Report on Final Project

submitted towards differentiation for the course of

Data Preparation & Visualization

by

Hung Nguyen Quang ,
Giang Duong Phuong,
Hanh Nguyen Thi Hong,
Phuong Anh Trinh

Group 6
School of Technology
National Economics University

Supervisor(s)
Giang Nguyen Thi Quynh

December 2024

Contents

1 | Introduction

This report presents an end-to-end pipeline to compete in Kaggle's Home Credit Risk competition. Through extensive exploratory data analysis which gives a thorough understanding of the dataset, we have developed an appropriate and time-efficient model to **rank first (1/20)** in the final leaderboard.

The dataset encompasses credit application records and related information. It comprises eight distinct files, with the *application_train* and *application_test* files serving as the primary datasets. The other files provide supplementary details:

- *previous_application, credit_card_balance, POS_CASH_balance,* and *installments_payments* describe individual behaviors in prior credit applications.
- *bureau* and *bureau_balance* represent external credit data associated with the applicants.

EDA focuses on uncovering patterns and insights within the data, identifying relationships across variables, and assessing the quality of the datasets. The analysis includes a detailed examination and visualization of key features in each dataset, evaluating distributions, correlations, and trends to support hypothesis generation and feature engineering. The sections below provide a detailed analysis of each dataset.

Through meticulous data curation, strategic feature engineering, and fine-tuned hyperparameter optimization, we developed a high-performance machine learning pipeline with exceptional computational efficiency pipeline that operates within 20GB of RAM, executes in just 5 minutes on our setup with i7-12700 Intel CPU, or under 30 minutes on a Kaggle notebook, while maintaining top-tier performance on the leaderboard.

2 | Exploratory data analysis

2.1 Application Train

Defaulters are typically low-skill laborers, males, those with lower education, cash loan applicants, and individuals in rented housing or with mismatched addresses. Stability in occupation, education, housing, or credit scores offers greater repayment reliability. For lenders, these insights provide a roadmap for smarter risk models and tailored policies, benefiting both borrowers and institutions.

2.1.1 Basic Statistic

The *application_train* file contains 246,009 records, offering a comprehensive client profile by integrating financial, personal, and behavioral attributes. It provides valuable insights for identifying patterns and key predictors of loan repayment difficulties, enabling accurate predictions of default probabilities.

The dataset has **8.08%** default rate, indicating a class imbalance. High-variability features like *AMT INCOME TOTAL* and *AMT CREDIT* reveal diverse financial profiles and repayment capacity, while stable features like *FLAG MOBIL* confirm baseline traits but have limited predictive value.

Property-related features such as *COMMONAREA AVG* and *APARTMENTS AVG* have high missingness (50%-70%), often irrelevant to renters or non-asset owners, highlighting client segmentation. Medium-missingness features like OCCUPATION TYPE (31%) and *EXT SOURCE 3* (19.8%) hold strong predictive potential despite data gaps, while low-missingness features such as *EXT SOURCE 2* and *AMT GOODS PRICE* provide reliable insights, making them essential for modeling.

2.1.2 Loan Applicants and Their Risky Profiles

The majority of loan applicants seek cash loans, making up over 90% of all applications. However, these loans come with a higher default rate (above 8%) compared to their safer counterpart, revolving loans (just over 5%). This disparity hints at the inherent risk of cash loans, often tied to larger amounts or more flexible terms.

Moving beyond loan types, we look at the individuals themselves. Women represent the majority of applicants, yet men default more often, making them riskier borrowers. This gender dynamic may hint at differences in financial habits or income stability. Additionally, the dataset contains three invalid *CODE GENDER* values labeled as XNA, all belonging to non-defaulters. These entries should be removed to maintain data quality.

Where applicants live speaks volumes about their financial stability. Those in rented apartments or co-op housing show alarmingly high default rates, a likely reflection of financial instability. Even those living with parents, often perceived as a safety net, aren't spared from heightened risk. In contrast, applicants owning their homes tend to be more secure, underscoring the protective effect of stable housing.

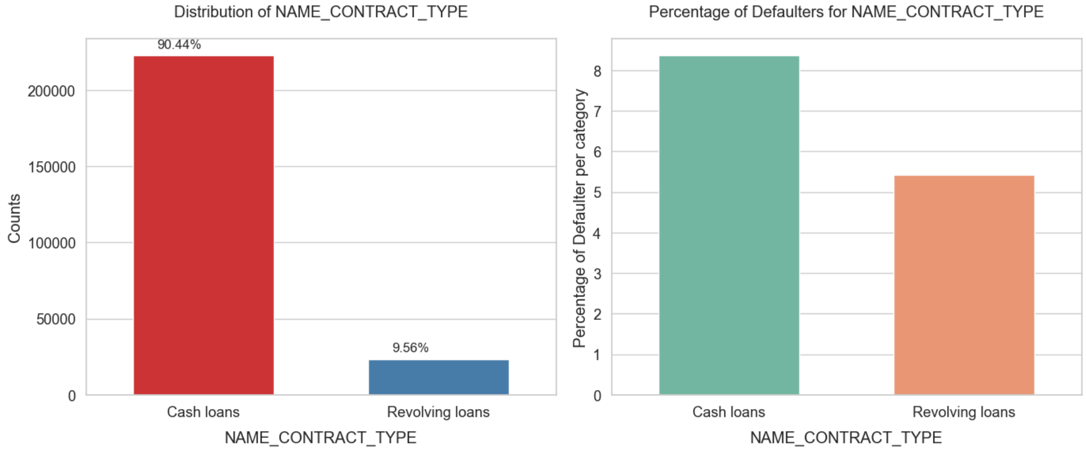

Figure 2.1: Distribution of NAME CONTRACT TYPE.

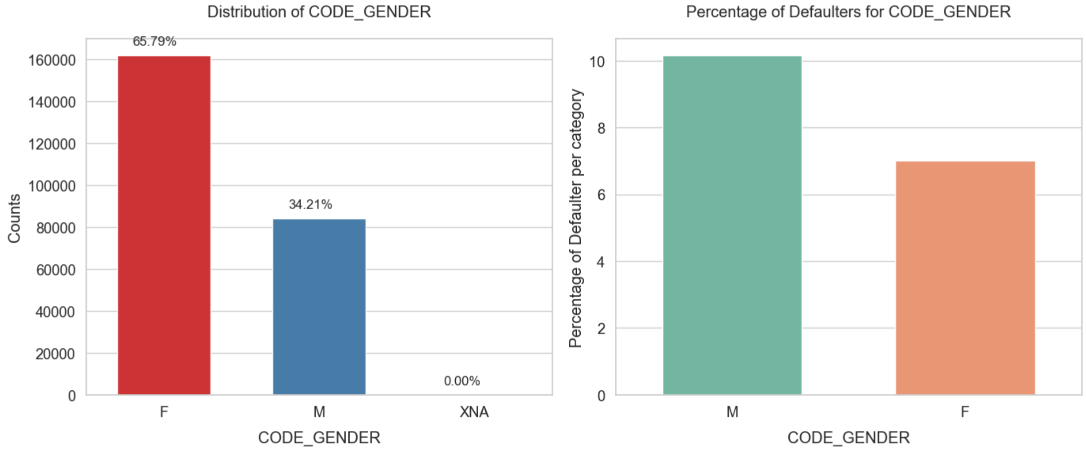

Figure 2.2: Distribution of CODE GENDER.

2.1.3 Occupation and Education

Delving deeper into the working lives of these applicants, we find that laborers are the largest group, accounting for 26.17% of applications. But size doesn't always mean safety—laborers, especially those in low-skill roles, face the highest risk of default. Other occupations, like sales staff and drivers, also carry notable risks, while those in high-skill technical roles, like IT professionals, are much safer bets.

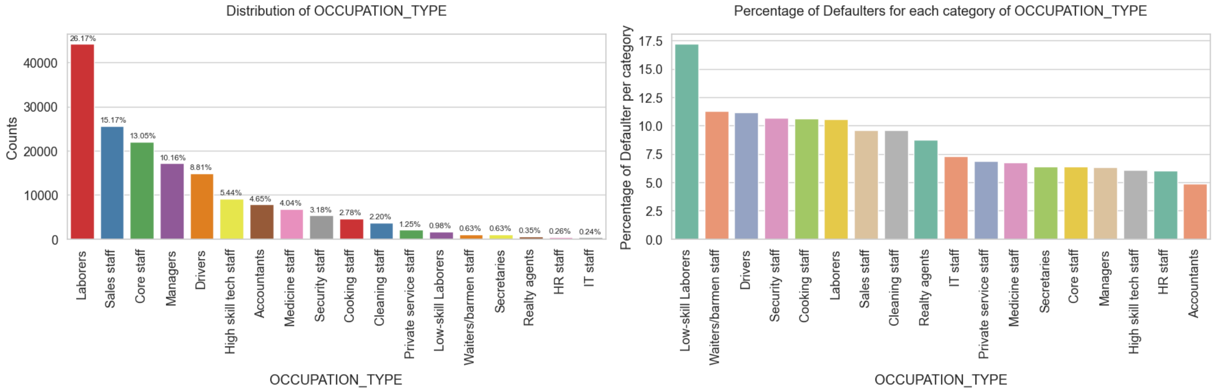

Figure 2.3: Distribution of OCCUPATION TYPE.

Education tells a similar story. The majority of applicants have secondary education, but those with lower secondary education are significantly more likely to default. On the flip side, individuals with higher education, particularly academic degrees, enjoy much lower default rates, showing how education often acts as a shield against financial vulnerability.

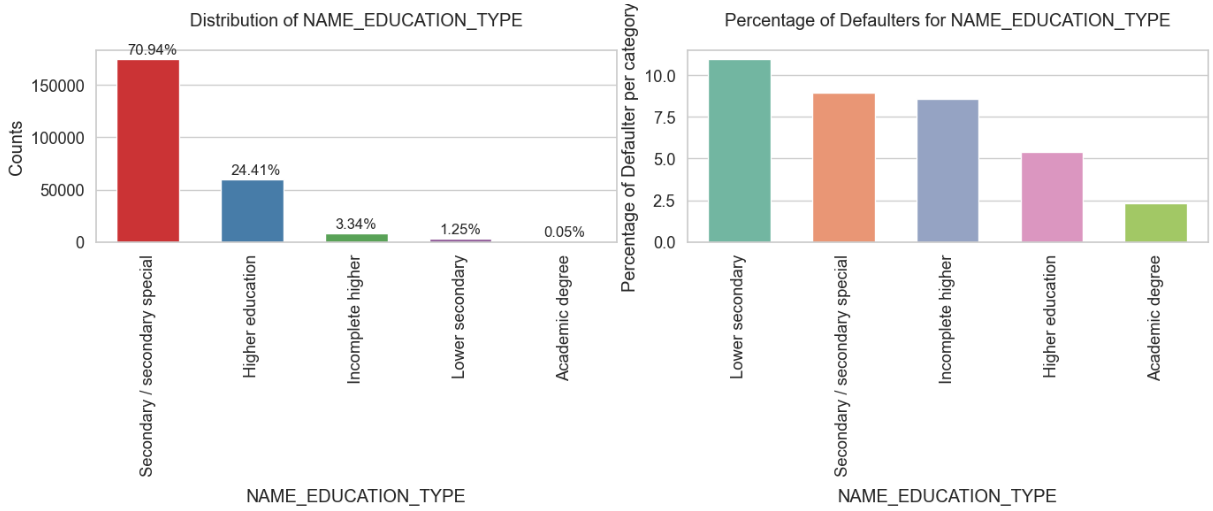

Figure 2.4: Distribution of NAME EDUCATION TYPE.

2.1.4 EXT SOURCE

EXT SOURCE scores stand out as the strongest indicators of creditworthiness. Defaulters consistently have lower scores, while non-defaulters shine with higher peaks, painting a clear picture of financial reliability.

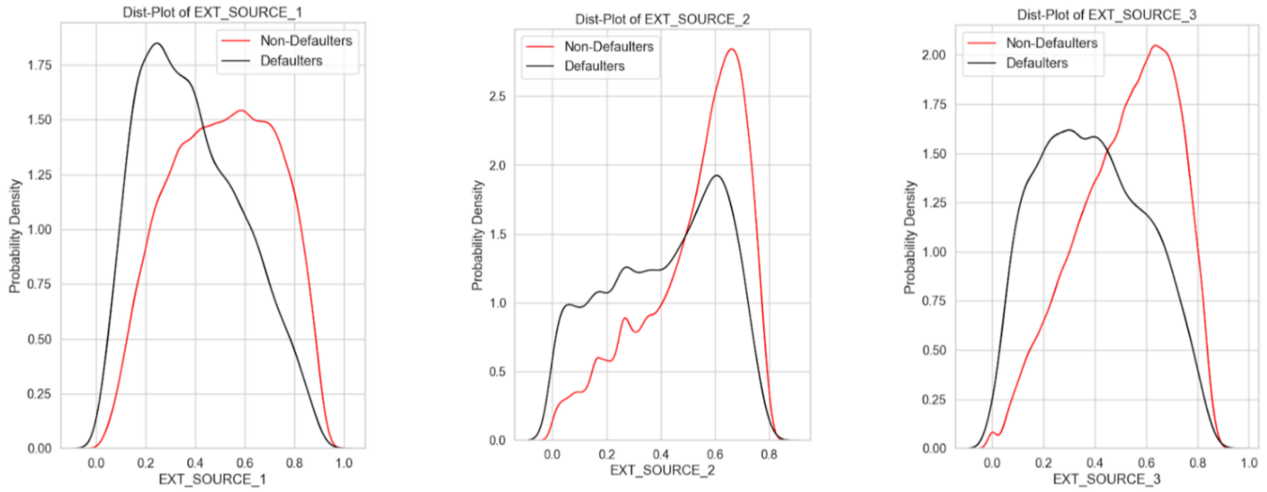

Figure 2.5: EXT SOURCE Images

2.2 Previous Application

Defaulters in the dataset typically include applicants for cash or revolving loans, those with prior rejections, incomplete records, or high-risk borrowing purposes. These traits reveal links between financial instability and repayment risk. For lenders, leveraging these insights to target risk-prone areas, enhance data quality, and tailor loan products can optimize risk management and support smarter lending strategies that benefit both lenders and borrowers.

2.2.1 Basic Statistic

The *previous application* dataset contains 1.41 million records, offering a detailed view of clients' past loan applications and capturing financial, behavioral, and loan-specific attributes. With 232,826 overlapping *SK ID CURR* values between *application train* and *previous application*, this dataset enriches the analysis by linking historical and current loan data.

High-variability features like *AMT APPLICATION* and *AMT CREDIT* highlight diverse borrowing needs, while features like *AMT DOWN PAYMENT* and *RATE DOWN PAYMENT* emphasize variability in loan structures, though anomalies like negative values require cleaning. Behavioral features such as *HOUR APPR PROCESS START* align with typical working hours and temporal features like *DAYS DECISION* span a wide range, requiring careful weighting in modeling.

Despite missingness in features like *AMT ANNUITY* and *AMT DOWN PAYMENT*, reliable attributes such as *CNT PAYMENT* and *AMT GOODS PRICE* provide strong predictive value, essential for modeling repayment behaviors and client segmentation. The Phi-K analysis reveals better the correlation of variables.

2.2.2 Loan Applications and Their Hidden Risks

While both cash loans and revolving loans are prevalent, revolving loans hold a slightly higher share and exhibit the highest default rates. The default rates between these two loan types are close, highlighting that neither offers a significant safety margin. Interestingly, records tagged with unclear contract types, such as 'XNA', exhibit the highest default rates, underscoring the potential risks tied to incomplete or ambiguous data. These findings highlight the importance of robust data validation and cautious risk modeling.

Diving into contract statuses, we see distinct patterns. Consumer loans boast the highest approval rates yet maintain a moderate default rate, reflecting their perceived stability. In stark contrast, canceled or refused applications show lower approval rates but alarmingly higher default rates when reconsidered, suggesting that applicants with prior rejections often carry greater financial instability.

2.2.3 The Role of Loan Purpose and Behavior

The reasons behind loan applications reveal much about risk. Borrowers refusing to name the goal or seeking loans for urgent needs, car repairs, or hobbies are more likely to default compared to those financing household goods or education. These high-risk purposes suggest that financial distress or impulsivity often drives these applications.

Behaviorally, applicants with a history of rejections or cancellations in previous loans tend to default more frequently. Repeat applicants, categorized under *NAME CLIENT TYPE*, are particularly notable—especially those labeled as 'XNA', who show the highest default rates. This indicates that repeaters in ambiguous or unstable categories carry heightened risks.

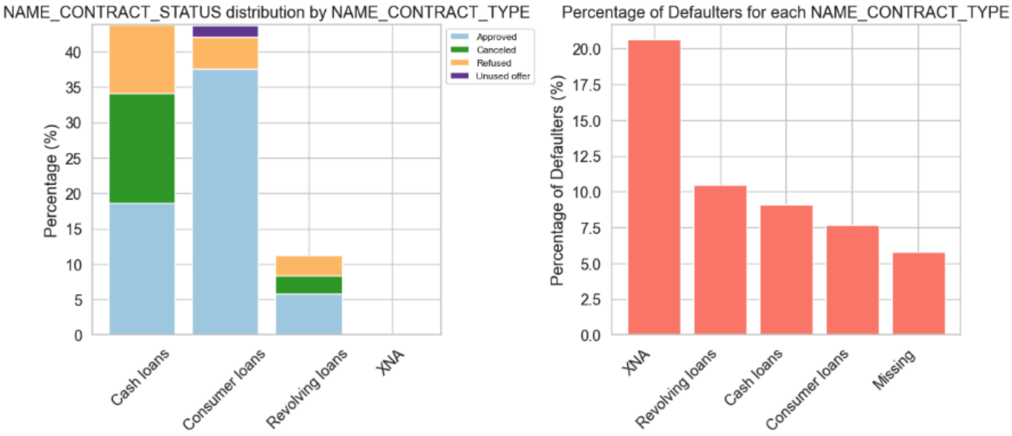

Figure 2.6: Distribution of NAME CONTRACT TYPE.

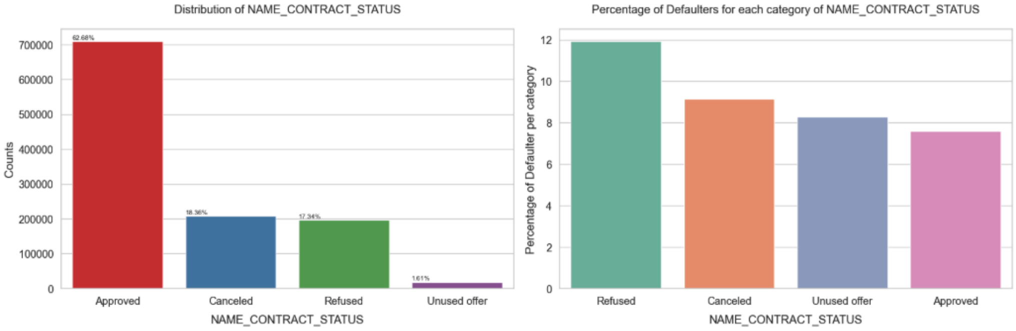

Figure 2.7: Distribution of NAME CONTRACT STATUS.

2.2.4 Additional Features

Analysis of defaulters' last five applications: From the analysis, it is evident that defaulters predominantly engage in Cash loans and Consumer loans in their last 5 transactions. These loan types are significantly more common than others, such as Revolving loans. The high frequency of these loan products suggests that defaulters often seek loans designed for immediate or short-term financial needs, possibly reflecting financial stress or a liquidity preference.

Interest Rates: The distribution of interest rates for defaulters shows most loans cluster near 0%, with a median around 0.32%. However, a significant tail of higher interest rates (above 5%, even exceeding 20%) reflects risk-based pricing for high-risk borrowers.

Down Payment Ratios: The analysis of down payment ratios reveals that most defaulters contributed minimal or no upfront payments for their loans. The majority of transactions have a down payment ratio close to zero, with very few exceeding a 10% threshold. This indicates a tendency among defaulters to minimize initial financial commitments, which could reflect constrained cash flow or riskier borrowing behavior.

2.3 Credit Card Balance

The *credit_card_balance* dataset contains monthly balance snapshots of previous credit cards held by applicants with Home Credit. With 87.4k rows and 86.9k unique SK ID CURR values, most applicants have only one credit card, while a few have more. This dataset will be explored to analyze credit card usage patterns and their relationship with loan behavior.

In the initial analysis of the dataset, we noticed some variables such as *AMT DRAWINGS ATM CURRENT, AMT DRAWINGS CURRENT, AMT RECEIVABLE PRINCIPAL, AMT RECEIVABLE,* and *AMT TOTAL RECEIVABLE* have negative minimum values, which may require further investigation to understand the cause. Regarding missing values, we identified 9 features with missing data. When exploring these, we found that when *AMT DRAWINGS ATM CURRENT* is null, *AMT DRAWINGS CURRENT* is always 0, indicating that the person did not withdraw money using an ATM. The same pattern applies to *AMT DRAWINGS OTHER CURRENT* and *AMT DRAWINGS POS CURRENT*.

We also created new features such as RATE OF PAYBACK (calculated as *AMT PAYMENT TOTAL CURRENT / AMT INST MIN REGULARITY*), which shows how much the monthly payment exceeds the minimum required. A deeper analysis using groupings by 6, 12, and 36 latest months revealed an interesting finding: As the loan repayment deadline approaches, *RATE OF PAYBACK* and the remaining balance for defaulters remain high, while for non-defaulters, these values decrease steadily and are nearly zero in the final 6 months. This supports the hypothesis that defaulters tend to have larger loan amounts and delay payments. As the repayment deadline approaches, non-defaulters have typically paid off their loans, while defaulters still owe a significant amount, leading them to pay more than the minimum installment. However, the debt is too large to be repaid on time.

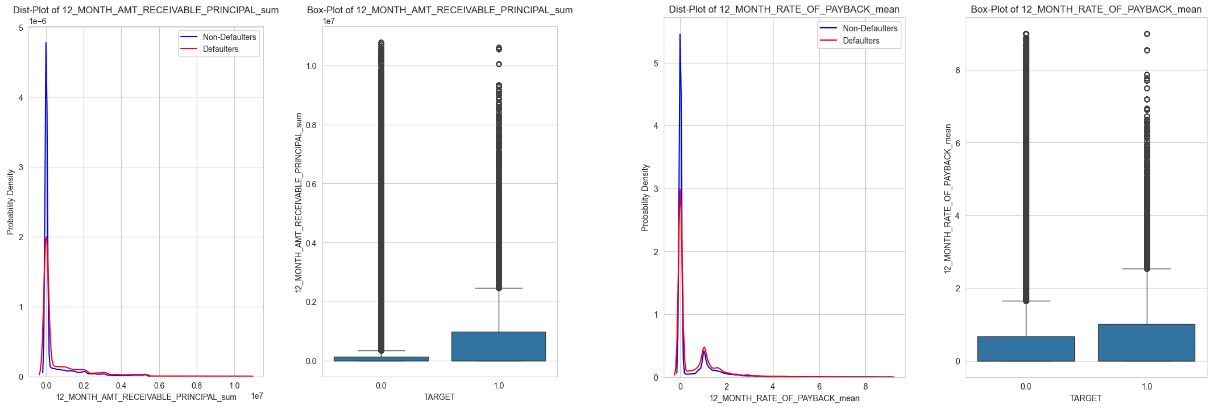

Figure 2.8: 12 Month Receivable Principal vs Rate of Payback

2.4 POS CASH Balance

The *POS_CASH_balance* dataset contains monthly balance snapshots of previous POS and cash loans with Home Credit. Each row represents the balance for a given month and credit, providing insights into the applicants' past loan behavior and its connection to their current loan applications.

The *POS_CASH_balance* dataset has only 2 columns with NaN values: *CNT_INSTALMENT* (remaining installments) and *CNT_INSTALMENT_FUTURE* (loan term), which account for just 0.26% of the data, so they are not a major concern.

Further analysis of distributions for variables like *MONTH_BALANCE*, *CNT_INSTALMENT*, and *CNT_INSTALMENT_FUTURE* reveals that most POS_CASH loans are short-term, typically lasting less than two years. For *SK_DPD* (Days Past Due), the majority of loans (96.97%)

have no overdue days, indicating timely repayments. However, there are notable outliers, with some loans showing overdue periods exceeding 3000 days. These extreme cases could signify bad debts or long-term defaults, warranting further investigation (Figure 2.9).

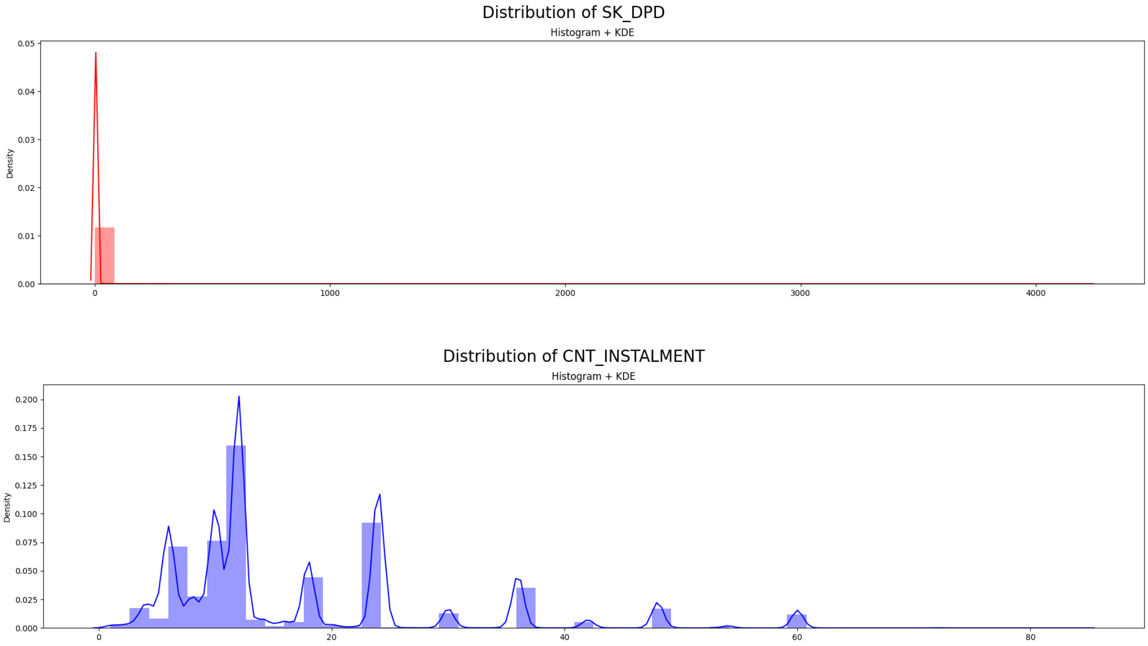

Figure 2.9: Distribution of DPD and CNT Installment POS CASH

Finally, when examining the box plot for *CNT_INSTALMENT_FUTURE*, we observe that defaulters generally have more installments remaining compared to non-defaulters, with higher median and upper whisker values (Figure **2.12**). This suggests that defaulters tend to have longer repayment periods. Additionally, defaulters are more likely to have canceled or amortized debt contracts, indicating difficulties in repayment or debt renegotiation. These patterns may help to better understand the differing repayment behaviors of defaulters and non-defaulters.

2.5 Installment Payment

The *Installment_Payment* dataset records repayment history for previous Home Credit loans related to our sample. Each row represents either a payment made or a missed payment for a single installment, offering insights into applicants' repayment behavior and financial habits.

A basic analysis of the dataset reveals that seven features have missing values, but they account for a minimal proportion (max 0.02%), so they do not pose a significant concern. A deeper analysis of the feature distributions with respect to the TARGET variable reveals that Defaulters tend to have *DAYS ENTRY PAYMENT* and *DAYS INSTALMENT PAYMENT* closer to the application date of the current loan, suggesting that they often make payments shortly before applying for new credit.

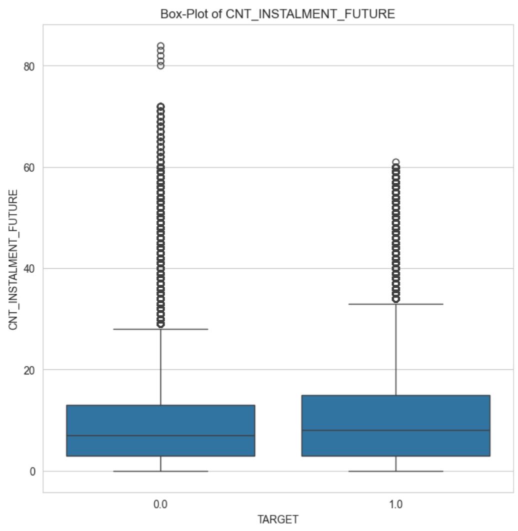

Figure 2.10: Box plot of count installment feature

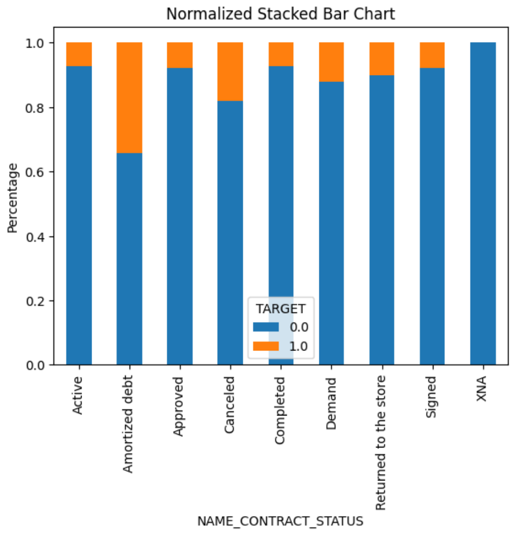

Figure 2.11: Normalize stacked bar chart of Name Contract Status

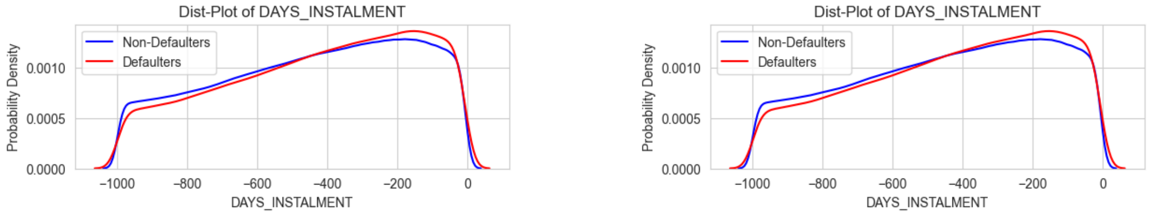

Figure 2.12: Distribution of Day Installment and Payment

2.6 Bureau and Bureau Balance

The *Bureau* and *Bureau_balance* datasets provide insights into clients' credit histories from other financial institutions before applying for Home Credit loans. The *bureau* dataset has 1.4M data points and 17 features, including *SK ID CURR* and *SK ID BUREAU* (Bureau ID). A single *SK_ID_CURR* can have multiple *SK ID BUREAU* values. Out of 263k unique *SK ID CURR* in Bureau, 210k appear in *application_train*, meaning some applicants lack previous credit history.

Firstly, we check missing values of 2 datasets. The *Bureau* dataset has 7 features containing missing values. Among them, the highest NaN values are observed with the column *AMT ANNUITY*, which has over 70% missing values. It can be seen that in all cases where *AMT ANNUITY* is missing, *DAYS ENDDATE FACT* is negative, which may indicate that the customer has fully paid off the loan and no longer has to make fixed monthly payments.

For the categorical variables in the *Bureau* dataset, *CREDIT ACTIVE* shows that most values are either ”Close” or ”Active”, while ”Bad debt” and ”Sold” have much smaller counts compared to ”Closed” and ”Active”. Additionally, ”Bad debt” and ”Sold” are less relevant to the *CREDIT ACTIVE* column, so we will replace these values with ”Active”. The *CREDIT CURRENCY* variable shows that ”currency 1” accounts for 99.88% of the data, offering little value for analysis. Therefore, we may remove the *CREDIT CURRENCY* column. Finally, for the *CREDIT TYPE* column, while it contains various types of credit, some categories appear only a few times. To simplify the dataset and avoid issues with sparse data, we will consolidate all credit types into a single category ”Other”, except for ”Consumer credit”, ”Credit card”, "Car loan", and "Mortgage".

For the numerical variables in the Bureau dataset, we made the following key observations:

The variables *DAYS REDIT* and *DAYS CREDIT UPDATE* show similar trends. Defaulters generally have shorter credit histories and more recent credit activity compared to non-defaulters. *DAYS ENDDATE FACT* shows the number of days since the client's previous credit ended. The box plot reveals that defaulters have a shorter period since their credit ended, while non-defaulters' credits typically ended longer ago.

An interesting finding is that some records in *DAYS ENDDATE FACT*, *DAYS CREDIT ENDDATE*, and *DAYS CREDIT UPDATE* show end dates over 100 years old, indicating those debts were likely settled a long time ago and may require further investigation.

3 | Modelling

For the final model, we use a total of 905 features. This number is concluded from multiple feature elimination tests using Recursive Feature Extraction (RFE). We tested the model with 500, 400, 300, and 250 features consecutively and found that there was not any significant model performance.

In this report, there will be several training techniques that are worth sharing. Intrinsically, Logistic Regression has a lot of weaknesses, such as only working well with linearly separated data and underforming under imbalanced data. Hence, we had to extensively adjust the dataset to overcome these shortcomings. Without these modeling techniques, we could only achieve a score of 0.55 on the public test. The pipeline consists of feature engineering, filling missing values, numerical transformation, handling imbalanced data, and k-fold evaluation.

3.1 Feature engineering

It is a relatively challenging task to convert time-series data to non-time-series data. From the original dataset, we created some linear and non-linear new variables by calculating ratios and differences between already existing features. After that, we used different types of Groupby functions including but not limited to min(), max(), mean(), count(), var(), skew(), nunique() for these newly created features. We did some experiments and discovered that using variance instead of standard deviation in these Groupby functions increases the model performance by 0.002 points (0.5675 - 0.5677).

With ratio calculation, we have to be extremely careful, as denominators can be 0, and failure to address this can entail a Division by Zero Error. For example, the new feature `Rate_of_Payback` is calculated as:

$$\texttt{Rate\_of\_Payback} = \frac{\texttt{AMT\_Payment\_Total\_Current}}{\texttt{AMT\_INST\_Min\_Regularity}},$$

and the `AMT_INST_Min_Regularity` feature can be 0 in situations where customers don't have to maintain a minimal installment.

We decided to convert zero values in this feature to null values; otherwise, they will be misunderstood as customers not paying anything. Doing this alone boosted our public test score from 0.567 to 0.569.

We discovered an innovative technique on Kaggle that substantially improved our model's predictive performance. Our approach involved leveraging two weak K-Nearest Neighbors (KNN) models with 500 nearest neighbors to generate unique feature variables, essentially answering the question: "How do the 500 closest customers behave in terms of loan default?" We focused on three primary feature categories for our KNN-derived insights: money (Ext source, credit/annuity ratio), background (Gender, Education, etc), and previous loans (AMT, CNT payment, etc). This KNN-derived feature proved to be a critical enhancement, elevating our public score from 0.54 to approximately 0.55—a meaningful improvement in model accuracy.

3.1.1 Filing NaN values

With such a huge dataset from the competition, using the KNN imputer to fill NaN values is not an ideal solution, so we rely mostly on filling NaN values with 0 and the mean value of each feature. However, there is no such "1 size fits all" solution for the problem of filling NaN values.

Logistic Regression is built on the foundation of each feature assuming a normal distribution, hence filling NaN values with the mean value from each feature can be a fast and reasonable solution. Based on this normal distribution assumption, samples that have values far from the mean value will trigger the weight better than those near the mean value. Additionally, when applying the Standard Scaler to NaN values, by filling them with the mean value they will become zero, or their weights are not activated, in other words. Indeed, this method of filling NaN values with the mean value does perform better than filling NaN values with 0.

However, on features with a high proportion of null values, filling these NaN values with the mean value completely changes their distribution. We conducted a test by filling features containing 80% of null values with zero and the model accuracy improved significantly. We assume that with highly null features, filling null values with zero makes the model learn better on those unique cases, and the parameter weights can be deactivated when they meet zero values.

3.2 Data Transformation

3.2.1 Categorical encoding

To deal with categorical features, we employed 2 different strategies: One-hot Encoding and Ordinal Encoding.

For Ordinal Encoding, we decide to encode features with only binary categories (e.g. Yes/No and Male/Female) and those that can be ranked (e.g. Education).

For One-hot Encoding, we tried to group similar variables into a single group, to avoid creating too many unnecessary features that could make the dataset spare and consequently hurt the model performance. Our initial model had roughly 1100 features, but later we squeezed the number of features down to 905.

3.2.2 Numerical transformation

At first, we thought using a Standard Scaler was enough for the numerical transformation. However, on testing our number transformation pipeline, we found some very interesting insights. A Standard Scaler, turns out, was not a one-side-fits-all solution, and most of our hyperparameter turning happened in this part.

We noticed that there were a lot of features in the datasets that were highly skewed. We defined some thresholds for feature skewness or kurtosis and applied Power Transformation to them accordingly. If the threshold is too low, most of the data will be converted to a normal distribution, which however overwrites all the meaningful distributions, leading to a decrease in model performance. If the threshold is too high, it will overlook the skewed features that could potentially harm the model. We decided to use a threshold of 3 for skewness and 23 for kurtosis. Applying this correct configuration in Power Transformation raised the model performance from .55 to .56 in the public test.

When preprocessing our data, we discovered that the Standard Scaler could potentially eliminate important zero values that lack extensive record depth. To address this issue, we strategically applied Min-Max scaling for features with minimal unique records (less than 50) and a baseline zero value. This approach ensures that zero values remain meaningful within the model, preserving their original significance while allowing appropriate scaling. By carefully selecting the Min-Max Scaler, we maintained the critical zero-value information and prevented these features from being overshadowed by outliers, thus preserving their potential predictive power.

3.3 Handling the imbalanced dataset

Logistic Regression is sensitive to imbalances in the dataset, so we need to balance the dataset back properly. We have tried numerous methods to address data imbalance, and none of them is better than the random resampling method. Even though SMOTE generates synthetic data, it does not work well on complex datasets and drags the model performance up to a .05 in GINI score decrease. We tried using a weak classifier to find weak samples (miss-classified samples, samples near the classification hyperplane, etc) and apply different sampling ratios on the weak and strong samples, accordingly. This method works in the early phases of the pipeline, but due to the uncertainty of threshold and ratios and limited time, we are unable to control the sampling effectively. We might have a detailed test conducted later. Since both train and test sets are imbalanced on default clients, we decided to up-sample the defaulters up to 15 times, or 1.5 times greater than non-default clients. This trick increased some points in the Gini score.