This represents a disproportionate financial impact from a small portion of the customer base, indicating limitations in current credit risk evaluation methods.
  1. Approving loans to applicants who are unlikely to repay their loans resulting in financial losses for the company.
  2. Disapproving loans to applicants who are likely to repay the loan resulting in business losses.
- **Goals:**
  The goal to be achieved in this case is to develop a machine learning model capable of accurately predicting whether a customer is likely to default or not, enabling proactive risk management and better-informed lending decisions to reduce default rate through data-driven client assessment using machine learning.
- **Objectives:**
  1. Analyze and preprocess historical credit and customer data to identify key features influencing default behavior.
  2. Build and evaluate multiple classification models to determine the best-performing algorithm.
  3. Achieve a balance between precision and recall to minimize false positives (unnecessarily rejecting good customers) and false negatives (missing high-risk defaulters).
- **Metrics:**
  - Default Rate Reduction (%)
  - Loss Avoided ($)
  - Model Accuracy / Recall on Bad Clients (%)
  - Cost Saved from Bad Loans ($)

Who is Applying for Loan?


Status Bin Distribution by Loan Default Status

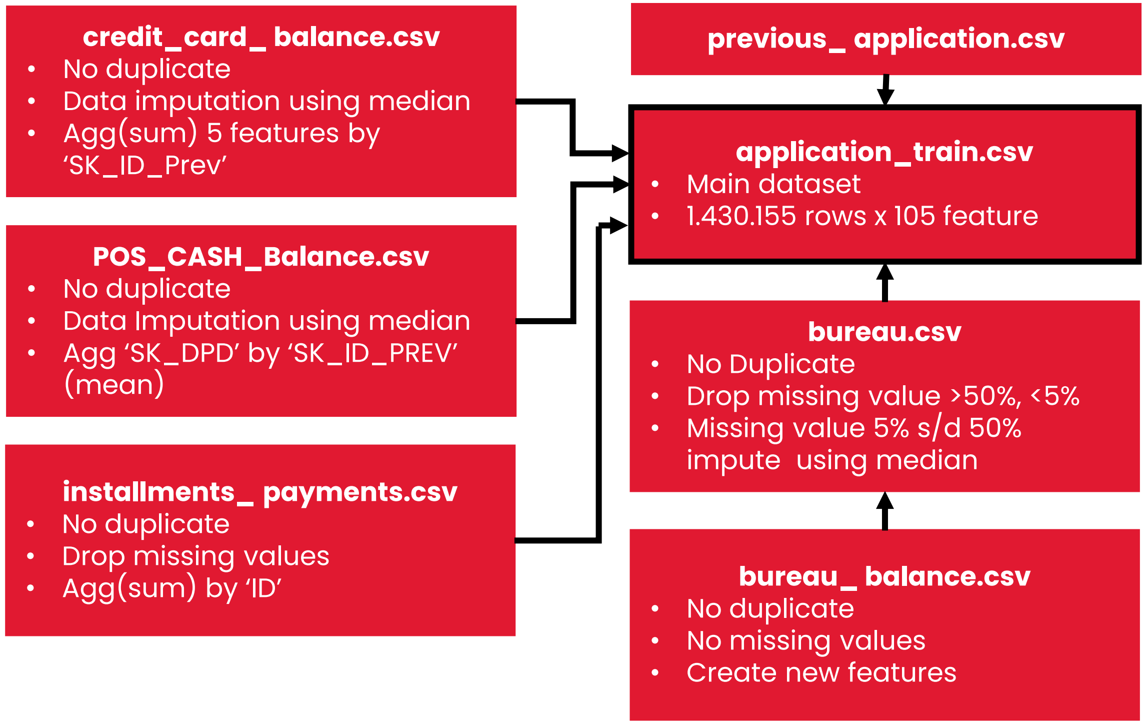

Distribution of Age


Gender Distribution by Loan Default Status

Insights:
- Heavy late payers dominate & show higher defaults
- Core customers are 30–55
- Females dominate the portfolio, with similar risk to males

Recommendations:
- Segment and apply differentiated credit policies based on payment behavior.
- Focus product design & marketing on this group; develop age-specific offers.
- Maintain neutral risk policy but tailor offers to the female market.

Data Preprocessing



credit_card_ balance.csv
- No duplicate
- Data imputation using median
- Agg(sum) 5 features by ‘SK_ID_Prev’

POS_CASH_Balance.csv
- No duplicate
- Data Imputation using median
- Agg ‘SK_DPD’ by ‘SK_ID_PREV’ (mean)

installments_ payments.csv
- No duplicate
- Drop missing values
- Agg(sum) by ‘ID’

previous_ application.csv

application_train.csv
- Main dataset
- 1.430.155 rows x 105 feature

bureau.csv
- No Duplicate
- Drop missing value >50%, <5%
- Missing value 5% s/d 50% impute using median

bureau_ balance.csv
- No duplicate
- No missing values
- Create new features

- The dataset contains **no duplicate** records.
- **Missing values** have been imputed using appropriate strategies—either the median or mode, depending on the nature of each feature.
- **Feature engineering** was applied to derive new variables from existing ones, and data types for datetime fields were properly adjusted.
- **Inconsistent values** were corrected.
- **Outliers** were capped at the **IQR** boundaries.
- For **feature selection**, high-missing-value features were removed, variables with strong Information Value (IV) were selected, and multicollinear features (correlation ≥ 0.7) were dropped.
- **Feature encoding** and scaling were performed using feature binning and Weight of Evidence (WoE) transformation.

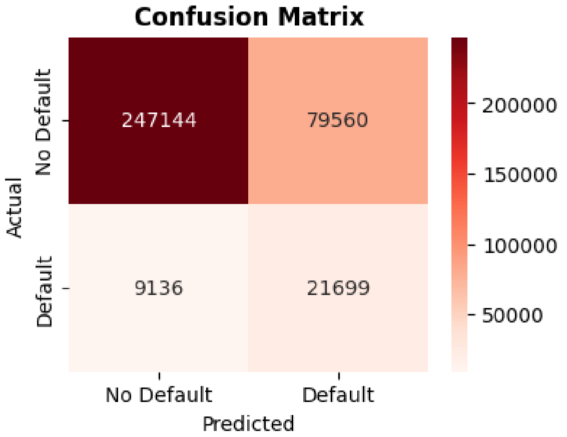
Model Development

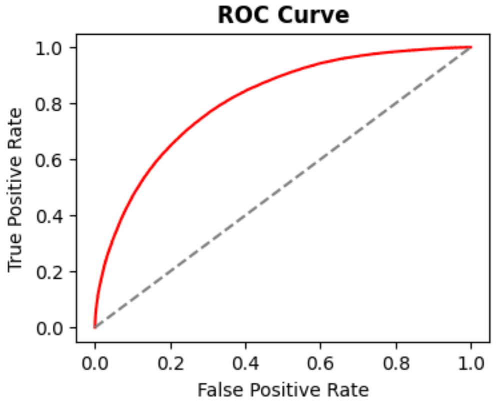
I used two models in this case: Logistic Regression and LightGBM for model development.

Model Evaluation

- To balance the trade-off between false positives and false negatives, **recall will be used as the primary evaluation metric.** This is because in our case, minimizing false negatives is more critical—missing a true positive could lead to missed opportunities. Recall is especially valuable in imbalanced classification tasks where the minority class is of greater importance.
- However, **accuracy** will also be considered, as it offers an easily interpretable measure of overall model correctness.
- In credit risk modelling, test performance is calculated using **the AUC metrics**.

Model	Accuracy	Recall	ROC AUC
Logistic Regression	0.68	0.67	0.74
Logistic Regression (Tuned)	0.68	0.67	0.74
LightGBM	0.70	0.68	0.76
LightGBM (Tuned)	**0.75**	**0.70**	**0.81**





- True Negatives **(TN)** - correctly predicted bad loans: 247144
- False Positives **(FP)** - predicted bad, actually good: 79650
- False Negatives **(FN)** - predicted good, actually bad: 9136
- True Positives **(TP)** - correctly predicted good loans: 21699

XGBoost gives the best performance.
Highest AUC **(81%)**, Strong generalization
High Recall **(70%)**, Effective high-risk detection
High Accuracy **(75%)**, Balanced and reliable
Stable in both train & test, Low overfitting risk

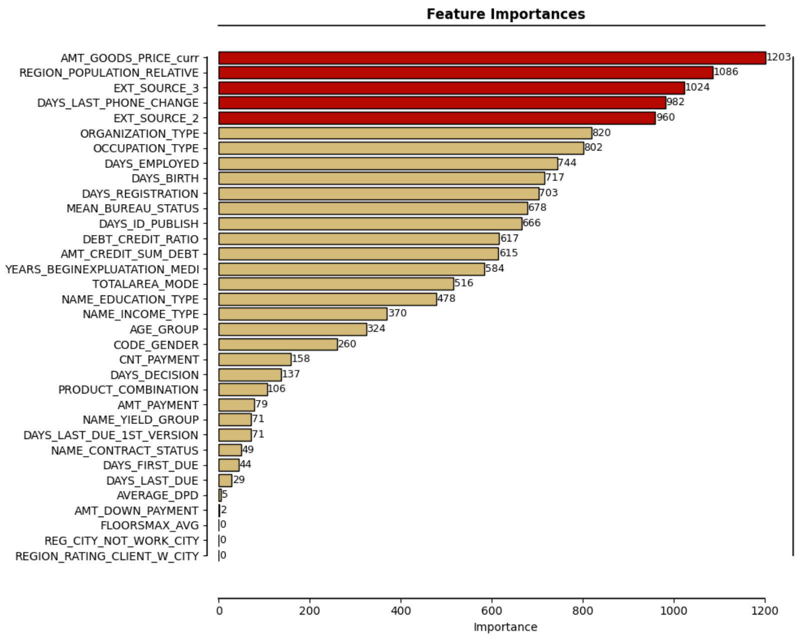
Features Importance & SHAP Value Distribution

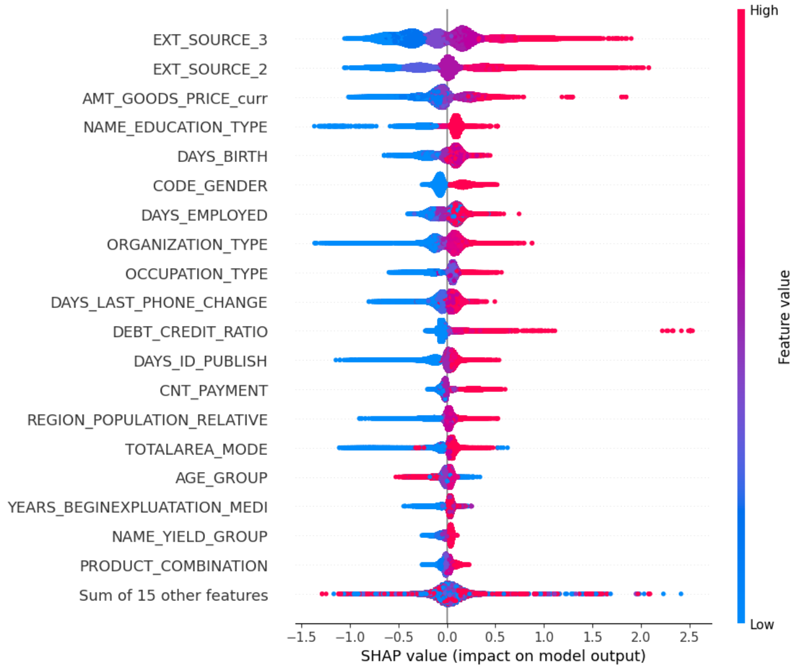
Most influential features:

- 'AMT_GOODS_PRICE_CURR' (The total price of the goods for which the loan is requested. Higher loan amounts may increase the financial burden on the applicant, potentially raising default risk).
- 'REGION_POPULATION_RELATIVE' (The relative population of the applicant's region. Applicants from less populated areas may be seen as riskier due to lower economic activity).
- 'EXT_SOURCE_3' (An external risk score, likely from a third-party credit bureau. Higher values typically indicate better creditworthiness).
- 'DAYS_LAST_PHONE_CHANGE' (Number of days since the applicant last changed their phone number. A recent change might signal instability or potential fraud).
- 'EXT_SOURCE_2' (Another external credit score used to assess risk. Like EXT_SOURCE_3, higher values are typically associated with lower credit risk).

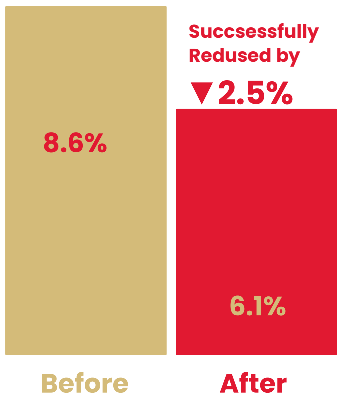
SHAP (SHapley Additive exPlanations) explains how each feature shifts the model's prediction:

- High feature values (shown in red) may reduce risk and push the prediction lower (e.g., high 'EXT_SOURCE_3' or 'EXT_SOURCE_2' values lower the risk).
- Low feature values (shown in blue) may increase risk, shifting predictions higher (e.g., low 'DAYS_LAST_PHONE_CHANGE' values—indicating a recent phone change—can raise risk).
- The direction and magnitude of the SHAP values show how much and in what way each feature affects individual predictions.

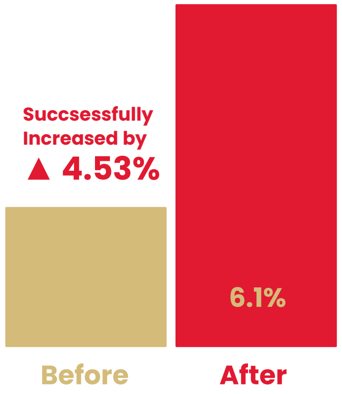




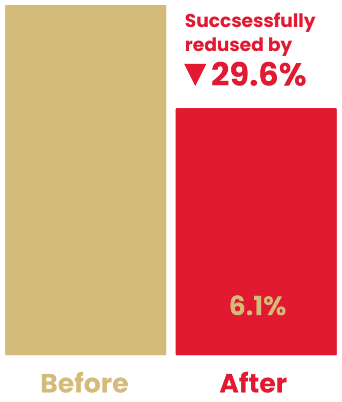
Default Rate



Net Revenue



Cost Save from Bad Loan



Business Recommendation

- **Enhance Risk Models with External Credit Scores**
  Insight:
  Features like 'EXT_SOURCE_3' and 'EXT_SOURCE_2' are the most powerful predictors of default.
  Recommendation:
  Ensure consistent access to high-quality third-party credit scoring data. Consider strengthening partnerships with bureaus that provide these scores and using them as a core component of risk-based pricing or automated approval systems.

- **Incorporate Loan Amount Tiers into Approval Rules**
  Insight:
  Higher 'AMT_GOODS_PRICE_CURR' is strongly associated with higher risk.
  Recommendation:
  Implement tighter approval thresholds or require stronger supporting documentation for high-loan-amount applicants. Consider tiered interest rates or collateral requirements based on loan size.

- **Use Behavioral and Lifecycle Features for Early Risk Detection**
  Insight:
  Age ('DAYS_BIRTH'), employment history ('DAYS_EMPLOYED'), and education type ('NAME_EDUCATION_TYPE') also impact predictions.
  Recommendation:
  Incorporate lifecycle and behavioral segments into marketing, underwriting, and collection strategies. For example:
  - Younger applicants or those with unstable employment may benefit from smaller initial loans with quicker repayment schedules.
  - Education level can be used to refine customer segmentation and tailor financial product offerings.

Click Here to See My Code

Linked In | GitHub | Portfolio