Predicting Loan Default Risk Using Machine Learning

A data-driven approach to enhance credit risk management for financial institutions.

Author:

Santosh Nagh Sirimalla

Data Analyst

Santoshnagh1@gmail.com | LinkedIN

Abstract

Financial institutions face significant challenges in managing loan default risks, which directly impact profitability and operational stability. This paper explores the use of advanced machine learning techniques, including logistic regression and Random Forest models, to predict default probabilities. By leveraging robust data preprocessing and predictive modeling, the analysis identifies key risk factors such as external credit scores and debt-to-income ratio. Key findings include an AUC of 0.9905 and 98.18% accuracy with Random Forest models. These results underscore the value of data-driven decision-making for tailored financial strategies.

Keywords:

- **Loan Default Risk**: Central focus of the analysis.
- **Machine Learning**: Methods applied to predictive modeling.
- **Random Forest**: The most effective model used in this study.
- **Financial Risk Management**: Practical application of insights.
- **Predictive Modeling**: Core methodology for risk assessment.
- **Data Analysis**: End-to-end process from preprocessing to insights.

INDEX

Executive Summary

The purpose of our capstone project, "Assessing Loan Default Risk Through Predictive Modeling and Quantitative Analysis," is to examine the issue of loan default risk prediction in the context of a financial institution. Loan default is a big risk to lenders since it affects profitability as well as operational stability. Our project uses predictive analytics on loan application details and previous loan histories from a large dataset published by Kaggle in order to predict the likelihood that a client will default on a given loan.

The core objective of this kind of study is to create a model for predicting loan defaults from the existing data streams. In analyzing this dataset we used exploratory data analysis (EDA) to find patterns and correlations in the values of things like income, loan amount, employment history, and so on. EDA key findings describe borrower behaviors and important features that increase the likelihood of challenging the loan payment.
For default prediction, which is a well-known binary classification problem, we use logistic regression, a really useful statistical technique when the task is to predict a category that is binary. Data preprocessing, feature selection, and validation are the process of the model training. Since it is an efficient way of managing data, SQL is used to combine and query large datasets. By integrating into SQL, we guarantee our data handling follows industry standards used in real-world data processing. The project findings table critical default risk predictors, including income stability, prior loan history, and demographics.

If the model identifies these factors, it can help financial institutions to refine their loan approval criteria, and reduce the risk of the default. This study emphasizes that its implication is not only predicting but also laying the foundation to help informed decisions in credit risk management, thereby contributing to sustainable financial practice.
This project concludes by showing the application of predictive analytics and quantitative techniques to solve a real business problem. The paper indicates the use to which data based decision making can be leveraged to improve the accuracy of lending assessment by improving lending institutions' resilience and profitability.

I. Introduction

In 2008, the global financial crisis was considered among the most severe economic downtrends in the history consequence of widespread loan defaults and poor risk management of the financial sector. Accurate assessment of loan default risk is critical as the ripple effects of the crisis illustrate – as that risk was not understood the collapse of major financial institutions, and a shock to the global economy ensued. For financial risk management, effective loan default prediction models have become an indispensable cornerstone – for lenders, a protective wall; for maintaining economic stability, a useful instrument.

In this capstone project, "Assessing Loan Default Risk Through Predictive Modeling and Quantitative Analysis," we address the critical task of predicting loan defaults using data driven methods. Our project seeks to identify the leading factors for default in loan applications, by utilizing a rich dataset of loan application records, and an applied combination of exploratory data analysis (EDA), statistical modeling, and predictive approaches. In particular, we use logistic regression, a strong binary classification method, to predict the likelihood of default based on a host of demographic, financial, and historical variables of loan applicants.

This also shows how SQL integrations help to work with large datasets and also standardizes our data processing pipeline for best efficiency and accuracy analysis in our study. Further, our findings are intended to not only better predict default loans, but also to provide insight into more generic areas of financial risk management. This project learns from past crises and leverages the most recent advances in predictive techniques to help lending institutions prepare for future financial disruptions.

Objectives:

- **Predict Loan Default Probability:** Apply one of the predictive modeling techniques to estimate the probability that a borrower will default on his or her loan based on his or her demographic, financial, and historical information.
- **Identify Key Risk Indicators:** Understand patterns and correlations between variables using exploratory data analysis (EDA). This should highlight important characteristics such as income, employment, past consumption behavior, and loan that play a significant role in the threat of default.
- **Leverage SQL for Efficient Data Management:** Create and append big data sets from involving files and files to merge appropriately, manage and picture loan application details effectively. This way we would learn about database management in the operational environment.
- **Demonstrate Application of Analytics Techniques:** Encourage the use of different business analytics methods, such as the technique of classification by means of logistic regression, Decision Trees, and Random Forest and feature engineering for data pre-processing and to carry out the assessment of model accuracy.
- **Provide Actionable Insights for Financial Institutions:** Based on model outcomes, the paper provides insights that can help financial institutions refine loan approval criteria, and effective risk management practices, and perhaps reduce default rates.

II. Data Description

With financial downturns such as the 2008 global crisis in mind, financial institutions are becoming more serious about predicting loan default risk to secure risk management practices and a lasting future. To accomplish this, we chose a dataset that is both high volume about loan applicants while providing real world challenges for financial data analysis. The data is from Kaggle and is fit for what we wish to accomplish. The data also contains rich information about borrowers without just loan default data and we can look into the drivers of default based on the borrower data and loan application history.

The details on demographics, financial status, employment details, and prior behaviors of borrowers can be viewed in this dataset. These features are great for understanding borrower profiles but also for developing predictive models. In addition, the data includes details that are often found in the real world, including missing values, and various types of data, all of which provides an added dimension for our analysis.

The dataset consists of three key files:

1. **Application Data:** This file lists information about the ongoing loan applications of a number of clients and each record represents a different application. Key variables include:
    - **SK_ID_CURR:** Unique ID for each loan application.
    - **TARGET:** The target variable indicating default risk (1 for default, 0 for non-default).
    - **Demographics:** Information such as CODE_GENDER (gender), FLAG_OWN_CAR (car ownership), CNT_CHILDREN (number of children), and NAME_INCOME_TYPE (income type) which shed light on the borrower's profile.
    - **Financial Details:** Variables like AMT_INCOME_TOTAL (total income), AMT_CREDIT (loan amount), and AMT_ANNUITY (annuity amount) that offer insights into the client's financial capacity.

- **Employment & Housing:** Other official values include: DAYS_EMPLOYED – employment length; NAME_HOUSED_TYPE – housing type.

2. **Previous Application Data:** This file contains details of previous loan applications made by each client in history. Overall, every row contains data about a specific previous loan application, granted or rejected. Key fields include:
    - **SK_ID_PREV:** A Unique number given to each of the past applications for identification purposes.
    - **SK_ID_CURR:** Identifier linking the past application to the current loan.
    - **Contract Information:** Other such parameters are NAME_CONTRACT_TYPE (type of loan contracted), AMT_APPLICATION (amount of money applied for), and AMT_CREDIT (amount of credit given).
    - **Application Status:** Fields like NAME_CONTRACT_STATUS (status of the previous application) and DAYS_DECISION (days relative to current application) provide a historical record of the client's interactions with the lender.
3. **Columns Description:** This file is a metadata document that offers detailed descriptions of each column across the other two files. It serves as a valuable reference to ensure accurate interpretation of each feature, particularly those with industry-specific terms or encoded values.

III. Methodology

In our project, "Assessing Loan Default Risk Through Predictive Modeling and Quantitative Analysis," we perform a structured analysis of our loan data set, a combination of loan application data and historical loan data, to assess patterns in default. This methodology describes the actions taken to develop a strong predictive model for loan default risk, which leverages different business analytics techniques. Our methodological framework includes the following stages:

1. **Data Collection and Understanding**

   The dataset used in this study is sourced from Kaggle and includes three primary files: previous_application, application_data, and columns_description. This file application_data includes the target variable for default risk, along with details of each loan application. columns_description shows the feature definitions and the previous_application, which is how we've seen each applicant in historical terms. With this initial understanding of the data we can then ascertain what are the key variables for analysis and align our goal.

2. **Data Preprocessing and Cleaning**

   Our first step is data preprocessing as we have large volume of records and a diverse set of variables. That involves dealing with missing values, throwing out duplicates, and all records being coupled consistently. In addition, we handle outliers which might distort the model predictions, and where appropriate, convert to numerical forms categorical variables. Also, data cleaning prepares the data to be trained on a correct noise model.

3. **SQL Integration for Data Management**

   We use SQL to do data management jobs to clean up large scale data processing. We join application_data together with previous_application using key identifiers to ensure we capture every applicant's history. This integration also allows us to synchronously retrieve and optimize the dataset handily across various stages of the analysis.

4. **Exploratory Data Analysis (EDA)**

   In this stage, we do EDA to understand the structure of the data and its relationship to the variable. We use visualizations and statistical summaries to discover patterns in demographics, finance, and behavioral features as well as their relation to loan default. The features mentioned above (i.e. income level, employment stability and previous loan performance) are identified as important for predicting default probability through EDA.

5. **Feature Engineering**

   We create new features and modify existing ones in order to improve model performance. Time based variables are transformed, data is normalized, and derived features are created to capture deeper insight into applicant risk factors, this is called feature engineering. For example, having engineered employment duration or even the loan to income (LTV) ratios to increase the predictive power of the model.

6. **Predictive Modeling with Logistic Regression**

   Since we are trying to predict loan default, which is a binary classification task, we realize that the best binary classification model to use is logistic regression. This method is appropriate for binary outcomes and enables interpretability of the influence of various features on the probability of default. Additionally, Decision trees and Random Forest techniques are used to complement and guarantee accuracy and reliability.

7. **Model Validation and Evaluation**

   We then cross-validate the model using accuracy, precision, recall and the AUC to evaluate the model's effectiveness. Validation is done to ensure that not only our model is predictive, but also will be generalizable to unseen data so as to avoid overfitting risk.

8. **Application of Findings to Financial Decision-Making**

   We then use our model's insights to formulate actionable recommendations for financial institutions. Based on practical applications of financial risk

management, these recommendations aim to identify high risk customers and derive improved ways to approve loans with an eye towards reducing default risk.

Initial Data Exploration and Cleaning

The main objective we considered in data cleaning is to make datasets more suitable for analysis leading to accurate and reliable results. With many records and variables in raw form, cleaning enables us to extract accurate insights that may be hindered by issues to do with missing or inconsistent data. Here's a brief explanation of the key steps we've taken:

IV. Findings and Analysis

Process, Analysis and Findings

Data Preprocessing

Effective data preprocessing is really important for accurate and interpretable analytics. We began with two primary datasets: Application Data (app_data) representing details of current loan applications and Previous Application Data (prev_data) with information on past loan applications. The following sections describe our preprocessing steps in detail, including the rationale for decisions taken to ensure data quality.

1. Initial Data Exploration

Before beginning data cleaning, we used the str() and summary() functions to see the data structures, data types, and summary statistics. This provided us with an understanding of each variable, including potential outliers and missing data patterns. Conducting this initial exploration was essential for designing an effective cleaning strategy, as it provided us with a holistic view of the data's state and highlighted specific areas needing attention.

app_data Dataset

- **Dataset Size:** Contains 307,511 observations across 122 variables.
- **Demographic and Loan-Related Variables:**
    - **AMT_INCOME_TOTAL:** Ranges from 25,650 to 117 million, with a median of 147,150. The large spread suggests potential outliers or a highly diverse income range.
    - **AMT_CREDIT:** Ranges widely from 45,000 to 4,050,000, with a median of 513,531, indicating significant variability in loan sizes.
    - **DAYS_BIRTH:** Ranges from -25,229 to -7,489 (in days), corresponding to an age range from approximately 20 to 69 years.
    - **DAYS_EMPLOYED:** This variable has a maximum value of 365,243, which suggests erroneous or placeholder data, most likely referring to specific circumstances or missing values.

- **Data Quality Issues:**

  Over 150,000 observations are missing, particularly in variables like OWN_CAR_AGE (202,929 missing) and other building attributes like APARTMENTS_AVG, BASEMENTAREA_AVG, YEARS_BEGINEXPLUATATION_AVG, and others.

- **Flag Variables:** While binary, flags like FLAG_OWN_CAR, FLAG_PHONE, and FLAG_EMAIL show imbalances. For example, FLAG_PHONE (mean of 0.2811) showed that only a very small fraction of applicants gave their phone number.

prev_data Dataset

- **Dataset Size:** 1,670,214 observations of data spanning 37 variables that relate to prior applications for the present applicants are included.
- **Loan and Payment Variables:**
  - **AMT_APPLICATION** and **AMT_CREDIT** display considerable variability, with median values of 71,046 and 80,541, respectively, but maximum values reaching over 6.9 million, indicating significant diversity in the credit needs of previous applications.
  - **CNT_PAYMENT:** Ranges from 0 to 84, with a median of 12 payments, which may represent various loan terms and repayment plans.
- **Rate Variables:** The **RATE_DOWN_PAYMENT**, **RATE_INTEREST_PRIMARY**, and **RATE_INTEREST_PRIVILEGED** fields have numerous missing values, with **RATE_INTEREST_PRIMARY** missing in 1,664,263 observations, suggesting incomplete rate data.

2. Handling Missing Values

If not handled properly, missing data might create bias and reduce the accuracy of model predictions. We addressed missing variables in a methodical manner, keeping in mind the requirement to balance data completeness with information retention.

- **Percentage Calculation:** We calculated the proportion of missing values in each dataset and for each column. Columns with substantial missing data often add noise rather than value; thus, this step allowed us to quantify the degree of incompleteness and decide how to proceed.
- **Overall Missing Values:**
    - **app_data:** About 22.36% of the data values are absent.
    - **prev_data:** Missing values make up about 16.65% of the data.
- **Column-Wise Missing Values in app_data:**
    - Most columns in **app_data** had negligible to no missing values. Notable columns with significant missing data include:
        - Many of the app_data's columns had little to no missing values. Among the notable columns with noteworthy missing data are OWN_CAR_AGE, which has 65.99% missing data.
          APARTMENTS_AVG (50.75%), BASEMENTAREA_AVG (58.52%), YEARS_BEGINEXPLUATATION_AVG (48.78%), and YEARS_BUILD_AVG (66.50%) are columns that describe building characteristics.
        - Additionally, more than 50% of the data in other building-related columns (such as COMMONAREA_AVG, ELEVATORS_AVG, and ENTRANCES_AVG) is missing.
        - **EXT_SOURCE_1** (56.38%) and **EXT_SOURCE_3** (19.83%) exhibit notable missingness among external factors.
- **Column-Wise Missing Values in prev_data:**
    - Columns with significant missing values include:
        - **AMT_ANNUITY** with 22.29% missing.
        - **AMT_GOODS_PRICE** with 23.08% missing.

        - **RATE_DOWN_PAYMENT** and *RATE_INTEREST_* have high missingness, with **RATE_INTEREST_PRIMARY** and **RATE_INTEREST_PRIVILEGED** missing 99.64% of their data.

- DAYS_FIRST_DRAWING, DAYS_FIRST_DUE, DAYS_LAST_DUE_1ST_VERSION, DAYS_LAST_DUE, and DAYS_TERMINATION all have around 40.30% missing values.
- **Handling Missing Values:**
    - To enhance the quality of the data, columns with more than 40% missing values were eliminated from both datasets. Because imputing values at such high degrees of missingness could contaminate the data and produce erroneous correlations, this criterion was used. We made sure the remaining data would still represent the majority of the original information while retaining more significant patterns by eliminating certain columns.
    - To ensure consistency, missing values in **categorical columns** that were converted to factors were filled up with "Missing".
    - **Numeric columns** with missing values were identified and assessed for appropriate imputation:
        - **AMT_ANNUITY** and **AMT_GOODS_PRICE** were among those imputed with mean or median values, while columns like **CNT_FAM_MEMBERS** and social circle counts (OBS_*DEF_*) were imputed with mode values.

3. Data Transformation

In order to express the number of days in respect to a reference point (such as days since birth or job start date), the date-related variables in the dataset were initially encoded as negative numbers. We used the following transformation to make these variables easier to understand and streamline further analysis:

Absolute Value Transformation: Turning **DAYS_BIRTH**, **DAYS_EMPLOYED**, **DAYS_REGISTRATION**, and **DAYS_ID_PUBLISH** from negative to positive. For example, the original format of **DAYS_BIRTH** ranged from -25,229 (73 years) to -7,489 (20 years), which was age in days. This can now be converted into a positive range that promotes readability and minimizes the chance of misinterpretation in analytical models since they can now represent age directly in days or be translated to years for demographic analysis.

4. Feature Engineering and Aggregation

Feature engineering was performed in an effort to expose the existing patterns within the data set and improve the predictive models by creating new features. The following engineered features were developed based on specific insights drawn from both datasets:

4.1 Aggregated Address Mismatch

- **Rationale:** Six attributes in app_data were identified to reflect the discrepancies; REG_REGION_NOT_LIVE_REGION, REG_REGION_NOT_WORK_REGION, LIVE_REGION_NOT_WORK_REGION, REG_CITY_NOT_LIVE_CITY, REG_CITY_NOT_WORK_CITY, LIVE_CITY_NOT_WORK_CITY. To do this, we combined these columns into a single variable which is simply a count of address mismatches for an applicant, the variable is named ADDRESS_MISMATCH_COUNT.
- **Potential Implications:** This feature raises questions concerning stability of address, where less address mismatch count might be correlated to stability in finances, or lower risk. Stability of address can observe certain lack of stability in the lifestyle or other conditions which can affect the credit standing.
- **Values**: The numerical range of ADDRESS_MISMATCH_COUNT is between 0 and 6; a majority of applicants, 59%, did not record more than one mismatch pointing to their residential stability and 1 % of applicants had a high score of 6 mismatches.

4.2 Credit Bureau Inquiry Weighting

- **Rationale:** The general dataset of the graphical application has six fields concerning credit bureau inquiries as follows: AMT_REQ_CREDIT_BUREAU_HOUR, AMT_REQ_CREDIT_BUREAU_DAY, AMT_REQ_CREDIT_BUREAU_WEEK, AMT_REQ_CREDIT_BUREAU_MON, AMT_REQ_CREDIT_BUREAU_QRT, and AMT_REQ_CREDIT_BUREAU_YEAR. A new feature, WEIGHTED_CREDIT_BUREAU_INQUIRIES, was derived from the total credit bureau inquiries with the most recent three months taking more than one time weight.
- **Calculation:** Each of the columns corresponding to the inquiries was multiplied by the recency weight estimate, the results were added together and then divided by the sum of the weights and given a normalized value. It gives a total weighted score of recent credit inquiry activity that person has undergone.
- **Values: WEIGHTED_CREDIT_BUREAU_INQUIRIES** averages at 0.25 and has a median value of 0.21 which means most of the applicants had one or two recent inquiries at most. A peak at two means most users have low or no recent usage rates.
- **Implications:** Recent and more frequent inquiry is related to credit-seeking behavior, which is commonly attributed to credit crunch. It can help models take into account the applicant's last several inquiries without giving much importance to older ones that are less likely to be relevant.

4.3 Social Circle Default Rate

- **Rationale:** The variables, number of contacts 30 days before and 60 days before default are (OBS_30_CNT_SOCIAL_CIRCLE, DEF_30_CNT_SOCIAL_CIRCLE, OBS_60_CNT_SOCIAL_CIRCLE, DEF_60_CNT_SOCIAL_CIRCLE). We came up with the creation of SOCIAL_CIRCLE_DEFAULT_RATE to define the importance of an applicant's social circle by dividing the total defaults by total observations.
- **Calculation:** If the sum of **OBS_30_CNT_SOCIAL_CIRCLE** and **OBS_60_CNT_SOCIAL_CIRCLE** was non-zero, the ratio of defaults to total social observations was calculated. For applicants without social observations, the rate was set to zero.

- **Values: SOCIAL_CIRCLE_DEFAULT_RATE** has a mean of 0.06 and 94% of the applicants have a default rate of zero in their social circle. However, a small portion falls in the high social default rate, which could mean a high risk of default.
- **Implications:** High default rates in an applicant's social circle could signal financial strain or indicate riskier financial behaviors, as research suggests financial behaviors can influence social networks.

4.4 Document and Contact Aggregations

Two other aggregation features were developed to aggregate document submission and contact completeness into one metric.

- **Documents Provided (NUM_DOCUMENTS_PROVIDED):**
    - **Rationale:** This feature aggregates binary document flags (from FLAG_DOCUMENT_2 to FLAG_DOCUMENT_21), flags signaling the existence of special documents in the application. These flags add up to give a measure of document completeness for each application.
    - **Values:** The document count on a per applicant base ranges from 0 to 4 and averages 0.93, indicating that most turn in at least one document, and a smaller sub segment turns in more documents.
    - **Implications:** As a result, loan applications with greater and more transparent documentation are perceived as reducing the risk profile, as being thorough and transparent are typically viewed as positive, indicators by financial institutions.
- **Contact Information (CONTACT_PROVIDED_SUM):**
    - **Rationale:** Several FLAG columns regarding contact methods (FLAG_MOBIL, FLAG_EMP_PHONE, FLAG_WORK_PHONE, FLAG_CONT_MOBILE, FLAG_PHONE, FLAG_EMAIL) are aggregated as CONTACT_PROVIDED_SUM to give a single metric which captures if the applicant is accessible.

- **Values:** It varies from 3 to 10 with an average count of 6.39. The more count there is, the higher it means that the applicant is likely to have provided more than one contact option to make it contactable.
- **Implications:** The higher contact information completeness of an applicant may also sway financial institutions to regard such applicants more reliably, lessening default risk owing to better-wired communication.

5. Treatment of Outliers and Outlier Analysis

This was first an important step in outlier handling so that we do not end up doing skewed, analytical results. We analyzed outliers in the **app_data** dataset using the **dlookr** package for R and capped extreme values to reduce their effect, without eliminating important data.

- We had to respect applications with possibly atypical financing structure and for this we left rows in **prev_data** in the **AMT_ANNUITY** column with a value of 0.
- We excluded rows that had 0 values in **AMT_APPLICATION** and **AMT_CREDIT**, which meant that those zeros would point to invalid applications, and mislead the financial calculations.

6. Column Reduction and Consolidation

To improve analytical efficiency, we reduced the dataset size by decreasing the number of redundant columns.

- To keep it simple, we didn't include columns of FLAG_DOCUMENT_2 through FLAG_DOCUMENT_21, since these weren't adding much additional analytics value and were creating some complexity.
- As a result of this, we were able to isolate address mismatch columns into one feature, which then helped us avoid o.verloading the dataset with too many location-based indicators.

7. Normalization and Scaling Decisions

After some thought though, we didn't pursue normalization as the scale worked for us we had the capability to work with R and SQL. This decision did not require justification based on the completion of normalization, though it could be undertaken later in the road based on the model requirements.

8. Final Data Preparation Summary

It was so painstaking getting it right so that we could have a cohesive, high quality dataset to enable reliable predictive modeling. Balancing structural modifications with thorough cleaning of our data preparation strategy were the goals of our project. To that end, the data was cleaned in R, combined using SQL and visualized in Tableau to have a solid foundation of data to work on.

9. SQL Integration and Final Export

Moreover, we integrated multiple datasets and applied data cleaning techniques before preparing the data to be used in predictive modeling for analyzing loan default risk. This was a process where SQL integration was such an important part of it, and we were able to merge and query several data tables simultaneously with ease and make sure that we have processes that are systematic with regard to working with large datasets. We used SQL to manage data quickly and associate tables to jog, filter, and transform without a problem. Data import followed by table creation, merging operations, and preparing a unified table was done to then be analyzed further in R and Tableau.

Objectives of SQL Integration

The SQL integration process aimed to achieve the following:

1. **Centralize Data Management:** The different cleaned and processed datasets are stored in a single database for version control and traceability of datasets.
2. **Perform Efficient Joins:** We can leverage the fact that SQL is able to optimize joins between current loan application data and prior loan application data.

3. **Ensure Data Consistency:** It resolves naming conflicts on attributes similar in meaning, but in different contexts, in different tables and null values.
4. **Prepare Data for Analysis:** We finally merge this dataset with another one that we will use next when we start doing more statistical analysis, machine learning, or visualization in R and Tableau.

Data Preparation

1. **Exporting Cleaned Data:**

   We first exported our cleaned dataframes from R as CSV files, app_data_cleaned_2 and prev_data_cleaned_2. The latest versions of cleaned and processed current and previous application data, respectively, were these files.
2. **Importing Data into SQL Database:**

   We imported the CSV files into a new SQLite database and built tables according to each dataset using **DB Browser for SQLite**. Using SQL, we imported the tables into the database, allowing us to merge and query on them.
3. **Column Renaming:**

   It made it easier to reference columns with the same name in different rows, so we renamed them to indicate if they came from this application or from a previous loan application. For instance, I changed NAME_CONTRACT_TYPE to CURRENT_NAME_CONTRACT_TYPE and PREVIOUS_NAME_CONTRACT_TYPE.
4. **Handling NULL Values:**

   In both tables, we replaced null values in the **NAME_TYPE_SUITE** column with the help of the COALESCE function. To maintain data consistency, we set any missing values to 'Unknown.'
5. **Left Join:**

   In order not to lose any records of the current application table (**app_data_cleaned_2**), we used a left join to ensure that the previous application table (**prev_data_cleaned_2**) did not contain any records that may affect the current application.

Challenges and Solutions

1. **Handling Null Values:** In both datasets, some fields were missing, for example NAME_TYPE_SUITE. Since they can be null, we replaced them using COALESCE with 'Unknown'.
2. **Naming Conflicts:** In both datasets, there were similar column names, which referred to different contexts. To eliminate any ambiguity, we renamed these columns so they clearly point to either the current or the previous application.
3. **Database Performance:** Computational intensity arose given the size of the datasets, which caused bottlenecks. We reduced processing time by optimizing our database configuration and indexing relevant fields like **SK_ID_CURR**.

Exporting the Merged Table

Once we verified that the **merged_data** table was successfully created, we exported it to a CSV file, and started our analysis in R and Tableau. By merging the three datasets into this final one, we obtain a complete view of each of our client's loan application history, which is useful for further statistical and visual analysis.

Merged Dataset Structure:

1. **Dimensions:**
    - Observations: 1,153,805 rows.
    - Variables: 60 columns.
2. **Data Types:**
    - **Integer (int):** Has identifiers (SK_ID_CURR, SK_ID_PREV), numeric count (CNT_CHILDREN, CNT_PAYMENT) and date fields (DAYS_BIRTH, DAYS_EMPLOYED).

- **Numeric (num):** He describes continuous variables, which include income (AMT_INCOME_TOTAL), loan amount (CURRENT_AMT_CREDIT), and external credit scores (EXT_SOURCE_2, EXT_SOURCE_3).
- **Factor:** CODE_GENDER, CURRENT_NAME_CONTRACT_TYPE, NAME_EDUCATION_TYPE variables were converted to factors. All categorical columns have been converted to factors. Systematically handled, these have included the proper use of placeholders when categories are missing and to impute numerical values unless otherwise specified. All of this matches machine learning requirements; for example, a logical feature like NO_PREVIOUS_LOANS (a binary flag) is encoded as an int.

Summary Statistics

1. **Target Variable:**
    - The TARGET variable, representing loan default (1) and non-default (0), is highly imbalanced:
        - Mean: 0.08335, that means ~8.3% of applicants defaulted.
2. **Demographic Variables:**
    - **Gender:**
        - CODE_GENDER: There are a lot of 'XNA' entries, however the majority of applicants are female (~67%).
    - **Family and Housing:**
        - **NAME_FAMILY_STATUS:** More than 64% of people are married.
        - **NAME_HOUSING_TYPE:** "The majority of our clients live in 'House / apartment' (~89%)."
    - **Income:**

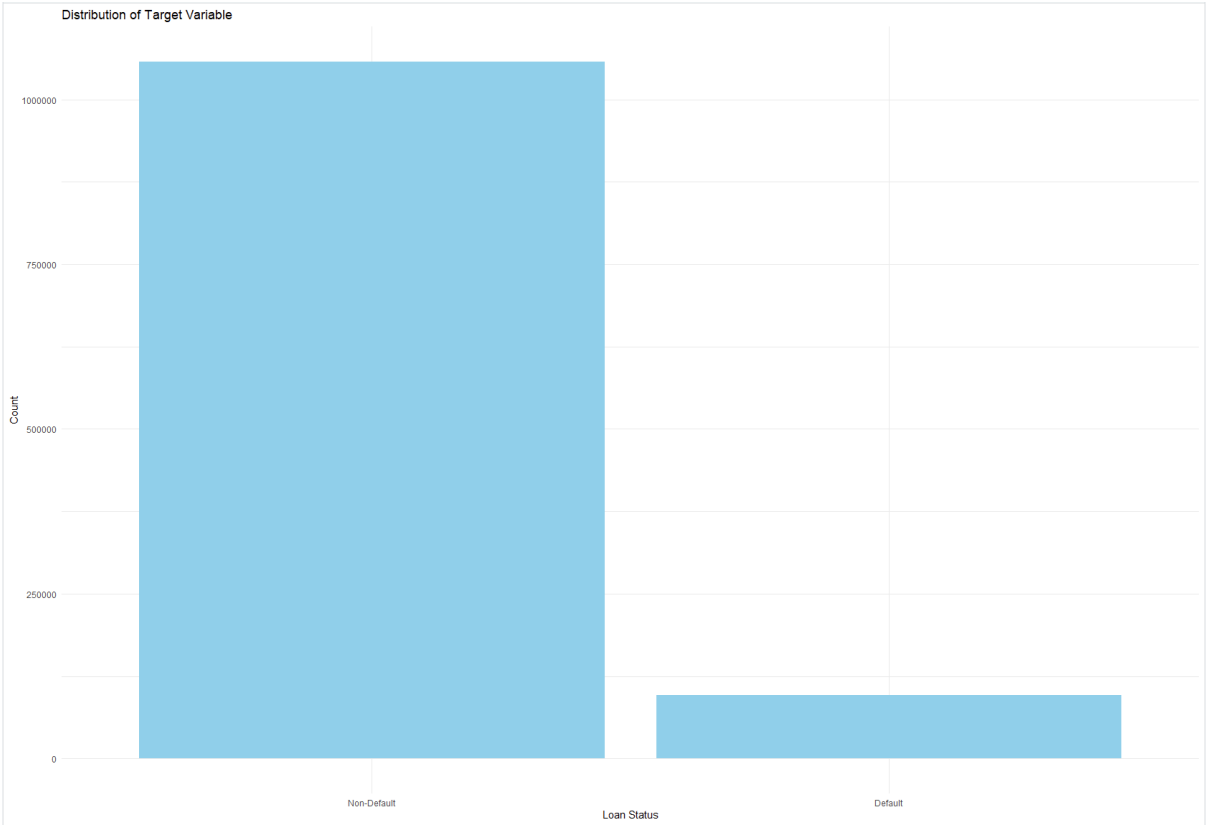
        - AMT_INCOME_TOTAL vary from 25,650 to a large maximum of 117,000,000, which indicates the outliers.

3. **Financial Data:**
    - **Income:**
        - AMT_INCOME_TOTAL ranges from 25,650 to an extreme maximum of 117,000,000, suggesting the presence of outliers.
        - Median income: 157,500, indicating skewness.
    - **Credit Amount:**
        - CURRENT_AMT_CREDIT mean: 593,483, and has outliers up to 4,050,000.
    - **Goods Price:**
        - CURRENT_AMT_GOODS_PRICE closely tracks with credit values, which often mirrors credit practice.
4. **Historical Loan Data:**
    - They have entries missing SK_ID_PREV, about ~1.45% of entries. That means applications with no prior loans.
    - **Financial Attributes:**
        - PREVIOUS_AMT_CREDIT: A large range of up to 4,509,688, also mean of 240,583.
        - PREVIOUS_AMT_ANNUITY: Typical monthly payments Mean of 15,394.
5. **Categorical Levels:**
    - **NAME_CONTRACT_STATUS:** Most (~76%) are "Approved" loans, and a significant number (~19%) are "Refused."

- **NAME_CASH_LOAN_PURPOSE:** Emphasizes missing or problematic data, although "XAP", "XNA" categories dominate the data.
- **NAME_INCOME_TYPE:** Most common source, "Working" (~51%).

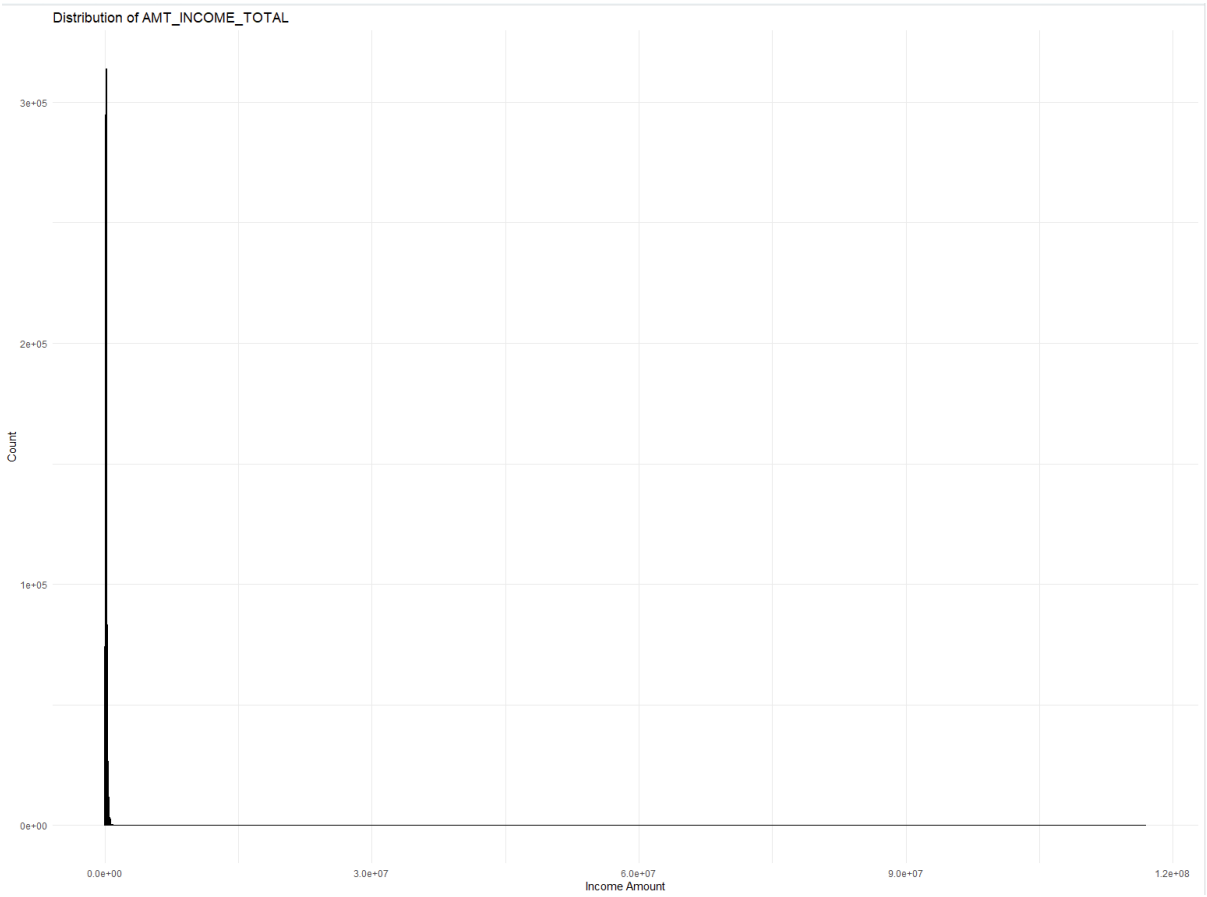
6. **External Scores:**
   - EXT_SOURCE_2 and EXT_SOURCE_3 show mean values (~0.512 and ~0.498, respectively), supporting their predictive potential.

Distribution of Target Variable



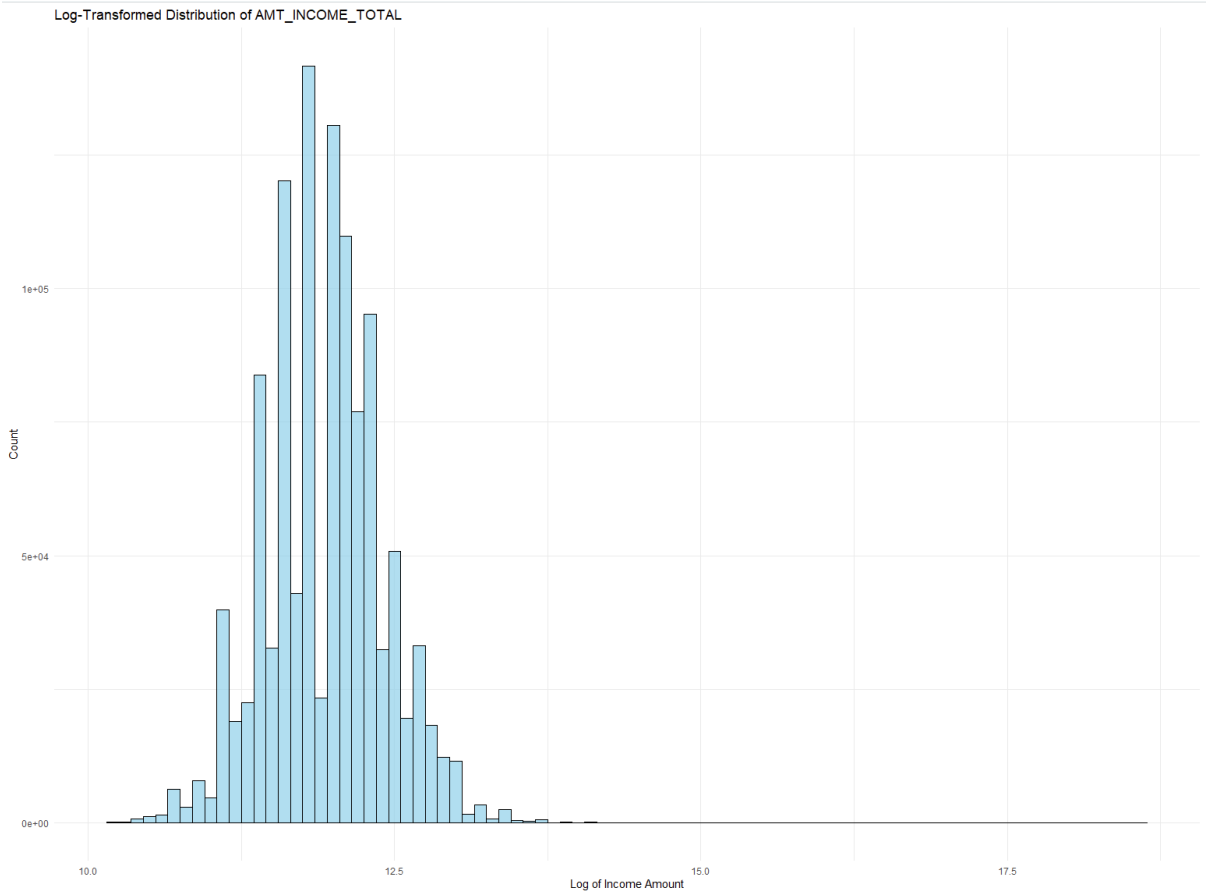
The number of defaults in the dataset is very low compared to the number of non-defaults. This could have some effects on our model predicting defaults.

Initial Histogram of AMT_INCOME_TOTAL:



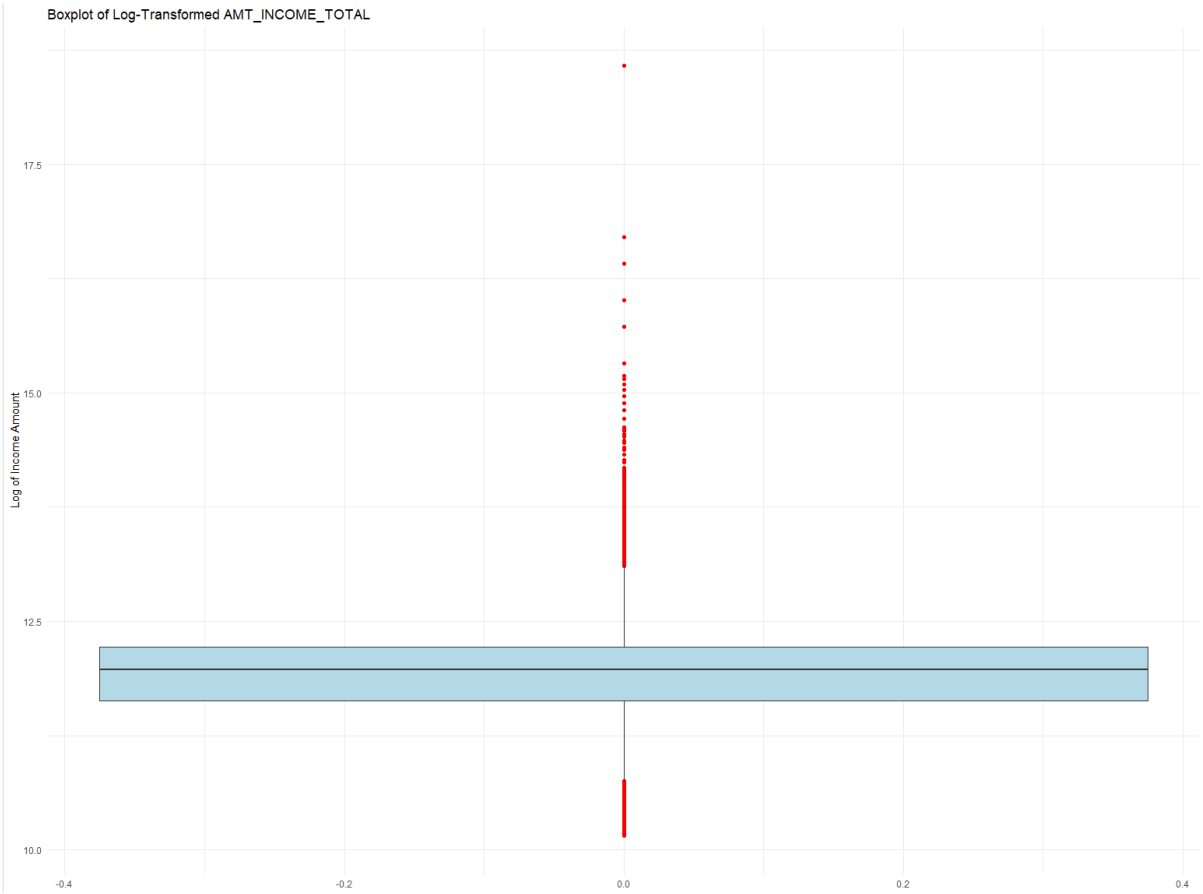
Raw income amounts were plotted to form a histogram to examine the raw distribution. The results implied an extreme right skew in that the bulk of the observations occurred at lower income levels. Therefore the data was subject to skewness that would make it difficult to interpret the variability and trends.

Log Transformation of AMT_INCOME_TOTAL:



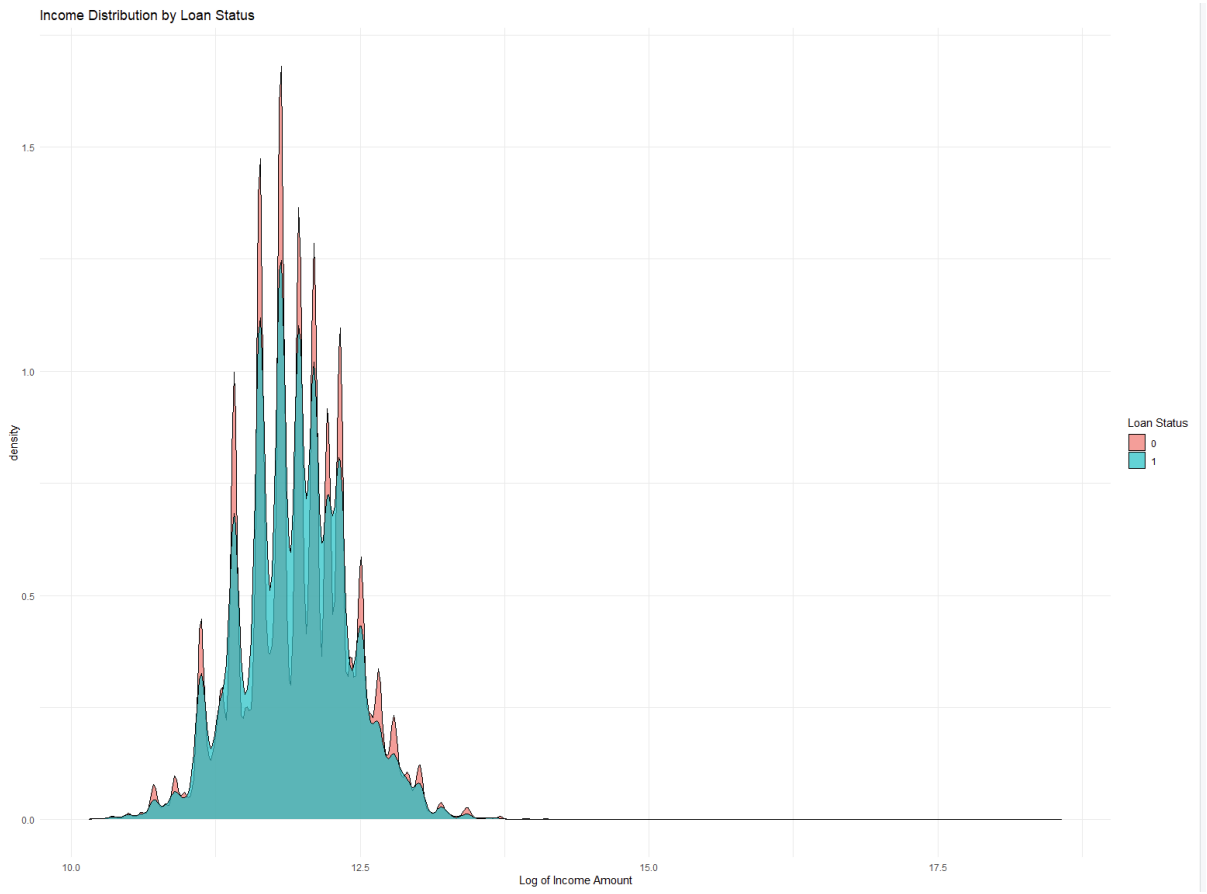
To deal with the skewness and make the distribution more visual, a log transformation (log1p, so as to avoid crazy numbers when there are zero values) was taken on the income data. The transformed histogram was quite symmetric and easy to interpret, so we could see the patterns at different income levels.

Boxplot of Log-Transformed Income:



Further analysis of the transformed income distribution was done using a boxplot to observe potential outliers. The points above the whiskers indicated some high-income outliers, but the bulk of the data was stuck tight around the median with little spread.

Density Plot of Income by Loan Status (TARGET):



To compare income distributions of clients with and without payment difficulties, a density plot was run.

- TARGET = 0 (Non-Default): This group was somewhat more dense than others at all income levels for representing clients who were not experiencing payment difficulties.
- TARGET = 1 (Default): This group represented clients with payment difficulties and exhibited a similar distribution, but with lower density, particularly in higher income levels.

What we know so far

Income Distribution:

- The raw AMT_INCOME_TOTAL is highly right skewed with the vast majority of values close to the lower end and an abnormally large number of outliers on the other end.
- Log transformation brings down skewness, normalizing the distribution and enhancing interpretability, especially for our machine learning models.

Outliers:

- Their influence is compressed by log transformation and their ranking is preserved. Log transformation compresses their influence while preserving their ranking.
- Boxplots highlight these outliers, which warrant further investigation for potential irregularities or unique borrower profiles.

Loan Default Relationship:

- High-income outliers represent affluent clients, but they can distort statistical summaries and model performance. Their influence is compressed by log transformation and their ranking is preserved. Boxplots identify these outliers, which may require further probing for possible irregularities or different borrower type.
- Financial vulnerability is signified by the fact that defaulting clients (TARGET = 1) are more likely in lower income groups.
- The higher incomes are dominated by non-defaulting clients (TARGET = 0), indicating more financial stability.

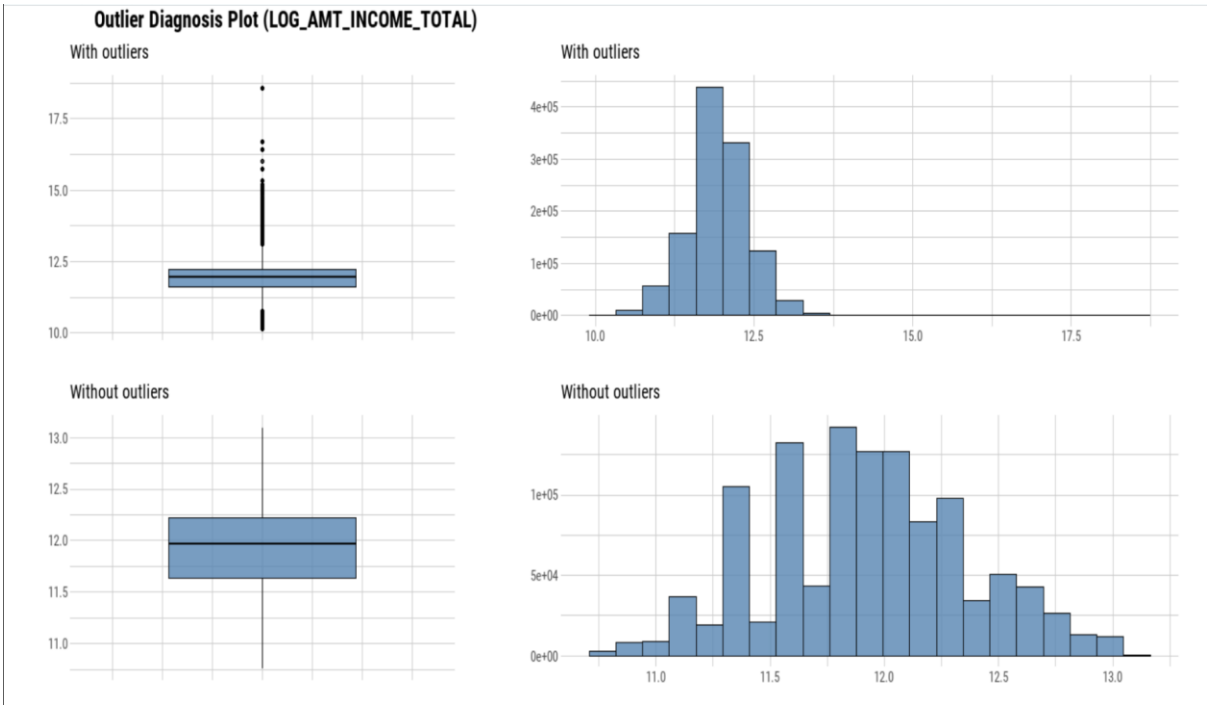
Key Takeaways:

- Income influences default risk but requires integration with other variables for accurate predictions.
- Outliers must be carefully managed to avoid skewing results.

- Both high- and low-income defaulting clients highlight the importance of broader socio-economic and behavioral factors in assessing risk.

```
> outlier_report_log
                         variables outliers_cnt outliers_ratio outliers_mean    with_mean without_mean
1                       SK_ID_CURR            0     0.00000000          NaN 2.784565e+05 2.784565e+05
2                           TARGET        96164     8.33451060 1.000000e+00 8.334511e-02 0.000000e+00
3                     CNT_CHILDREN        16541     1.43360447 3.187413e+00 4.133038e-01 3.729556e-01
4                AMT_INCOME_TOTAL        53569     4.64281226 4.577763e+05 1.726497e+05 1.587672e+05
5              CURRENT_AMT_CREDIT        20975     1.81789817 1.889415e+06 5.934830e+05 5.694880e+05
6             CURRENT_AMT_ANNUITY        25158     2.18043777 7.391765e+04 2.706825e+04 2.602396e+04
7         CURRENT_AMT_GOODS_PRICE        49502     4.29032635 1.535081e+06 5.335535e+05 4.886586e+05
8       REGION_POPULATION_RELATIVE        27647     2.39615880 7.250800e-02 2.080001e-02 1.953059e-02
9                       DAYS_BIRTH            0     0.00000000          NaN 1.621984e+04 1.621984e+04
10                 DAYS_EMPLOYED        83790     7.26205901 8.202957e+03 2.313398e+03 1.852202e+03
11             DAYS_REGISTRATION         2698     0.23383501 1.685229e+04 4.984974e+03 4.957159e+03
12               DAYS_ID_PUBLISH            0     0.00000000          NaN 3.033594e+03 3.033594e+03
13               CNT_FAM_MEMBERS        15539     1.34676137 5.191003e+00 2.158588e+00 2.117191e+00
14                  EXT_SOURCE_2            0     0.00000000          NaN 5.122738e-01 5.122738e-01
15                  EXT_SOURCE_3         4725     0.40951461 4.871838e-03 4.979930e-01 5.000207e-01
16         DAYS_LAST_PHONE_CHANGE          439     0.03804802 3.811034e+03 1.090494e+03 1.089459e+03
17         ADDRESS_MISMATCH_COUNT       287060    24.87942070 2.285944e+00 5.687296e-01 0.000000e+00
18 WEIGHTED_CREDIT_BUREAU_INQUIRIES        18853     1.63398495 1.537653e+00 3.351643e-01 3.151894e-01
19       SOCIAL_CIRCLE_DEFAULT_RATE       140216    12.15248677 5.038164e-01 6.122622e-02 0.000000e+00
20         NUM_DOCUMENTS_PROVIDED       111689     9.68005859 3.400424e-01 9.361157e-01 1.000000e+00
21           CONTACT_PROVIDED_SUM        67496     5.84986198 7.128378e+00 6.427448e+00 6.383897e+00
22                       SK_ID_PREV        16681     1.44573823 0.000000e+00 1.887126e+06 1.914810e+06
23            PREVIOUS_AMT_ANNUITY        77481     6.71525951 5.230676e+04 1.539435e+04 1.273715e+04
24                 AMT_APPLICATION       124079    10.75389689 9.651996e+05 2.147174e+05 1.242864e+05
25              PREVIOUS_AMT_CREDIT       122379    10.60655830 1.060436e+06 2.405829e+05 1.433071e+05
26         PREVIOUS_AMT_GOODS_PRICE       127530    11.05299422 9.525704e+05 2.191420e+05 1.280026e+05
27         HOUR_APPR_PROCESS_START          919     0.07964951 4.868335e+00 1.259482e+01 1.260107e+01
28                  DAYS_DECISION            0     0.00000000          NaN 1.039450e+03 1.039450e+03
29                    CNT_PAYMENT        94650     8.20329258 5.356503e+01 1.572345e+01 1.234179e+01
30              NO_PREVIOUS_LOANS        16681     1.44573823 1.000000e+00 1.445738e-02 0.000000e+00
31           LOG_AMT_INCOME_TOTAL        20486     1.77551666 1.200116e+01 1.194060e+01 1.193951e+01
32         LOG_CURRENT_AMT_CREDIT         3986     0.34546565 1.078972e+01 1.306442e+01 1.307230e+01
33     LOG_CURRENT_AMT_GOODS_PRICE         5653     0.48994414 1.074095e+01 1.295416e+01 1.296505e+01
34            LOG_AMT_APPLICATION        66145     5.73277114 1.663228e-01 1.103629e+01 1.169733e+01
35         LOG_PREVIOUS_AMT_CREDIT        16935     1.46775235 2.249236e-01 1.156918e+01 1.173817e+01
36    LOG_PREVIOUS_AMT_GOODS_PRICE        23125     2.00423815 1.824523e+00 1.149147e+01 1.168918e+01
```

Log AMT_INCOME_TOTAL



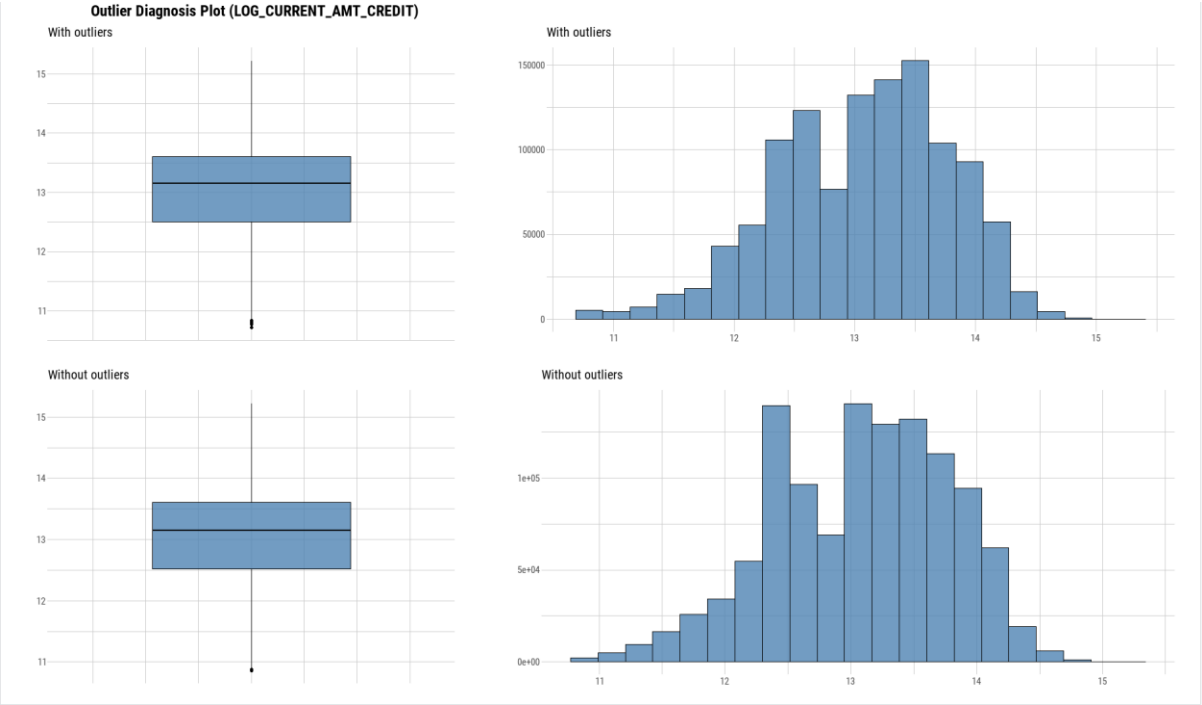
- **Observations**:

  Outliers are present in both the upper and lower ends of the distribution, but the log transformation has compressed the distribution to a manageable scale.

- **Action**:

  Retain the log-transformed column and consider capping or excluding the extreme outliers.

Log CURRENT_AMT_CREDIT



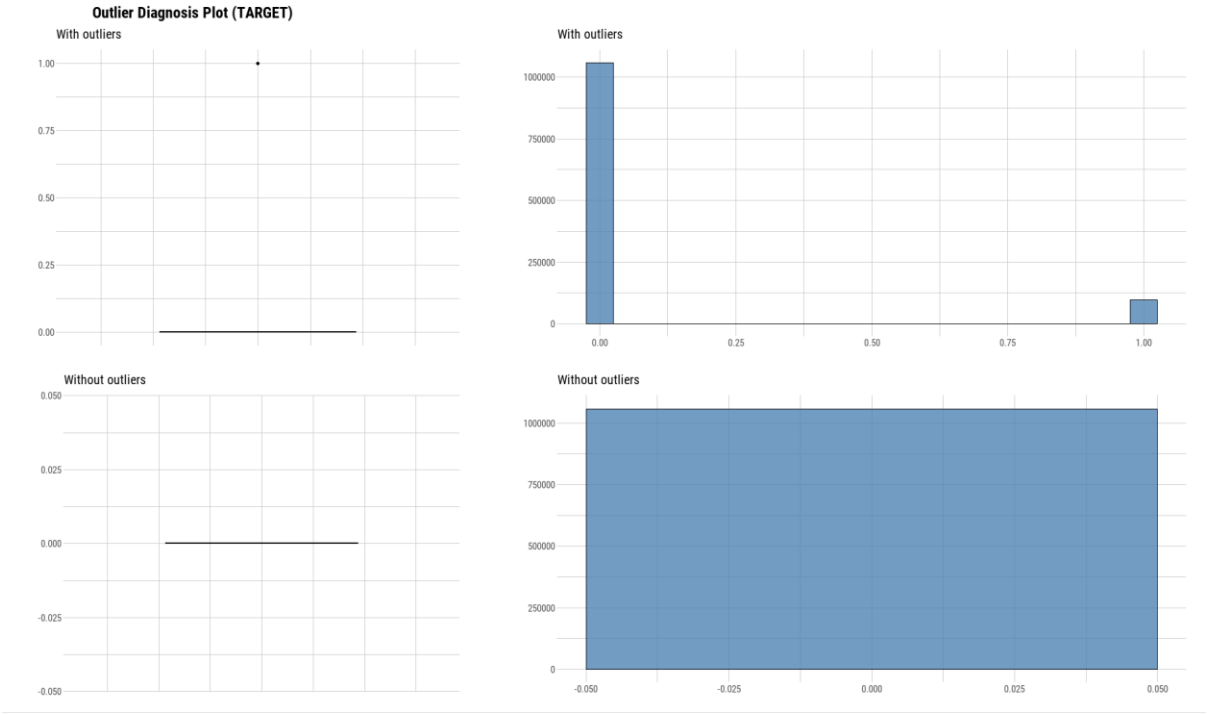
- **Observations**:

  The distribution looks normal after the log transformation, with very few extreme values.

- **Action**:

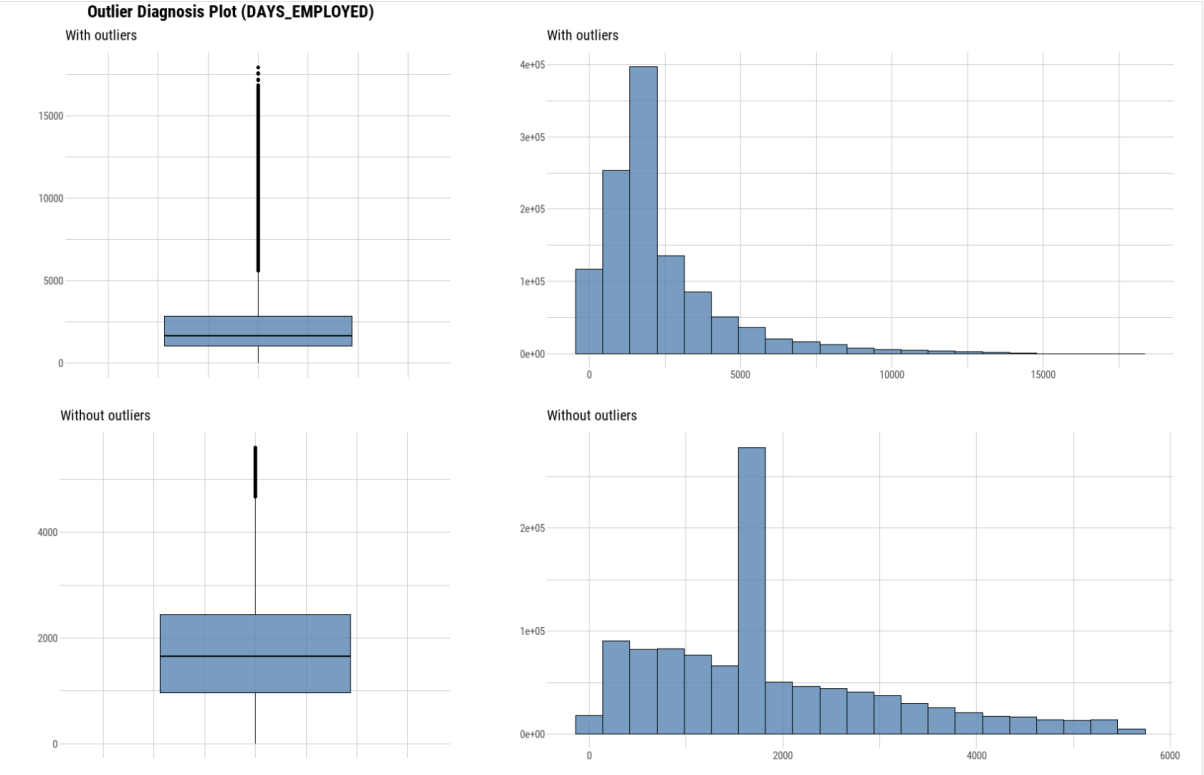
  No further action needed for this column. Retain the log-transformed data.

TARGET



- No further action is needed for this column.

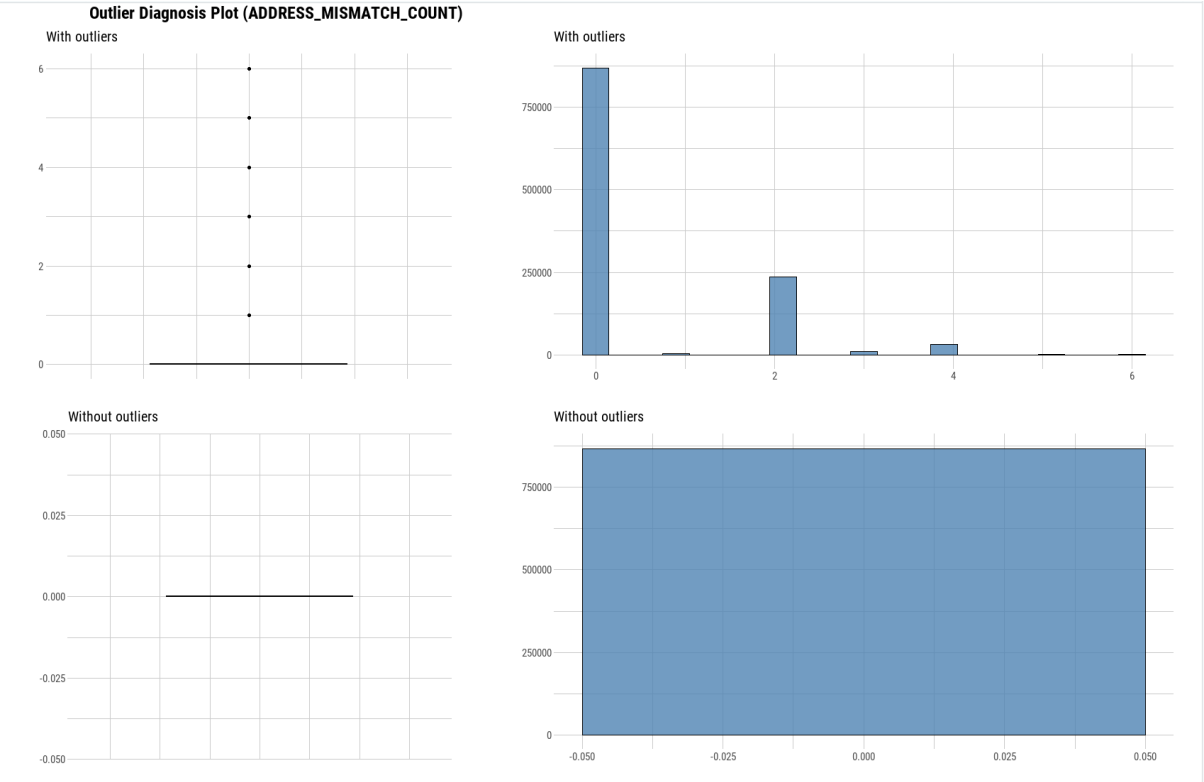
DAYS_EMPLOYED



Observation: There is a long right tail. It could be possible that many applicants have been employed for a long time.

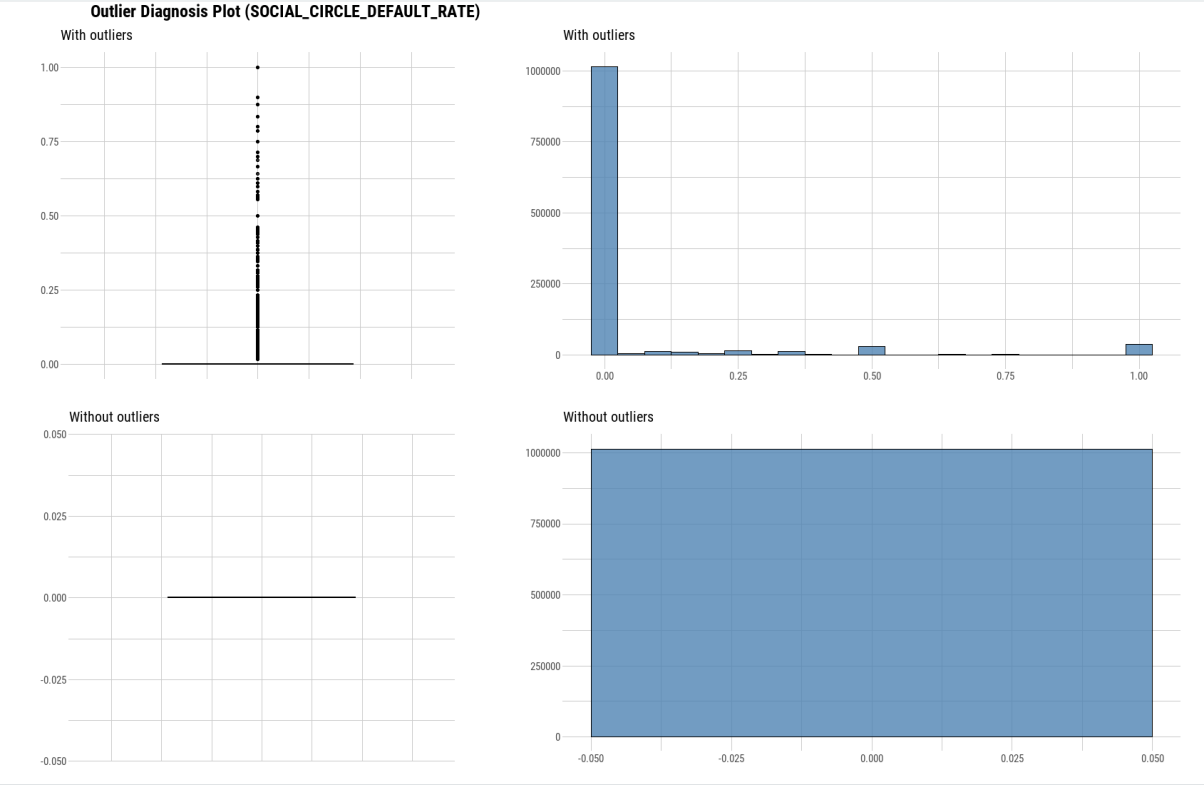
Action: None for now.

ADDRESS_MISMATCH_COUNT



Observation: There don't seem to be any exceptional outliers in this column. Because of a high number of 0s, the model categorized others as outliers.

Action: None



Social Circle Default Rate

- **Observations**:

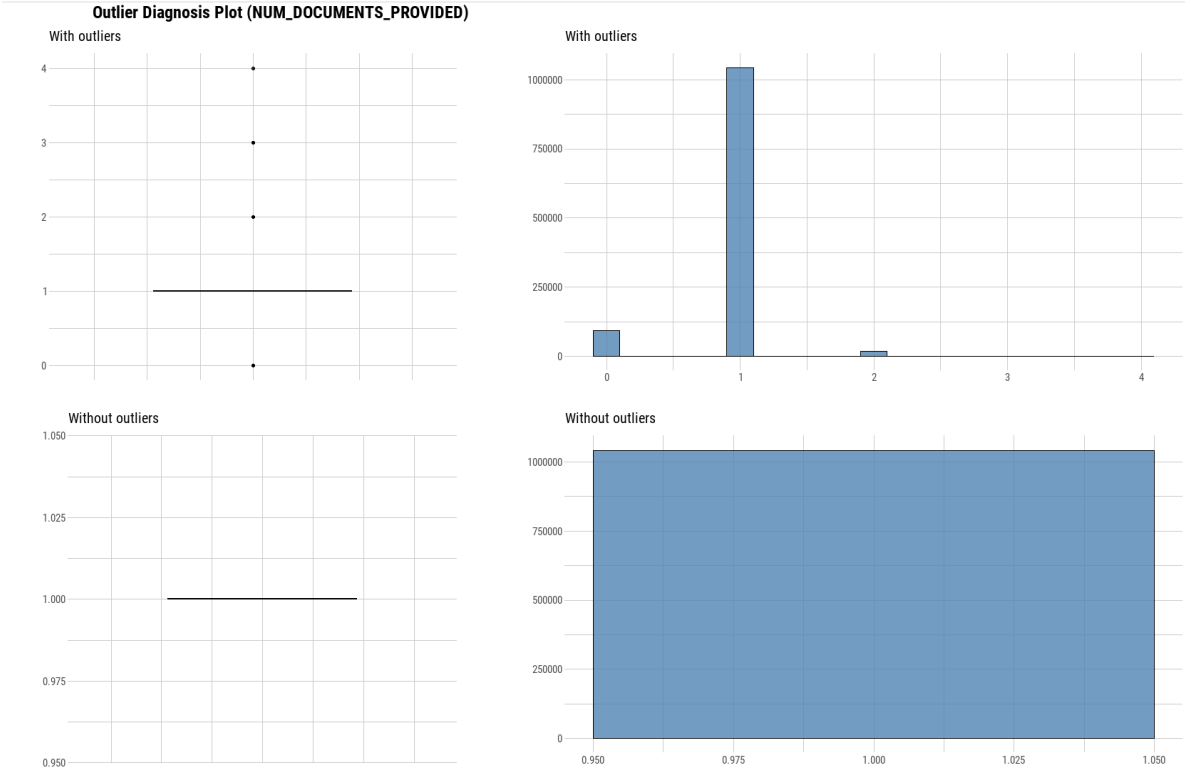
  The distribution is extremely skewed, but the outliers are part of the natural range.

  No major deviation is observed after removing outliers.

- **Action**:

  Retain the column as is without further transformations.

NUM_DOCUMENTS_PROVIDED



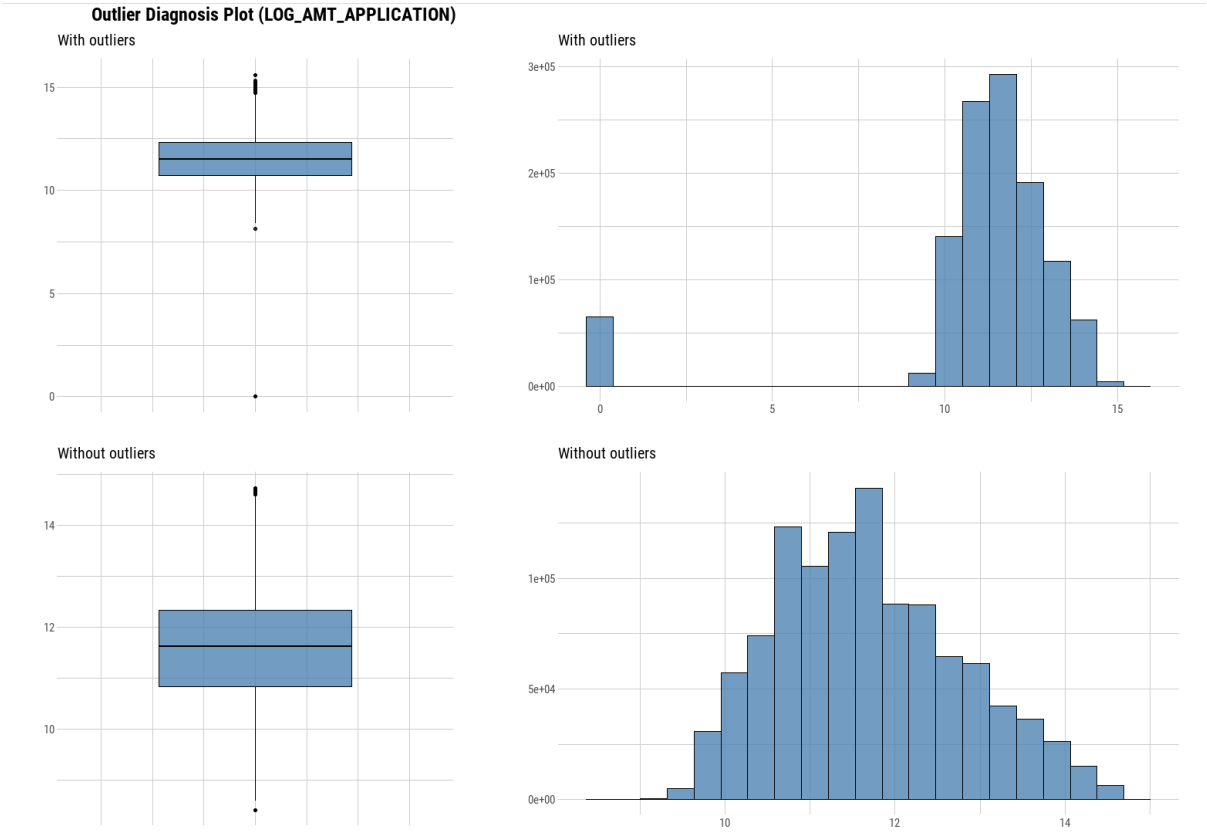
- **Observations**:

  Most of the values fall in a narrow range. The few outliers are a result of multiple documents being provided, which may not need correction.

- **Action**:

  Retain the column without transformations, as the outliers are just naturally occurring.

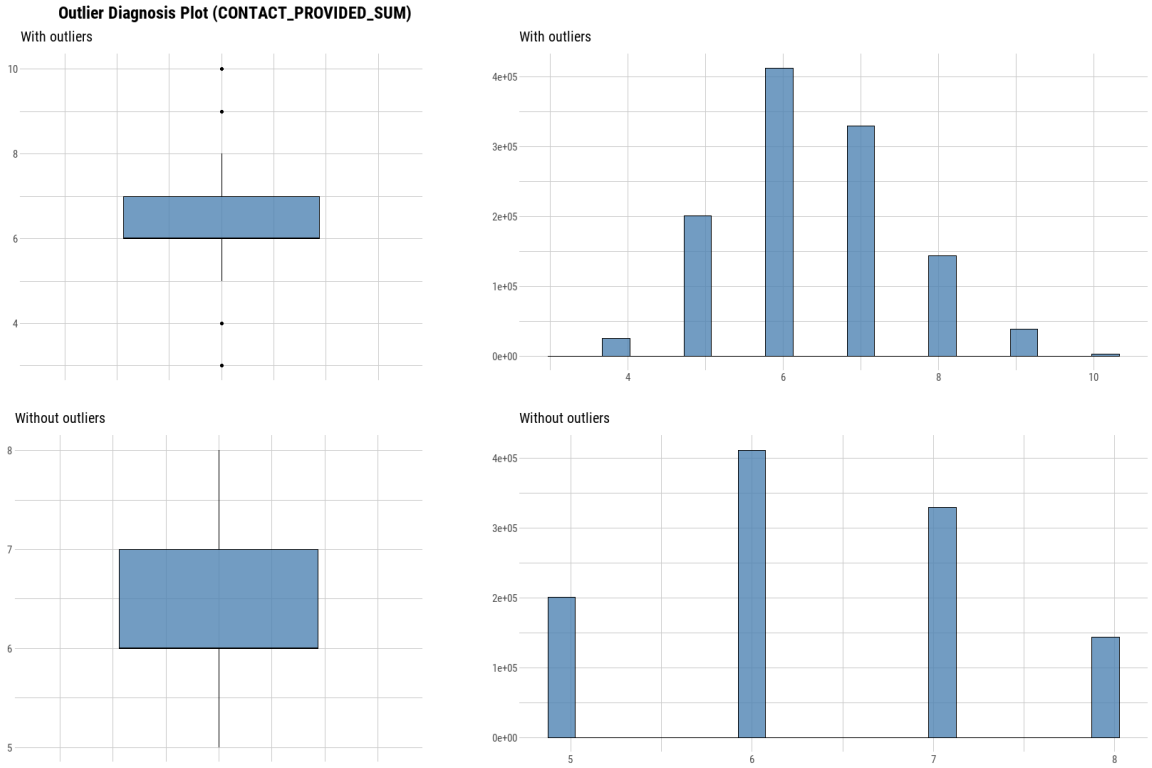
LOG_AMT_APPLICATION



Observation: The log transformation has helped in bringing down the outlier ratio of the column significantly and the data is more interpretable. The remaining outliers are the 0s in the column are a mix of New applications with no previous data and applications where no particular amount was applied but credit was still given.

Action: None.

CONTACT_PROVIDED_SUM



No action is needed as the range is small and there are no evident exceptions

CNT_PAYMENT

- **Observations**:

  Outliers are present on the higher side of the distribution even after log transformation. However, the majority of values are clustered at proper levels.

- **Action**:

  Retain the column as is. We can consider capping the extreme upper outliers if they adversely affect modeling.

Feature Selection

See Appendix for criteria.

Based on the above, we have removed the unnecessary columns to cut down to 45 variables.

```
> str(filtered_data)
'data.frame':   1153805 obs. of  45 variables:
 $ TARGET                           : int  1 0 0 0 0 0 0 0 0 0 ...
 $ CURRENT_NAME_CONTRACT_TYPE       : Factor w/ 2 levels "Cash loans","Revolving loans": 1 1 1 1 2 1 1 1 1 1 ...
 $ CODE_GENDER                      : Factor w/ 3 levels "F","M","XNA": 2 1 1 1 2 1 1 1 1 1 ...
 $ FLAG_OWN_CAR                     : Factor w/ 2 levels "1","2": 1 1 1 1 2 1 1 1 1 1 ...
 $ FLAG_OWN_REALTY                  : Factor w/ 2 levels "1","2": 2 1 1 1 2 2 2 2 2 2 ...
 $ CNT_CHILDREN                     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ AMT_INCOME_TOTAL                 : num  202500 270000 270000 270000 67500 ...
 $ CURRENT_AMT_CREDIT               : num  406598 1293503 1293503 1293503 135000 ...
 $ CURRENT_AMT_ANNUITY              : num  24701 35699 35699 35699 6750 ...
 $ CURRENT_AMT_GOODS_PRICE          : num  351000 1129500 1129500 1129500 135000 ...
 $ NAME_INCOME_TYPE                 : Factor w/ 8 levels "Businessman",..: 8 5 5 5 8 8 8 8 8 8 ...
 $ NAME_EDUCATION_TYPE              : Factor w/ 5 levels "Academic degree",..: 5 2 2 2 5 5 5 5 5 5 ...
 $ NAME_FAMILY_STATUS               : Factor w/ 6 levels "Civil marriage",..: 4 2 2 2 4 1 1 1 1 1 ...
 $ NAME_HOUSING_TYPE                : Factor w/ 6 levels "Co-op apartment",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ REGION_POPULATION_RELATIVE       : num  0.0188 0.00354 0.00354 0.00354 0.01003 ...
 $ DAYS_BIRTH                       : int  9461 16765 16765 16765 19046 19005 19005 19005 19005 19005 ...
 $ DAYS_EMPLOYED                    : int  637 1188 1188 1188 225 3039 3039 3039 3039 3039 ...
 $ OCCUPATION_TYPE                  : Factor w/ 19 levels "","Accountants",..: 10 5 5 5 10 10 10 10 10 10 ...
 $ CNT_FAM_MEMBERS                  : int  1 2 2 2 1 2 2 2 2 2 ...
 $ REGION_RATING_CLIENT             : Factor w/ 3 levels "1","2","3": 2 1 1 1 2 2 2 2 2 2 ...
 $ ORGANIZATION_TYPE                : Factor w/ 58 levels "Advertising",..: 6 40 40 40 12 6 6 6 6 6 ...
 $ EXT_SOURCE_2                     : num  0.263 0.622 0.622 0.622 0.556 ...
 $ EXT_SOURCE_3                     : num  0.139 0.511 0.511 0.511 0.73 ...
 $ ADDRESS_MISMATCH_COUNT           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ WEIGHTED_CREDIT_BUREAU_INQUIRIES: num  0.103 0 0 0 0 ...
 $ SOCIAL_CIRCLE_DEFAULT_RATE       : num  1 0 0 0 0 0 0 0 0 0 ...
 $ NUM_DOCUMENTS_PROVIDED           : int  1 1 1 1 0 1 1 1 1 1 ...
 $ CONTACT_PROVIDED_SUM             : int  7 6 6 6 9 6 6 6 6 6 ...
 $ PREVIOUS_NAME_CONTRACT_TYPE      : Factor w/ 4 levels "","Cash loans",..: 3 2 3 3 3 2 4 2 2 3 ...
 $ PREVIOUS_AMT_ANNUITY             : num  9252 65759 6737 64568 5357 ...
 $ AMT_APPLICATION                  : num  179055 900000 68810 337500 24282 ...
 $ PREVIOUS_AMT_CREDIT              : num  179055 1035882 68054 348638 20106 ...
 $ NAME_CONTRACT_STATUS             : Factor w/ 5 levels "","Approved",..: 2 2 2 2 2 2 2 4 2 2 ...
 $ NAME_CLIENT_TYPE                 : Factor w/ 5 levels "","New","Refreshed",..: 2 4 3 3 2 4 4 4 4 4 ...
 $ NAME_PORTFOLIO                   : Factor w/ 6 levels "","Cards","Cars",..: 5 4 5 5 5 4 2 4 4 5 ...
 $ CHANNEL_TYPE                     : Factor w/ 9 levels "","AP+ (Cash loan)",..: 9 7 6 9 8 7 7 7 7 6 ...
 $ CNT_PAYMENT                      : int  24 12 12 6 4 18 0 48 48 12 ...
 $ NO_PREVIOUS_LOANS                : int  0 0 0 0 0 0 0 0 0 0 ...
 $ LOG_AMT_INCOME_TOTAL             : num  12.2 12.5 12.5 12.5 11.1 ...
 $ LOG_CURRENT_AMT_CREDIT           : num  12.9 14.1 14.1 14.1 11.8 ...
 $ LOG_CURRENT_AMT_GOODS_PRICE      : num  12.8 13.9 13.9 13.9 11.8 ...
 $ LOG_AMT_APPLICATION              : num  12.1 13.7 11.1 12.7 10.1 ...
 $ LOG_PREVIOUS_AMT_CREDIT          : num  12.1 13.85 11.13 12.76 9.91 ...
 $ debt_to_income_ratio             : num  2.01 4.79 4.79 4.79 2 ...
 $ age_in_years                     : num  25.9 45.9 45.9 45.9 52.2 ...
```

As we are going to use Logistic Regression to predict load default, and the model only accepts numeric inputs, we have introduced dummy variables for the factor variables in the dataset. We have used the 'fastdummies' package for this purpose. After the transformation, we are left with 156 variables on the whole.

```
> str(final_data)
'data.frame':                1153805 obs. of  156 variables:
```

See Appendix for snapshot of output.

We have too many variables. For the purpose of developing an efficient model, we are going to determine and eliminate less relevant features using a machine learning technique called Lasso Regression. Lasso regression is commonly used for feature selection because it penalizes the absolute size of the regression coefficients, driving some of them to exactly zero. This results in a sparse model where only the most important features are kept. It helps automatically select a subset of the most relevant features and eliminate irrelevant ones.

Using the 'glmnet' package in R, we ran a cross validated Lasso Regression and found an optimal lambda to identify which features to retain. The below are the extracted lasso coefficients for the optimal lambda.

```
> print(lasso_coefficients)
156 x 1 sparse Matrix of class "dgCMatrix"
```

See Appendix for snapshot.

We have then removed the features with coefficients reduced to 0 and retained non-zero features.

See Appendix for selected features snapshot

Data Splitting

Splitting the Dataset into Training and Testing Sets

Data Splitting

- **Action**: The data was split using the a DataPartition function from **Caret** library into 80% training set and 20% test set.
- **Reason**: A typical split to have an 80–20 split, where it has enough data to train the model but still has enough unseen data to be robust in evaluating.
- **Outcome**:
    - **Training data size**: We verified using nrow(training_data).
    - **Testing data size**: We verified using nrow(testing_data).

```
> summary(logistic_model)

Call:
glm(formula = TARGET ~ ., family = "binomial", data = training_data)

Coefficients:
```

See Appendix for snapshot

Initial Logistic Regression

Action: All variables were used to fit a logistic regression model to the training data.

Model Fit and Significance:

- **Null Deviance:** 529,384
  Contains the extent of deviation for a model without predictors and just the mean response. Lower residual deviance compared to null deviance suggests a better model fit.
- **Residual Deviance:** 474,914
  Measures the model's deviance after fitting predictors. A substantial reduction from the null deviance confirms that predictors are useful.

- **AIC (Akaike Information Criterion):** 475,090
  A measure of the model's goodness of fit that penalizes model complexity. Lower AIC values suggest better models.

Coefficients:

Highly Significant Variables (p-value < 0.001):

- EXT_SOURCE_2 and EXT_SOURCE_3: Extremely significant predictors, with very high z-values, suggesting strong explanatory power for the target.
- DAYS_BIRTH and DAYS_EMPLOYED: Age and employment duration remain critical, with consistently high significance.
- ADDRESS_MISMATCH_COUNT, CNT_FAM_MEMBERS: Relatively strong significance, highlighting their role in prediction.
- Numerous categorical variables like NAME_INCOME_TYPE_Pensioner, NAME_FAMILY_STATUS_Married, and REGION_RATING_CLIENT_2 are significant.

Interpretation of Some Coefficients:

- **EXT_SOURCE_2 (-2.028) and EXT_SOURCE_3 (-2.808):**
  - Negative coefficients indicate higher external credit sources reduce the likelihood of default (lower TARGET).
- **DAYS_EMPLOYED (-6.374e-05):**
  - Longer employment periods reduce default risk, as seen from the negative coefficient.
- **ADDRESS_MISMATCH_COUNT (2.205e-02):**
  - Higher mismatches increase default risk slightly, evidenced by the positive coefficient.
- **Categorical Variables:**

- Categories like CURRENT_NAME_CONTRACT_TYPE_Revolving loans (-7.175e-01) and NAME_INCOME_TYPE_Pensioner (-1.739e-01) offer insights into specific customer segments prone to risk.

Reason: It gave us a baseline model and we could then pinpoint the statistically significant predictors.

Outcome:

- Further details were obtained in the summary(logistic_model) output, which included coefficients and p-values.
- **Significant Variables:** Variables with p-values < 0.05 (except intercept) were selected for further analysis. These included:
  - DAYS_BIRTH, EXT_SOURCE_2, EXT_SOURCE_3, AMT_CREDIT, AMT_ANNUITY
  - Categorical variables like CODE_GENDER, FLAG_OWN_CAR, and others.

```
> significant_vars
```

See Appendix for snapshot

```
> high_cor_pairs
                        row col
LOG_CURRENT_AMT_CREDIT     16  2
LOG_CURRENT_AMT_GOODS_PRICE 17  2
CURRENT_AMT_GOODS_PRICE     2 16
LOG_CURRENT_AMT_GOODS_PRICE 17 16
CURRENT_AMT_GOODS_PRICE     2 17
LOG_CURRENT_AMT_CREDIT     16 17
```

Refining the Dataset

Action:

- gsub was used to remove ugly backticks from variable names, and important variables were retained.
- The correlation matrix for numeric variables was used to assess multicollinearity. Highly correlated pairs (|correlation| > 0.9) were found. For instance:
    - CURRENT_AMT_GOODS_PRICE and LOG_CURRENT_AMT_GOODS_PRICE were highly correlated with AMT_CREDIT.
    - We removed redundant variables for efficiency.

Reason: Multicollinearity reduction makes coefficient estimates stable and increases model interpretability.

Outcome: The dataset was refined containing significant predictors with no multicollinearity issues.

Logistic Regression on Refined Dataset

Action: A logistic regression model was fit on the refined dataset.

Reason: To evaluate the impact of feature selection and multicollinearity reduction on model performance.

Outcome:

- Probabilities for the test data were predicted.
- Using a threshold of 0.5, binary predictions were generated.
- A confusion matrix and accuracy were computed with Accuracy calculated as Correct predictions/Total predictions.

1. Default Threshold (0.5):

- **Accuracy:** 91.69%
- **Sensitivity:** 1.21%
  The model identifies only 1.2% of defaults, which is extremely low and problematic for use cases where identifying defaults is critical.
- **Specificity:** 65.63%
  The model correctly identifies 65.63% of non-defaults, which is decent.

Key Issue: The high accuracy is misleading, as the model is heavily biased toward non-default due to high class imbalance. This threshold, however, is too low in sensitivity for risk management in the real world.

2. Youden's Optimum Threshold (0.078):

Youden's Index is a statistical measure used to evaluate the performance of a diagnostic test, particularly for classification problems. It helps in selecting the optimal cut-off point (threshold) for distinguishing between two groups, such as positive and negative test results.

- **Accuracy:** 65.99%
- **Sensitivity:** 69.92%
  The model captures nearly 70% of defaults, which is a major improvement.
- **Specificity:** 65.63%
  The proportion of correctly classified non-defaults remains the same.

Key Advantage: Because of the much higher sensitivity, the model works better in detecting defaults, which is usually the main objective of credit risk management, albeit at the cost of overall accuracy.

Trade-Offs:

- **Default Threshold:**
  - **Pros:** High overall accuracy.

- **Cons:** It almost never defaults (very high insensitivity), so is not really suitable for applications where you need to identify which of your clients might be risky.

- **Youden's Threshold:**
  - **Pros:** Strikes a balance between sensitivity and specificity, dramatically improving the identification of defaults.
  - **Cons:** The overall accuracy drops due to an increased number of false positives (non-defaults misclassified as defaults). Also, the number of default predictions was low.

Addressing Class Imbalance with Weights

Action:

- Class imbalance was quantified using a weight ratio count out non-defaulters/count of defaulters, which was 11.
- A weighted logistic regression was fit using weights = 11 for defaulters (TARGET = 1) and weights as 1 for non-defaulters.

Reason: Assigning higher weights to the minority class ensures that the model prioritizes learning patterns for defaulters, reducing false negatives.

By using a lot of trial and error for the combination of weightage of minority and threshold for classification, we arrived at weightage of 11 and a threshold of 0.5.

Outcome:

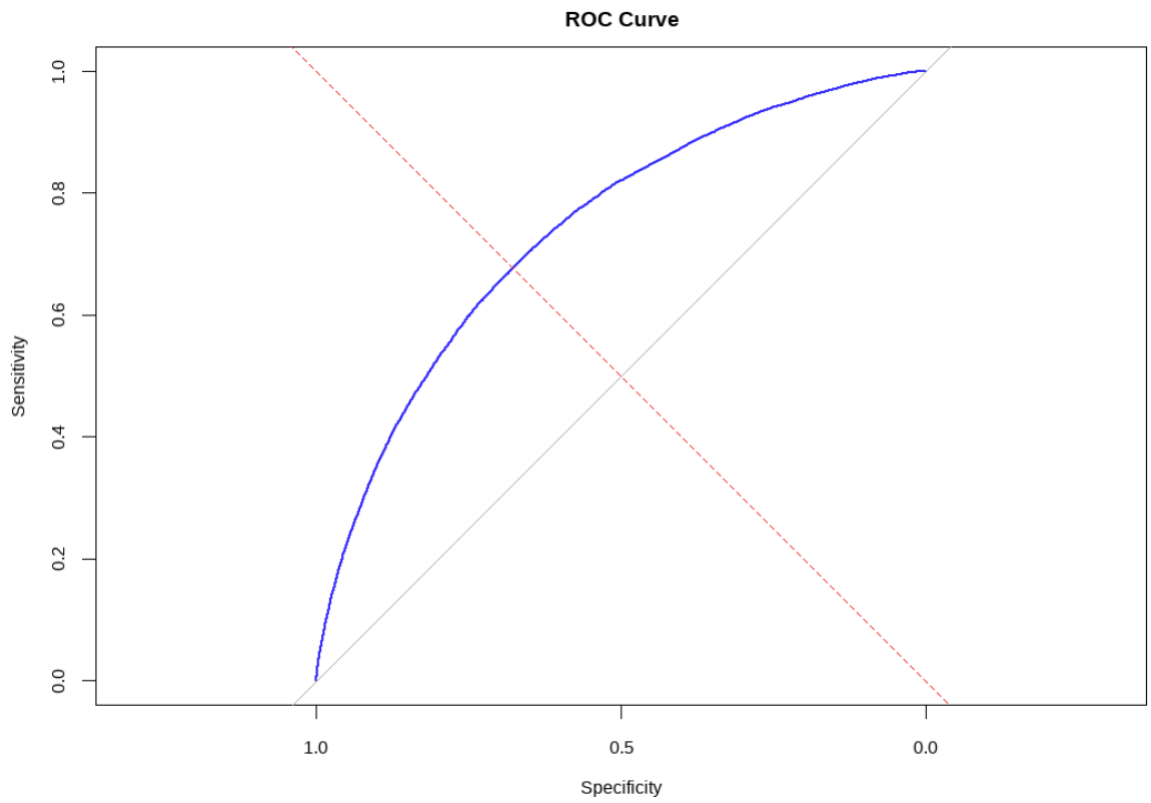
- The weighted model improved sensitivity while maintaining reasonable specificity.
- The final model was evaluated using a threshold of 0.5.

```
> conf_matrix
      Actual
```

```
Predicted     0     1
      0 144486  6350
      1  67002 12923
> # Calculating accuracy
> accuracy <- sum(diag(conf_matrix)) / sum(conf_matrix)
> accuracy
[1] 0.68213
> sensitivity <- conf_matrix[2, 2] / sum(conf_matrix[, 2])
> specificity <- conf_matrix[1, 1] / sum(conf_matrix[, 1])
> sensitivity
[1] 0.6705235
> specificity
[1] 0.6831877
> # Generating ROC curve and calculating AUC
> roc_curve <- roc(testing_data$TARGET, predicted_probs)
Setting levels: control = 0, case = 1
Setting direction: controls < cases
> auc_score <- auc(roc_curve)
> auc_score
Area under the curve: 0.7392
```

Performance Metrics:

- **Accuracy:** 68.21% — It means that this model will likely classify 68% of the cases correctly (21% indicates that).
- **Sensitivity:** 67.05% — The model can correctly identify defaulters (positive cases) in 67% of instances.
- **Specificity:** 68.32% — Shows the model's ability to correctly classify non-defaulters (negative cases).
- **Area Under the Curve (AUC):** 0.7392 — Suggests good discriminatory power in distinguishing between defaulters and non-defaulters.



AUC-ROC: The **AUC of 0.7392** indicates a moderate ability to distinguish between defaulters and non-defaulters.

Overall Evaluation

The model's metrics confirm a reasonable trade-off between detecting defaulters (high sensitivity) and avoiding excessive false positives (moderate specificity). These results align well with practical objectives in financial risk assessment.

Significance of Key Variables

From the model's coefficients and statistical significance:

1. **Highly Influential Predictors (Negative Impact on Default Probability):**

- **EXT_SOURCE_2 (-2.023) and EXT_SOURCE_3 (-2.837):** The two important indicators of high default risk are the external credit scores. Better creditworthiness corresponds to higher values.
- **DAYS_BIRTH (-2.342e-05):** For older applicants, there's less of a chance they'll default.
- **DAYS_EMPLOYED (-6.203e-05):** An increase in the employment duration is associated with a decrease in the default risk.

2. **Key Positive Predictors (Increase Default Risk):**
   - **ADDRESS_MISMATCH_COUNT (+0.027):** The default risk is associated with more address mismatches.
   - **SOCIAL_CIRCLE_DEFAULT_RATE (+0.336):** Higher defaults amongst the applicant's social circle predict higher default probability.
   - **NO_PREVIOUS_LOANS (-0.662):** Default risk increases as a result of a lack of previous loans.
3. **Sociodemographic Factors:**
   - **Gender (Male)** contributes positively to default risk (+0.307).
   - **Unemployment (+1.738)** significantly increases default likelihood.
4. **Loan Characteristics:**
   - **Type of Contract (Revolving loans):** A lower likelihood of default (-0.685) than other contract types.

This model is useful as an initial screening mechanism for the application of loans. To mitigate false positives, users flagged as defaulters should undergo additional manual review or other diagnoses.

Decision Tree Analysis

Objective

These decision tree models are developed and evaluated to classify borrowers into defaulters and non defaulters using a selected subset of the impactful features. This model is designed to help identify the determinants of loan default with fair predictive accuracy, particularly in the prediction of default cases.

Data Preparation

Feature Selection

The highly relevant variables that capture demographic and financial attributes as well as loan-specific attributes were included via a systematic feature selection process. The final features were:

1. **TARGET**: A binary variable (1 if client defaulted; 0 otherwise).
2. **CURRENT_AMT_ANNUITY**: Current loan annuity amount.
3. **DAYS_BIRTH**: Age of client in days.
4. **DAYS_EMPLOYED**: The number of days experienced by a client in employment.
5. **CNT_FAM_MEMBERS**: Depends on relative number of family members dependent on the borrower.
6. **EXT_SOURCE_2 & EXT_SOURCE_3**: Predictors of external credit score.
7. **WEIGHTED_CREDIT_BUREAU_INQUIRIES**: Financial scrutiny in terms of weighted number of credit bureau inquiries.
8. **SOCIAL_CIRCLE_DEFAULT_RATE**: Defaults ratio among group of people from the borrower's social circle.
9. **NUM_DOCUMENTS_PROVIDED**: Number of documents submitted during the application process.

10. **AMT_APPLICATION**: The amount applied for in the loan application.
11. **LOG_AMT_INCOME_TOTAL**: Better scale representation of the income amount so represented as log transformed.
12. **debt_to_income_ratio**: The ratio between the amount of total debt obligations and the income of the borrower.
13. **NO_PREVIOUS_LOANS**: Whether the borrower has taken loans in the past.

This selection takes into account balance between financial stability, credit history and socio definition variables.

Data Transformation

- We converted categorical features into factor variables (i.e. NUM_DOCUMENTS_PROVIDED, NO_PREVIOUS_LOANS, TARGET) but everything else is numeric.
- The training (80%) and testing (20%) subsets were split from the dataset in order to guarantee an unbiased evaluation of model performance.

Model Development

Initial Decision Tree

First, using rpart package we built the model using a decision tree. Key parameters included:

- **Method**: As the target variable was binary, Classification.
- **Complexity Parameter (cp)**: It is set to 0.001 so that the tree will get the details without overfitting.

The resulting tree struggled with making appropriate splits. This is mainly because of the class imbalance in TARGET. So, we needed an alternate approach to employ decision tree.

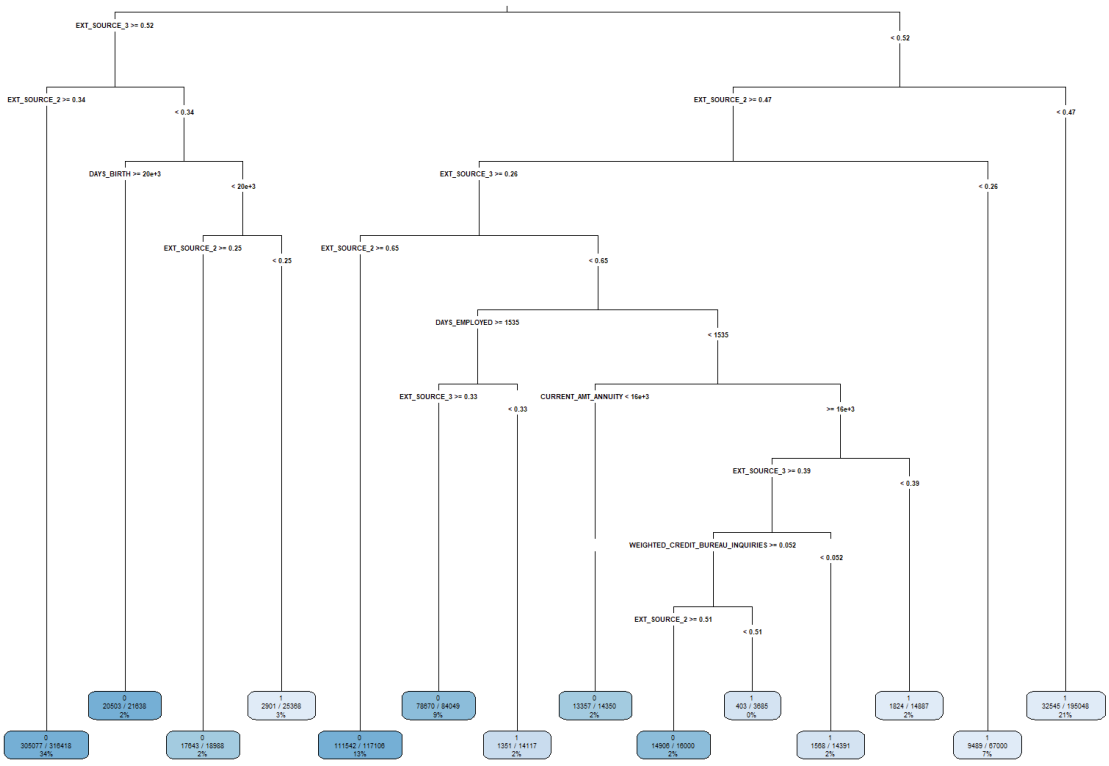
Weighted Decision Tree

To address the imbalance in the dataset (only ~8.3% defaulters), a weighted decision tree was implemented with a custom loss matrix:

- False negatives (defaulters predicted as non-defaulters): **11**
- False positives: **1**

This prioritization ensured the model gave higher importance to identifying defaulters, even at the cost of increased false positives.

Model Visualization:



The rpart.plot package was used to visualize the decision tree. The structure emphasized:

1. **Primary Splits**:

   EXT_SOURCE_2 and EXT_SOURCE_3 dominate the top-level splits:

   In identifying who the borrowers really are, these external credit scores are critical. Generally, higher scores of these indicate a lower risk, and are classified as non defaulters.

The split occurs upon EXTSOURCE_2 < 0.52 at the root node, splitting applicants into higher risk group and lower risk group.

2. **Secondary Factors**:

   DAYS_BIRTH:

   DAYS_BIRTH: >= 20,000 (~55 years old) changed to treat applicants differently, perhaps where older borrowers have different default behavior for whatever reason (e.g., stability, retirement).

   DAYS_EMPLOYED:

   Default risk is affected by employment duration. Borrowers that are employed for more than 1533 days (~4 years) are treated differently and are classified differently as employment stability is considered as a mitigating factor for risk.

   CURRENT_AMT_ANNUITY:

   Low annuity values of less than 16,000 imply a higher probability of default. And perhaps it's a sign of financial strain or reduced borrowing capacity.

3. Interaction Effects

- **Combination of Credit Scores and Demographics**:
  - For instance, age (DAYS_BIRTH) further partitions risky from non-risky applicants further in cases where both EXT_SOURCE_2 and EXT_SOURCE_3 are low, illustrating how this set of factors work together.
- **Credit Bureau Inquiries**:
  - That is, applicants are categorized differently when WEIGHTED_CREDIT_BUREAU_INQUIRIES >= 0.052. This suggests frequent inquiries may signal financial instability or higher credit-seeking behavior.

4. **Leaf Nodes**:

   Nodes with low external credit scores (EXT_SOURCE_2, EXT_SOURCE_3) and short employment durations have the highest default proportion.

   Low-Risk Groups: Other points are, nodes with higher external credit scores and better financial condition will have significantly low default.

Model Evaluation

Confusion Matrix

The weighted decision tree was evaluated using a confusion matrix:

- **True Negatives (TN)**: A total of 140,863 cases are correctly classified as non-defaulters.
- **False Positives (FP)**: This had resulted in incorrectly classifying 70,665 cases as defaulters.
- **True Positives (TP)**: We correctly identified 12,450 defaulters.
- **False Negatives (FN)**: 6,782 defaulters missed.

Key Metrics:

- **Accuracy**: 66.44% - Fraction of correct predictions
- **Sensitivity (Recall)**: 64.74% (proportion of defaulters correctly identified).
- **Specificity**: 66.59% (proportion of non-defaulters correctly identified).
- **Kappa**: 0.1248 (fair agreement between predicted and actual classes).

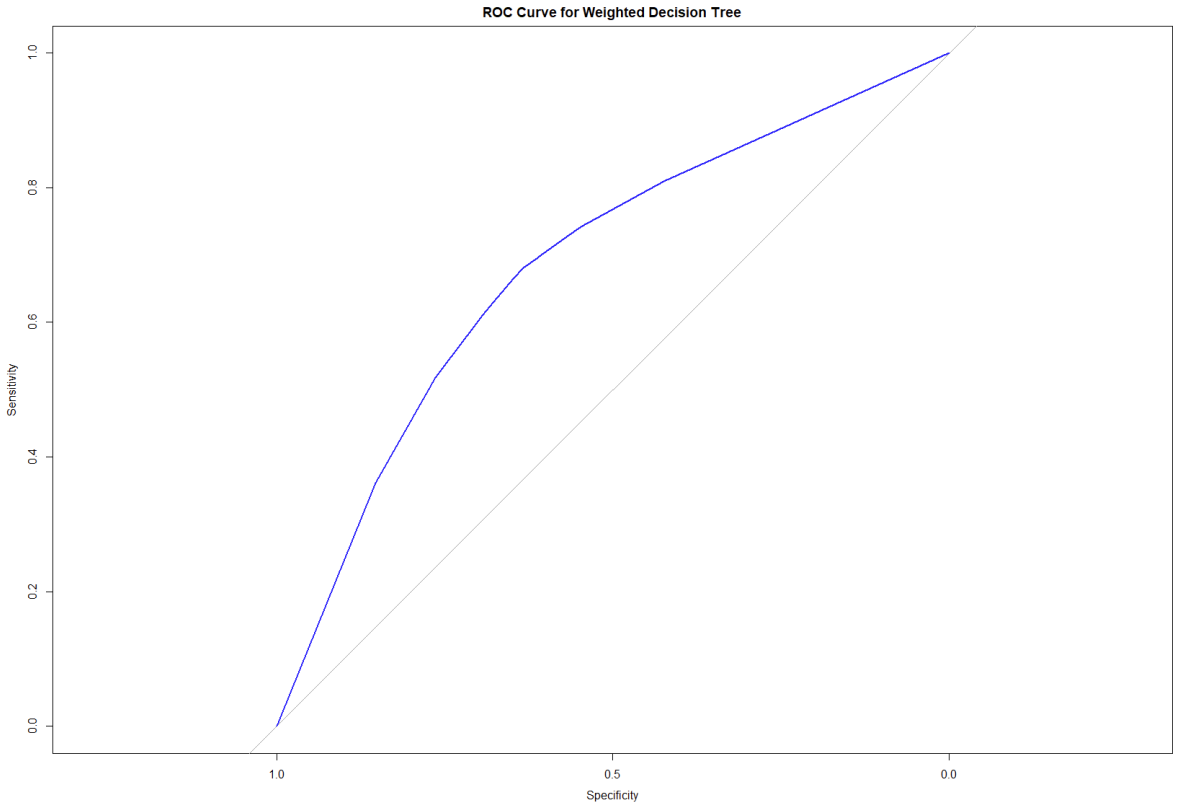
Although this model was not able to differentiate between defaulters and non-defaulters too well, the balanced accuracy of this model was 65.66.

Insights:

1. **Sensitivity vs. Specificity**: Due to higher weight assigned to defaulters, sensitivity was better but false positives increased.

2. **Practical Implications**: False positive may lead to unnecessary interventions, but increasing the number of defaulters corresponds to the project goal of minimizing financial risk.

ROC Curve and AUC Analysis



The model's performance was further evaluated using a Receiver Operating Characteristic (ROC) curve:

- **AUC (Area Under the Curve)**: Measures model's strength in ranking predictions.

The ROC curve of the weighted decision tree showed moderate discrimination of positive (default) as compared to negative (non-default) classes.

The ROC curve for the weighted decision tree highlighted a moderate level of discrimination between the positive (default) and negative (non-default) classes.

Visual Observations:

- The failure of the curve to even match the diagonal is evidence that the model beats random guessing.
- Nevertheless, AUC implies that improvement is possible, especially with respect to sensitivity and specificity trade-off.

Key Takeaways

1. **Feature Importance**: Like prior analyses, the decision tree confirmed that EXT_SOURCE_2 and EXT_SOURCE_3 were the top predictors.
2. **Class Imbalance Handling**: Identification of defaulters has been improved with the weighted decision tree, but specificity remains a challenge to maintain.
3. **Future Recommendations**: Ensemble methods, such as Random Forest or Gradient Boosting, could further enhance model performance. In particular, threshold optimization can also reduce false positives in operational contexts.

Random Forest Model Analysis

1. Here, we implemented a Random Forest model to classify clients as defaulters or non-defaulters on the basis of a number of demographic, financial, and behavioral features.

The choice of the Random Forest model was made in light of the fact that it can handle large data sets, measure the importance of features, and ensure good classifier performances through ensemble learning. This analysis delivers results, metrics, and feature importance on how loan defaults are driven and what our model's predictive capabilities are.

2. Model Development

2.1 Data Configuration

- **Training Dataset Size:** 923,045 observations
- **Testing Dataset Size:** 230,760 observations
- **Number of Features Used:** These include 13 financial ratios, external credit sources, demographic variables, and behavioral factors.

2.2 Model Parameters

- **Number of Trees (num.trees):** 100
- **Importance Measure:** Gini impurity
- **Probability Estimation:** Made probabilistic prediction enabled.
- **Tree Splitting Criterion:** Gini Index
- **Sampling Strategy:** With replacement Random sampling.

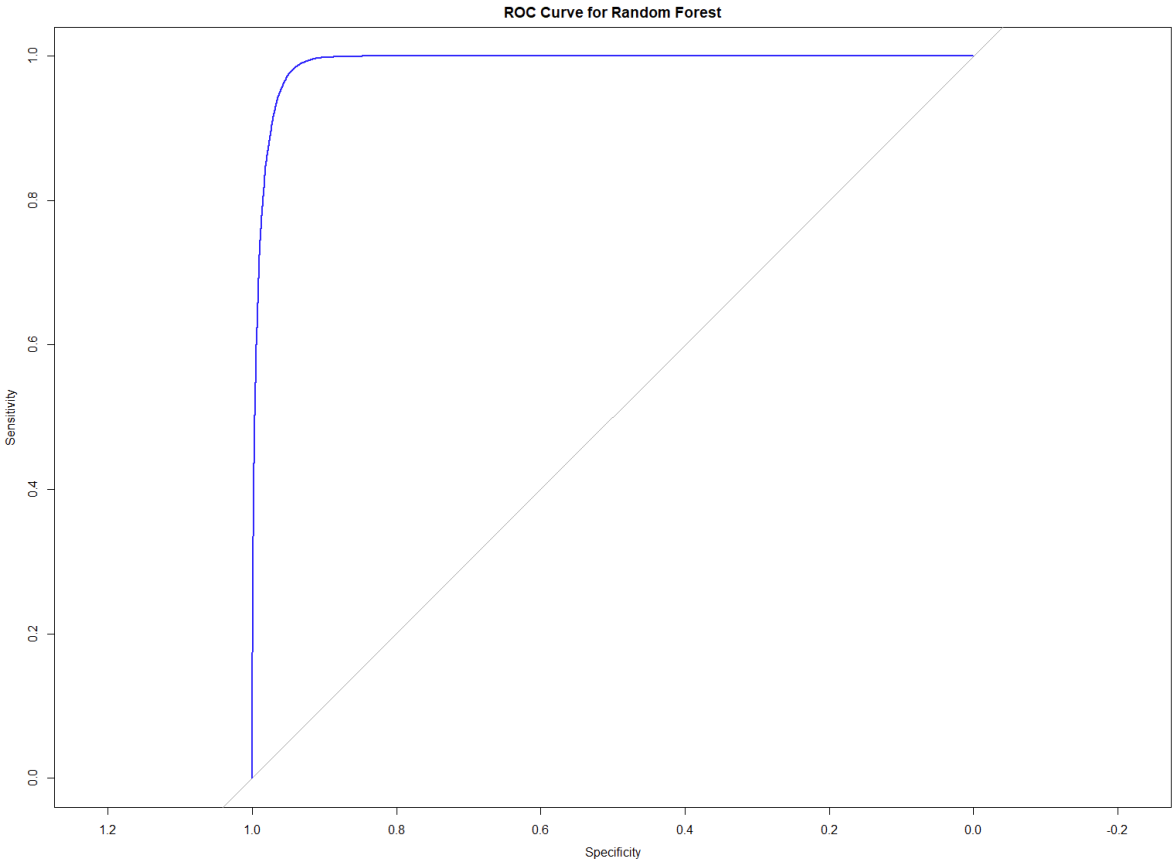
2.3 Implementation Details

- The implementation of the Random Forest was done using the ranger package of the R, which can run fast and efficiently tree construction on large datasets. Based

on this reduction in node impurity (Gini impurity), we extracted feature importance.

3. Model Performance Metrics

3.1 ROC and AUC Analysis



The Receiver Operating Characteristic (ROC) curve was plotted to evaluate the model's performance in distinguishing defaulters from non-defaulters.

- **Area Under the Curve (AUC): 0.9905**
    - **Interpretation:** The AUC value is close to 1, reflecting the model's excellent ability to differentiate between the two classes. This demonstrates that the model performs exceptionally well in predicting loan defaults.

3.2 Confusion Matrix

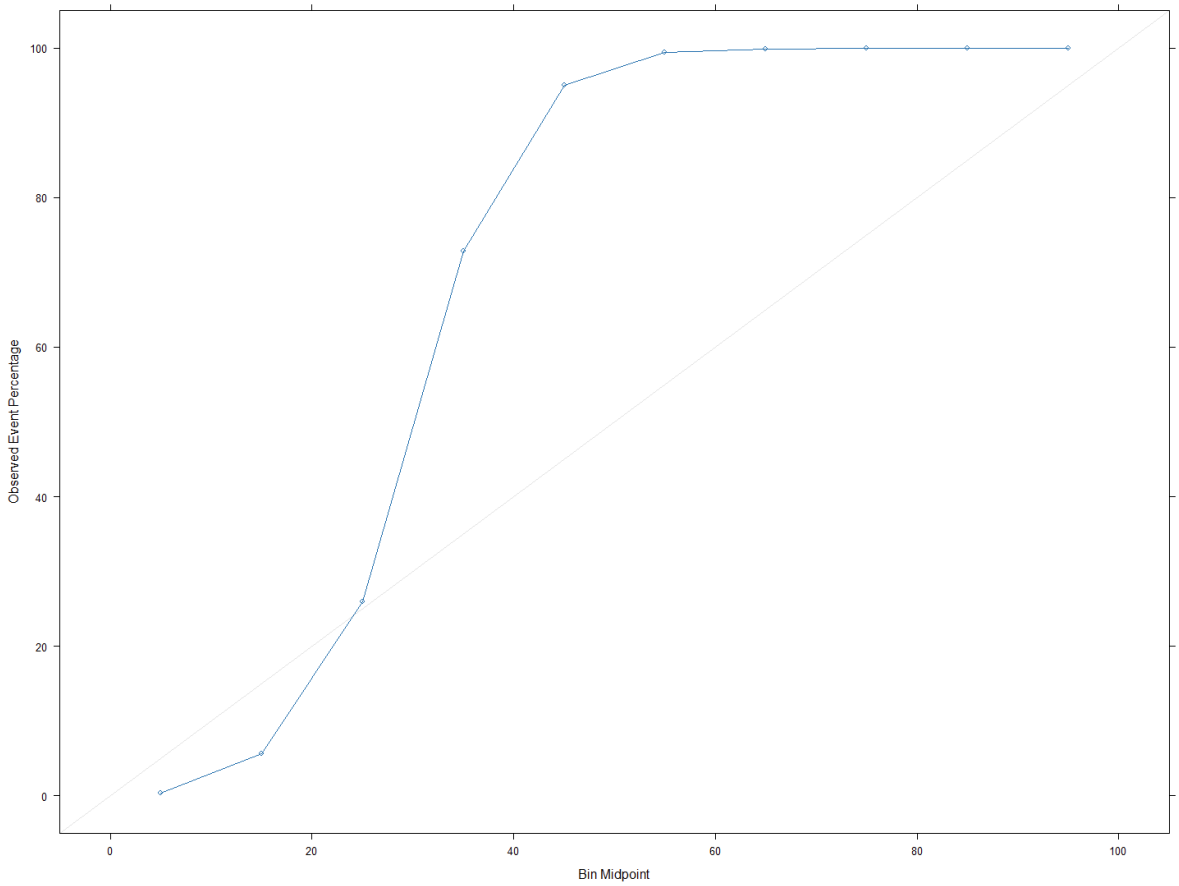
The confusion matrix summarizes the classification results on the test dataset:

	Predicted: Non-Defaulter (0)	**Predicted: Defaulter (1)**
Actual: Non-Defaulter (0)	211,514	14
Actual: Defaulter (1)	4,197	15,035

Metrics Derived from Confusion Matrix:

- **Overall Accuracy: 98.18%**
    - Indicates the proportion of correct predictions (both classes) across the test dataset.
- **Sensitivity (Recall): 78.18%**
    - Reflects the model's ability to correctly identify defaulters.
- **Specificity: 99.99%**
    - Indicates the model's capability to accurately identify non-defaulters.
- **Positive Predictive Value (Precision): 99.91%**
    - Shows the likelihood that a client predicted as a defaulter is indeed a defaulter.
- **Negative Predictive Value: 98.05%**
    - Highlights the likelihood that a client predicted as a non-defaulter is indeed a non-defaulter.
- **Kappa Score: 0.8675**
    - A strong measure of agreement between predicted and actual labels.

3.3 Calibration Analysis



- A comparison was plotted between the predicted probabilities and the default rates, across deciles.
- The result shows that predicted probabilities closely resemble actual outcomes, suggesting that the model is calibrated and reliable in predicting probability.

3.4 Brier Score

- **Brier Score: 0.0162**
    - This is measured as the average squared difference of predicted probabilities minus actual outcomes. High accuracy in probabilistic predictions corresponds to a low value.

4. Feature Importance Analysis

For each feature, Gini impurity reduction was used to derive feature importance.

Feature importance was derived using Gini impurity reduction for each feature. The top predictors of loan default risk are as follows:

Feature	Importance Score
EXT_SOURCE_2	19,194.72
DAYS_BIRTH	15,819.60
EXT_SOURCE_3	15,519.60
Debt-to-Income Ratio	14,888.60
CURRENT_AMT_ANNUITY	14,757.59
DAYS_EMPLOYED	14,206.44
LOG_AMT_INCOME_TOTAL	9,441.15
WEIGHTED_CREDIT_BUREAU_INQUIRIES	8,290.83
AMT_APPLICATION	4,820.08

Insights from Feature Importance:

1. **External Credit Scores (EXT_SOURCE_2 and EXT_SOURCE_3):**
    - Feature importance reveals that external credit data is very relevant in forecasting default risk.
2. **Age (DAYS_BIRTH):**
    - As representative larger DAYS_BIRTH clients will be older (more so) and therefore will have a lower default risk, showing that age has a stabilising effect on financial behaviours.
3. **Debt-to-Income Ratio:**

- This measure of financial health is a strong predictor, indicating that clients with higher ratios are more likely to default.

4. **CURRENT_AMT_ANNUITY:**
    - The default risk is significantly affected by loan-related features, for instance, the annuity amount.
5. **Behavioral Indicators (WEIGHTED_CREDIT_BUREAU_INQUIRIES):**
    - One important behavioral feature associated with default propensity is the number of weighted credit bureau inquiries.

5. Threshold Optimization and Predictions

Binary predictions were taken from probabilities with a threshold set to 0.5. The high predictive power of the model is shown by the confusion matrix, with few false positives and false negatives. By comparing the resulting performance metrics, it is determined that the chosen threshold corresponds well to the dataset.

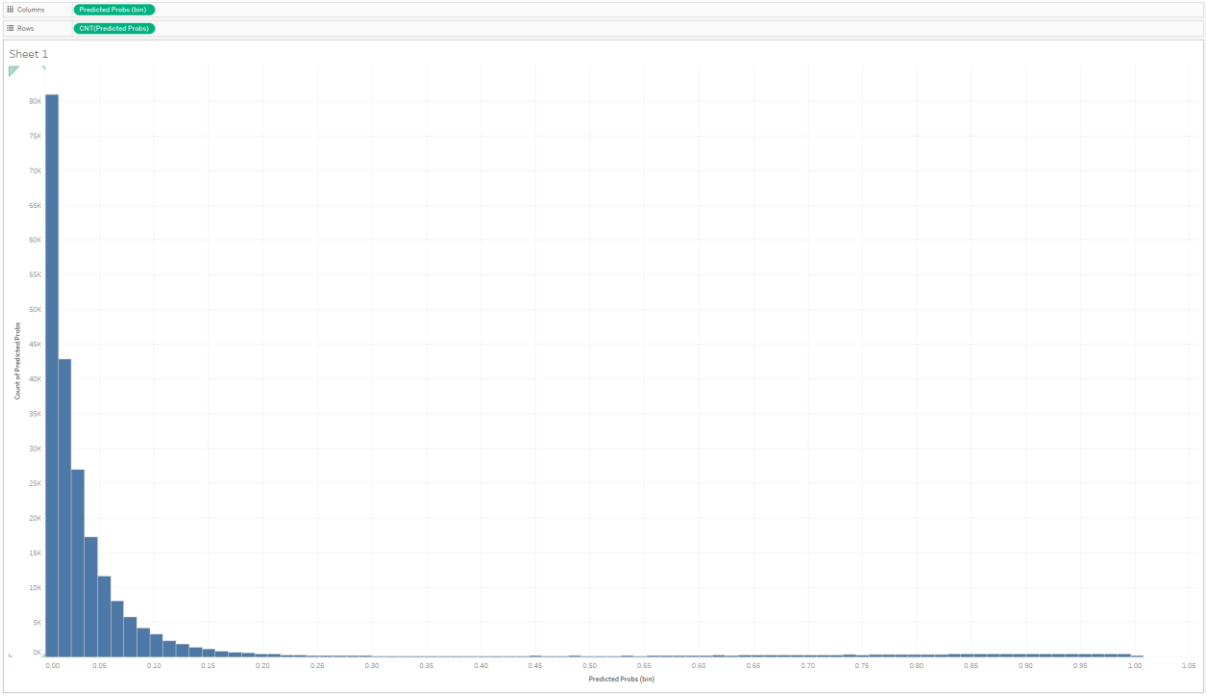
6. Conclusion

The Random Forest model performed very well on loan defaults prediction as shown good AUC, accuracy and well calibrated probabilities. Through this model, critical predictors and hence critical variables including external credit scores, debt-to-income ratio, and demographic variables were identified that would help in risk management strategy. The result of this analysis is the confirmation of Random Forest effectiveness to solve complex classification problems in the financial domain.

Complementary Visualizations

On top of the R data analysis and predictive modeling, Tableau visualizations were created to complement. These visualizations allow for an easy way to look into data trends, distributions and relationships as they exist across the various features. Below are the visualizations generated, along with their interpretations:

Distribution of Predicted Probabilities



Visualization Overview:

This Random Forest Model with only predicting loan default (class 1) produces this histogram showing their distribution of predicted probabilities. This distribution helps demonstrate the model's ability to discriminate high risk versus low risk customers.

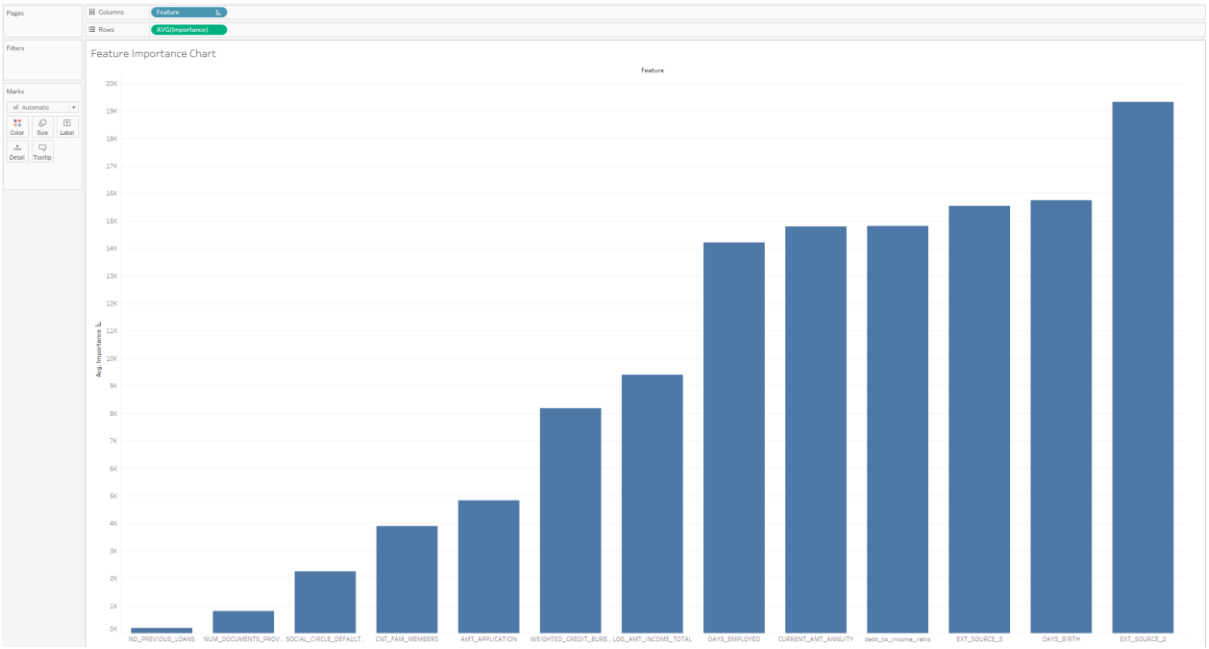
Key Observations:

1. **Concentration at Lower Probabilities:**
    - A large majority of the clients have predicted probabilities near 0, meaning the probability of default is very low.

- This is consistent with the overall dataset's imbalanced nature, with the majority being non-defaulters.

2. **Long Tail Distribution:**
    - As predicted probabilities go up, the count goes down gradually and creates a long tail.
    - The clients with higher predicted probabilities represent high-risk groups, which are critical for loan default mitigation strategies.
3. **Rare High-Risk Clients:**
    - The far-right bins (probabilities close to 1) are quite sparse in terms of number of observations.
    - That is, these clients are most likely to default and are, therefore, the clients that should be most prioritized for intervention strategies.
4. **Model Confidence:**
    - There is a clear separation between low and high probabilities indicating the Random Forest model knows what it is talking about.
    - This distribution highlights the model's ability to differentiate between high-risk and low-risk clients effectively.

Feature Importance Chart

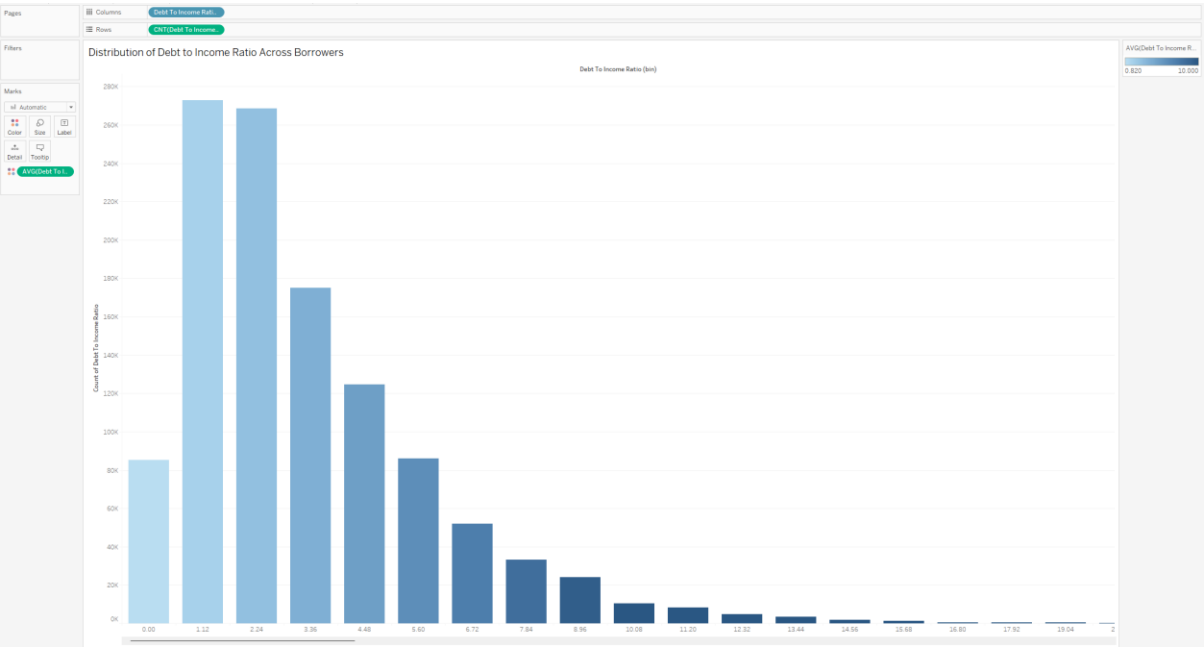


The presented bar chart represents the importance of the features implemented in the Random Forest model for evaluating the risk of loan default.

- The exterior sources 2 and 3 are most important, and the exterior source 2 has a very high degree of importance.
- DAYS_BIRTH (age of the applicant) is the third most important promising variable according to the ranking. In the analysis of the default risk, it should be noted that age can affect financial solvency and loan repayment experience.
- Some of the features, including debt_to_income_ratio, CURRENT_AMT_ANNUITY, LOG_AMT_INCOME_TOTAL, and DAYS_EMPLOYED, are placed of moderate importance. These features track aspects of financial data, loan characteristics, and borrower stability, which are essential considerations for measuring risk.
- It also moderately weights the criterion of WEIGHTED_CREDIT_BUREAU_INQUIRIES and the role of recent activity of a client in the assessment of the risk. While examining the importance score calculated for all the features used in this analysis, the absence of previous loans, namely, the NO_PREVIOUS_LOANS feature, occupies the last position, which means that prior loan history is not regarded as a key determinant of high default risk.

- The chart also shows significance decreasing sharply after the first few features, which implies that very few features are most useful in the model.
- Low-ranking features, although they may not influence the model significantly, may contain useful ancillary information that can be used to improve the accuracy of the model.

Distribution of Debt-to-Income Ratio Chart



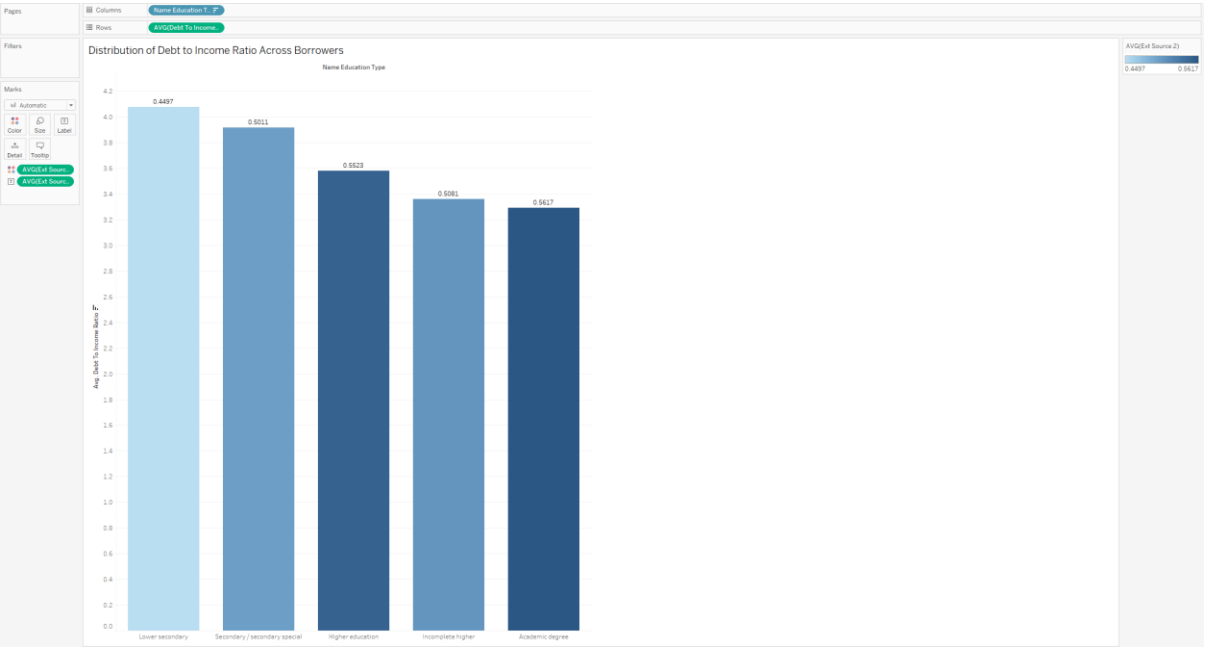
Visualization Overview:

This histogram illustrates the distribution of debt-to-income ratios (DTI) across borrowers. The x-axis represents the debt-to-income ratio (binned into intervals), while the y-axis displays the count of borrowers in each bin.

Key Observations:

1. **Concentration of Borrowers:**
    - The majority of borrowers have a debt-to-income ratio between **1.12 and 4.48**, as evidenced by the tallest bars in this range. This suggests that most borrowers have moderate levels of debt compared to their income.
2. **Tail Distribution:**
    - The frequency decreases sharply for borrowers with a debt-to-income ratio greater than **6.72**. A small portion of borrowers exhibit high debt-to-income ratios, indicating significant financial burden relative to income.
3. **Low Debt-to-Income Ratios:**
    - A notable portion of borrowers also have very low DTI ratios (near or below **1.12**), reflecting strong financial health or minimal debt obligations relative to their income.
4. **Average Debt-to-Income Ratio:**
   The average debt-to-income ratio of the bins with a low number of applicants is very high compared to those with a high number of applicants.

Debt-to-Income Ratio Across Borrowers by Education Level

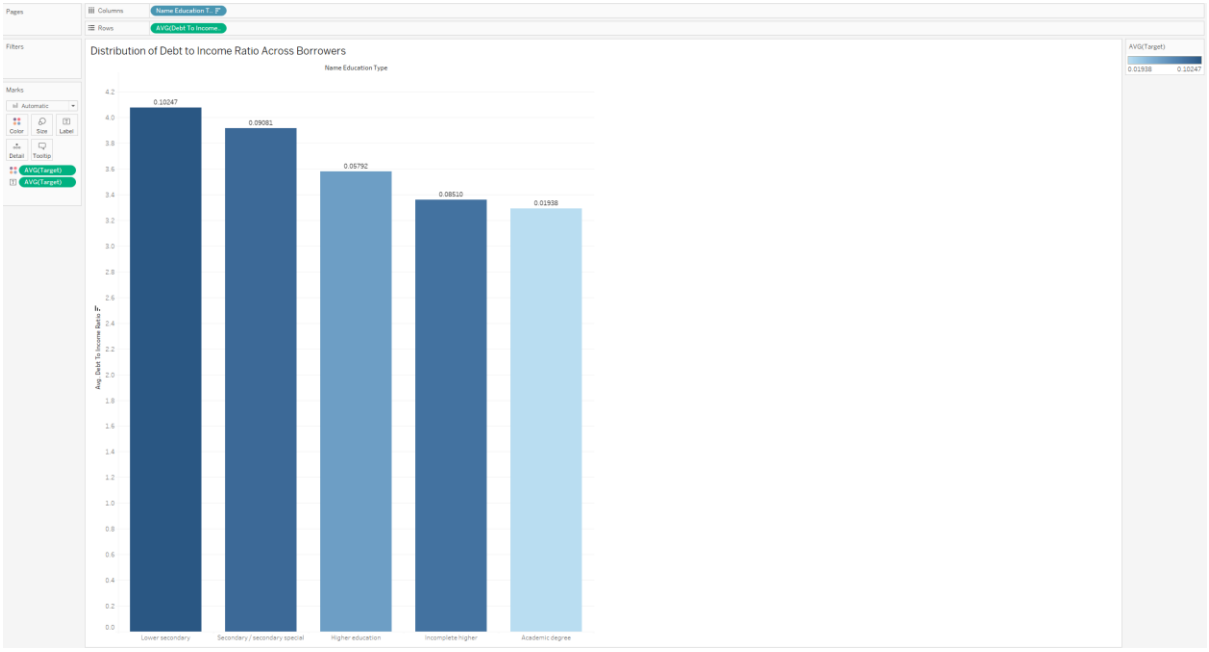


This bar chart illustrates the **average debt-to-income (DTI) ratio** across borrowers segmented by their **education levels**, with an additional **EXT_SOURCE_2 score** as a label and color-coded measure. The **y-axis** represents the DTI ratio, while the **label on each bar** corresponds to the average EXT_SOURCE_2 value.

Key Observations:

- The Average DTIs of Lower Secondary and Secondary groups are highest.
- Group with Academic degree has the lowest average DTI
- The EXT_ SOURCE_2 score can be seen to be the highest for the Academic degree group and lowest for Lower Secondary.

- That means, the lower the education level, the higher the risk for default.



Default Rates by Gender

- The result shows that **male borrowers** default at a higher rate of **10.68%** which means the chances of the individual to have a payment problem is greater than that of **female borrowers**.
- Although the default rate among female borrowers is **7.84 percent**, which is relatively closer to better financial responsibility or **risk management** (**risk management**).
- **Gender-specific credit risk evaluation** is an urgent need to narrow the default rate gap.
- Adapting lending criteria or providing customers with **tailored financial products**, may be desirable for **financial institutions** so as to offset the greater default risk observed in male borrowers.

Comparison of Predicted Probabilities and Default Rates

Comparison Against Defaults Rates and Probabilities Predicted.

- The chart displays a very similar relationship between **predicted probabilities** (**bin ranges**) and **observed default rates**.
- The more the predicted probabilities rise, the more proportionally the actual default rate rises, which confirms the **reliability of the prediction model**.
- The default rates are still remarkably low (e.g. **0.22**, at the range displayed above, of **0.15 → 0.25**) for bins where the predicted probabilities is less than **0.25**.
- These suggest the model identifies **low risk borrowers** appropriately.t Default rates show a steep increase in bins with predicted probabilities over **0.35**.
- For example, at the **0.40–0.45** range the default rate is **42.6%**, and it races up to **100%** beyond **0.85**.
- Bins with probability higher than **0.75** are accurately identified as those with **higher default risks**, where default rate corresponding to such bins is either close to or equaling **100%**.
- The establishment of **threshold criteria** for credit approvals using this visualization.

- For example, **higher-risk borrowers**, determined by a predicted probability > **0.40**, may be processed differently with respect to **lending terms** and / or **additional guarantees**.

V. Conclusion and Recommendations

Conclusion:

This project adequately provides a real-life example of how data analytics and advanced analytics modeling tools may be used in measuring and managing loan default risks in financial organizations. By integrating a wide range of data sources, using SQL for data combination, and R and Tableau for analysis and visualization, we covered a holistic approach to analyzing the causes for loan defaults.
In this perspective, the study showed that credit scores, credit to income and credit to current employment stability are among the most influential variables that inform these defaults. The logistic regression, decision tree as well as random forest models all had a high accuracy value, which is evidence for the suitability of machine learning approaches in financial risk assessment. In addition, Tableau insights gave specific details on borrowers' details, spending habits, and defaults which enriched the decision-making process.

The identified results show the effectiveness of data-driven solutions in changing the approaches to loan approval, advancing the financial risks management as well as making the lending process more efficient and fair.

Recommendations:

1. **Refine Loan Risk Assessment Criteria:** External credit scores, together with other important loan features such as the debt-to-income ratio and employment stability, should remain core factors in loan sanctioning and risk assessment by financial institutions. This will improve the opportunity to recognize such borrowers at the initial stage of credit delivery.
2. **Use Advanced Predictive Analytics:** To increase the chance and stability of the model across multiple segments and domains, ensemble models such as **random forest**, **gradient boosting**, and other equivalent **state-of-the-art machine learning** techniques should be utilized.

3. **Improve Data Management Systems:** Enhance **SQL** and other large-scale database system software as tools for handling large and dynamic datasets. Self-updating procedures for data integration will further improve operational efficiency.

4. **Develop Tailored Borrower Strategies:**
   Advanced borrower segmentation is critical for developing tailored solutions. For example, borrowers at high risk of default can receive **personalized financial advice**, revised loan terms, or additional incentives to reduce risk.

5. **Enhance Data Transparency and Completeness:**
   Ensure applicants provide **accurate and complete information** during loan application procedures. This includes identity details, such as phone numbers and addresses, income documentation, and credit scores, to enhance the reliability of risk assessments.

6. **Incorporate Dynamic Decision Dashboards:**
   Present risk statistics, borrower characteristics, and performance metrics using **real-time Tableau data visualization** widgets. These tools enable decision-makers to focus on critical areas and make data-supported decisions.

7. **Monitor and Update Predictive Models Regularly:**
   Ensure continuous monitoring of model performance to account for changes in borrower behavior, economic conditions, and regulatory requirements. Regular updates with new data will maintain model relevance and accuracy.

8. **Adopt Proactive Risk Mitigation Policies:**
   Introduce **early alerts** to identify potential high-risk customers. Proactive measures, such as early communication or adjustments to loan terms, can reduce default probabilities.

By implementing these recommendations, financial institutions can improve credit risk management, enhance customer trust, and increase operational efficiency. The integration of **data analytics** into risk assessment workflows ensures a forward-thinking, scalable, and impactful approach to addressing the challenges of loan default prediction.

VI. Limitations

Despite the substantial progress made in loan risk analysis and predictive modeling, several issues emerged that must be addressed:

1. **Limited Computational Resources:**
   A single laptop with limited computational power was used for this project. This constraint was evident in the processing speed for analyzing large datasets and in resource-intensive tasks like developing **Random Forest** and **Decision Tree** models.

2. **Large Dataset Complexity:**
   The project included a dataset with over a million observations and numerous features, making it significantly larger than typical datasets. Challenges included **memory usage**, computational time, and feature selection within constrained resources.

3. **Class Imbalance:**
   The target variable, loan default, exhibited a severe **class imbalance**, with non-default cases overwhelmingly outnumbering defaults. While weighting techniques and thresholds helped mitigate this issue, they may still have impacted the model's ability to accurately predict the minority class (default cases).

4. **Time Constraints:**
   Time limitations affected the granularity of **exploratory data analysis**, **model tuning**, and **hyperparameter optimization**. Additional time could have improved model performance and allowed for more comprehensive evaluations.

5. **Simplified Feature Engineering:**
   Although substantial effort was invested in feature selection, other potentially impactful feature transformations or combinations were not explored due to time and computational limitations. Advanced **domain-specific feature engineering** could have further enhanced results.

6. **Dependence on Imputation Techniques:**
   Median and mode imputation were used to address missing values. While effective, these techniques may have introduced biases, potentially skewing the dataset's distribution and affecting model accuracy.

7. **Static Analysis:**
   The dataset captured borrower attributes at a single point in time, omitting dynamic or temporal factors that might influence default risk. This limited the study's ability to assess how changes over time affect default probabilities.

8. **Practical Deployment Considerations:**
   Although the models were evaluated for accuracy, additional work is needed to test their **real-world applicability**, including integration into production environments, ensuring fairness, interpretability, and compliance with regulatory requirements.

These limitations highlight the challenges of working with large datasets and computationally intensive models in resource-constrained environments. While these obstacles may have restricted the generalization of findings, they do not diminish the project's ability to demonstrate the applicability of **predictive analytics** for loan default risk assessment. Moreover, this work lays a foundation for further empirical studies and real-world implementations.

Appendix

Feature selection criteria

Column Name	Decision	Reasoning
SK_ID_CURR	Remove	Identifier column, irrelevant for modeling.
TARGET	Keep	Target variable for prediction.
CURRENT_NAME_CONTRACT_TYPE	Keep	Loan type (Cash or Revolving); potentially relevant to default behavior.
CODE_GENDER	Keep	Gender may correlate with financial behavior.
FLAG_OWN_CAR	Keep	Indicates car ownership, potentially linked to income level or financial stability.
FLAG_OWN_REALTY	Keep	Indicates real estate ownership, relevant for financial stability.
CNT_CHILDREN	Keep	Number of children, may affect income burden and default risk.
AMT_INCOME_TOTAL	Keep	Income is a key determinant of loan repayment ability.
CURRENT_AMT_CREDIT	Keep	Loan amount, critical for understanding the financial obligation.
CURRENT_AMT_ANNUITY	Keep	Annuity amount indicates repayment burden.
CURRENT_AMT_GOODS_PRICE	Keep	Price of goods purchased, relevant for understanding loan use.
CURRENT_NAME_TYPE_SUITE	Remove	Who the loan is for (e.g., family, partner) is unlikely to significantly affect outcomes.
NAME_INCOME_TYPE	Keep	Employment status (e.g., working, pensioner) correlates with financial behavior.

Column Name	Decision	Reasoning
NAME_EDUCATION_TYPE	Keep	Education level often correlates with income and financial responsibility.
NAME_FAMILY_STATUS	Keep	Family status (e.g., married, single) may influence financial behavior.
NAME_HOUSING_TYPE	Keep	Housing type may indicate financial stability (e.g., owned vs. rented).
REGION_POPULATION_RELATIVE	Keep	Reflects the socio-economic environment, relevant for defaults.
DAYS_BIRTH	Keep	Age (derived from days_birth) is a critical demographic factor.
DAYS_EMPLOYED	Keep	Employment duration indicates job stability, relevant to risk.
DAYS_REGISTRATION	Remove	Likely redundant with other features, low relevance for default prediction.
DAYS_ID_PUBLISH	Remove	Duration since ID was published is unlikely to influence defaults.
OCCUPATION_TYPE	Keep	Occupation type can correlate with income and financial behavior.
CNT_FAM_MEMBERS	Keep	Family size may affect financial burdens and repayment ability.
REGION_RATING_CLIENT	Keep	Regional ratings provide contextual socio-economic data.
REGION_RATING_CLIENT_W_CITY	Remove	Redundant with REGION_RATING_CLIENT.
ORGANIZATION_TYPE	Keep	Employer type reflects financial stability.
EXT_SOURCE_2	Keep	External credit risk score, highly predictive.
EXT_SOURCE_3	Keep	Another external risk score, highly predictive.
DAYS_LAST_PHONE_CHANGE	Remove	Unlikely to influence defaults directly.
ADDRESS_MISMATCH_COUNT	Keep	Indicates address inconsistencies, relevant for stability.

Column Name	Decision	Reasoning
WEIGHTED_CREDIT_BUREAU_INQUIRIES	Keep	Reflects borrowing behavior and credit inquiry history.
SOCIAL_CIRCLE_DEFAULT_RATE	Keep	Indicates default rates within the borrower's social circle.
NUM_DOCUMENTS_PROVIDED	Keep	Number of documents provided may reflect borrower transparency.
CONTACT_PROVIDED_SUM	Keep	Number of contacts provided may indicate borrower effort.
SK_ID_PREV	Remove	Identifier for previous loans, irrelevant for modeling.
PREVIOUS_NAME_CONTRACT_TYPE	Keep	Type of previous contracts may correlate with financial behavior.
PREVIOUS_AMT_ANNUITY	Keep	Previous loan annuities, relevant to repayment history.
AMT_APPLICATION	Keep	Loan application amount, relevant for financial obligation.
PREVIOUS_AMT_CREDIT	Keep	Previous loan amounts are relevant to repayment history.
PREVIOUS_AMT_GOODS_PRICE	Remove	Redundant with PREVIOUS_AMT_CREDIT or CURRENT_AMT_GOODS_PRICE.
WEEKDAY_APPR_PROCESS_START	Remove	Day of the week is unlikely to influence defaults.
HOUR_APPR_PROCESS_START	Remove	Hour of application is unlikely to influence defaults.
FLAG_LAST_APPL_PER_CONTRACT	Remove	Unlikely to influence defaults.
NFLAG_LAST_APPL_IN_DAY	Remove	Unlikely to influence defaults.
NAME_CASH_LOAN_PURPOSE	Remove	Purpose of the loan is ambiguous and often non-predictive.
NAME_CONTRACT_STATUS	Keep	Status of the contract indicates borrower behavior.
DAYS_DECISION	Remove	Likely redundant with other features.

Column Name	Decision	Reasoning
NAME_PAYMENT_TYPE	Remove	Unlikely to significantly impact default prediction.
CODE_REJECT_REASON	Remove	Unlikely to significantly impact default prediction.
PREVIOUS_NAME_TYPE_SUITE	Remove	Low relevance to prediction.
NAME_CLIENT_TYPE	Keep	Indicates whether the client is new, repeated, or refreshed.
NAME_GOODS_CATEGORY	Remove	Too granular and likely redundant with loan details.
NAME_PORTFOLIO	Keep	Loan portfolio may influence risk profile.
NAME_PRODUCT_TYPE	Remove	Unlikely to significantly impact prediction.
CHANNEL_TYPE	Keep	Loan distribution channel may influence defaults.
NAME_SELLER_INDUSTRY	Remove	Low relevance for prediction.
CNT_PAYMENT	Keep	Number of payments made or scheduled is relevant.
NAME_YIELD_GROUP	Remove	Ambiguous impact on defaults.
PRODUCT_COMBINATION	Remove	Redundant with loan details.
NO_PREVIOUS_LOANS	Keep	Indicates whether the borrower is a new or existing client.
LOG_AMT_INCOME_TOTAL	Keep	Log-transformed feature to reduce skewness.
LOG_CURRENT_AMT_CREDIT	Keep	Log-transformed feature to reduce skewness.
LOG_CURRENT_AMT_GOODS_PRICE	Keep	Log-transformed feature to reduce skewness.
LOG_AMT_APPLICATION	Keep	Log-transformed feature to reduce skewness.
LOG_PREVIOUS_AMT_CREDIT	Keep	Log-transformed feature to reduce skewness.

Column Name	Decision	Reasoning
LOG_PREVIOUS_AMT_GOODS_PRICE	Remove	Redundant with other financial features.
debt_to_income_ratio	Keep	Engineered feature relevant to risk.
age_in_years	Keep	Engineered feature relevant to risk.

```
> str(final_data)
'data.frame':	1153805 obs. of  156 variables:
 $ CNT_CHILDREN                          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ AMT_INCOME_TOTAL                      : num  202500 270000 270000 270000 67500 ...
 $ CURRENT_AMT_CREDIT                    : num  406598 1293503 1293503 1293503 135000 ...
 $ CURRENT_AMT_ANNUITY                   : num  24701 35699 35699 35699 6750 ...
 $ CURRENT_AMT_GOODS_PRICE               : num  351000 1129500 1129500 1129500 135000 ...
 $ REGION_POPULATION_RELATIVE            : num  0.0188 0.00354 0.00354 0.00354 0.01003 ...
 $ DAYS_BIRTH                            : int  9461 16765 16765 16765 19046 19005 19005 19005 19005 19005 ...
 $ DAYS_EMPLOYED                         : int  637 1188 1188 1188 225 3039 3039 3039 3039 3039 ...
 $ CNT_FAM_MEMBERS                       : int  1 2 2 2 1 2 2 2 2 2 ...
 $ EXT_SOURCE_2                          : num  0.263 0.622 0.622 0.622 0.556 ...
 $ EXT_SOURCE_3                          : num  0.139 0.511 0.511 0.511 0.73 ...
 $ ADDRESS_MISMATCH_COUNT                : int  0 0 0 0 0 0 0 0 0 0 ...
 $ WEIGHTED_CREDIT_BUREAU_INQUIRIES      : num  0.103 0 0 0 0 ...
 $ SOCIAL_CIRCLE_DEFAULT_RATE            : num  1 0 0 0 0 0 0 0 0 0 ...
 $ NUM_DOCUMENTS_PROVIDED                : int  1 1 1 1 0 1 1 1 1 1 ...
 $ CONTACT_PROVIDED_SUM                  : int  7 6 6 6 9 6 6 6 6 6 ...
 $ PREVIOUS_AMT_ANNUITY                  : num  9252 65759 6737 64568 5357 ...
 $ AMT_APPLICATION                       : num  179055 900000 68810 337500 24282 ...
 $ PREVIOUS_AMT_CREDIT                   : num  179055 1035882 68054 348638 20106 ...
 $ CNT_PAYMENT                           : int  24 12 12 6 4 18 0 48 48 12 ...
 $ LOG_AMT_INCOME_TOTAL                  : num  12.2 12.5 12.5 12.5 11.1 ...
 $ LOG_CURRENT_AMT_CREDIT                : num  12.9 14.1 14.1 14.1 11.8 ...
 $ LOG_CURRENT_AMT_GOODS_PRICE           : num  12.8 13.9 13.9 13.9 11.8 ...
 $ LOG_AMT_APPLICATION                   : num  12.1 13.7 11.1 12.7 10.1 ...
 $ LOG_PREVIOUS_AMT_CREDIT               : num  12.1 13.85 11.13 12.76 9.91 ...
 $ debt_to_income_ratio                  : num  2.01 4.79 4.79 4.79 2 ...
 $ age_in_years                          : num  25.9 45.9 45.9 45.9 52.2 ...
 $ NO_PREVIOUS_LOANS                     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ CURRENT_NAME_CONTRACT_TYPE_Revolving loans : int  0 0 0 0 1 0 0 0 0 0 ...
 $ CODE_GENDER_M                         : int  1 0 0 0 1 0 0 0 0 0 ...
 $ CODE_GENDER_XNA                       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_OWN_CAR_2                        : int  0 0 0 0 1 0 0 0 0 0 ...
 $ FLAG_OWN_REALTY_2                     : int  1 0 0 0 1 1 1 1 1 1 ...
 $ NAME_INCOME_TYPE_Commercial associate : int  0 0 0 0 0 0 0 0 0 0 ...
 $ NAME_INCOME_TYPE_Maternity leave      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ NAME_INCOME_TYPE_Pensioner            : int  0 0 0 0 0 0 0 0 0 0 ...
```

$ NAME_INCOME_TYPE_State servant : int 0 1 1 1 0 0 0 0 0 0 ...
$ NAME_INCOME_TYPE_Student : int 0 0 0 0 0 0 0 0 0 0 ...
$ NAME_INCOME_TYPE_Unemployed : int 0 0 0 0 0 0 0 0 0 0 ...
$ NAME_INCOME_TYPE_Working : int 1 0 0 0 1 1 1 1 1 1 ...
$ NAME_EDUCATION_TYPE_Higher education : int 0 1 1 1 0 0 0 0 0 0 ...
$ NAME_EDUCATION_TYPE_Incomplete higher : int 0 0 0 0 0 0 0 0 0 0 ...
$ NAME_EDUCATION_TYPE_Lower secondary : int 0 0 0 0 0 0 0 0 0 0 ...
$ NAME_EDUCATION_TYPE_Secondary / secondary special: int 1 0 0 0 1 1 1 1 1 1 ...
$ NAME_FAMILY_STATUS_Married : int 0 1 1 1 0 0 0 0 0 0 ...
$ NAME_FAMILY_STATUS_Separated : int 0 0 0 0 0 0 0 0 0 0 ...
$ NAME_FAMILY_STATUS_Single / not married : int 1 0 0 0 1 0 0 0 0 0 ...
$ NAME_FAMILY_STATUS_Unknown : int 0 0 0 0 0 0 0 0 0 0 ...
$ NAME_FAMILY_STATUS_Widow : int 0 0 0 0 0 0 0 0 0 0 ...
$ NAME_HOUSING_TYPE_House / apartment : int 1 1 1 1 1 1 1 1 1 1 ...
$ NAME_HOUSING_TYPE_Municipal apartment : int 0 0 0 0 0 0 0 0 0 0 ...
$ NAME_HOUSING_TYPE_Office apartment : int 0 0 0 0 0 0 0 0 0 0 ...
$ NAME_HOUSING_TYPE_Rented apartment : int 0 0 0 0 0 0 0 0 0 0 ...
$ NAME_HOUSING_TYPE_With parents : int 0 0 0 0 0 0 0 0 0 0 ...
$ OCCUPATION_TYPE_Accountants : int 0 0 0 0 0 0 0 0 0 0 ...
$ OCCUPATION_TYPE_Cleaning staff : int 0 0 0 0 0 0 0 0 0 0 ...
$ OCCUPATION_TYPE_Cooking staff : int 0 0 0 0 0 0 0 0 0 0 ...
$ OCCUPATION_TYPE_Core staff : int 0 1 1 1 0 0 0 0 0 0 ...
$ OCCUPATION_TYPE_Drivers : int 0 0 0 0 0 0 0 0 0 0 ...
$ OCCUPATION_TYPE_High skill tech staff : int 0 0 0 0 0 0 0 0 0 0 ...
$ OCCUPATION_TYPE_HR staff : int 0 0 0 0 0 0 0 0 0 0 ...
$ OCCUPATION_TYPE_IT staff : int 0 0 0 0 0 0 0 0 0 0 ...
$ OCCUPATION_TYPE_Laborers : int 1 0 0 0 1 1 1 1 1 1 ...
$ OCCUPATION_TYPE_Low-skill Laborers : int 0 0 0 0 0 0 0 0 0 0 ...
$ OCCUPATION_TYPE_Managers : int 0 0 0 0 0 0 0 0 0 0 ...
$ OCCUPATION_TYPE_Medicine staff : int 0 0 0 0 0 0 0 0 0 0 ...
$ OCCUPATION_TYPE_Private service staff : int 0 0 0 0 0 0 0 0 0 0 ...
$ OCCUPATION_TYPE_Realty agents : int 0 0 0 0 0 0 0 0 0 0 ...
$ OCCUPATION_TYPE_Sales staff : int 0 0 0 0 0 0 0 0 0 0 ...
$ OCCUPATION_TYPE_Secretaries : int 0 0 0 0 0 0 0 0 0 0 ...
$ OCCUPATION_TYPE_Security staff : int 0 0 0 0 0 0 0 0 0 0 ...
$ OCCUPATION_TYPE_Waiters/barmen staff : int 0 0 0 0 0 0 0 0 0 0 ...
$ REGION_RATING_CLIENT_2 : int 1 0 0 0 1 1 1 1 1 1 ...
$ REGION_RATING_CLIENT_3 : int 0 0 0 0 0 0 0 0 0 0 ...
$ ORGANIZATION_TYPE_Agriculture : int 0 0 0 0 0 0 0 0 0 0 ...
$ ORGANIZATION_TYPE_Bank : int 0 0 0 0 0 0 0 0 0 0 ...
$ ORGANIZATION_TYPE_Business Entity Type 1 : int 0 0 0 0 0 0 0 0 0 0 ...
$ ORGANIZATION_TYPE_Business Entity Type 2 : int 0 0 0 0 0 0 0 0 0 0 ...
$ ORGANIZATION_TYPE_Business Entity Type 3 : int 1 0 0 0 0 1 1 1 1 1 ...
$ ORGANIZATION_TYPE_Cleaning : int 0 0 0 0 0 0 0 0 0 0 ...
$ ORGANIZATION_TYPE_Construction : int 0 0 0 0 0 0 0 0 0 0 ...
$ ORGANIZATION_TYPE_Culture : int 0 0 0 0 0 0 0 0 0 0 ...
$ ORGANIZATION_TYPE_Electricity : int 0 0 0 0 0 0 0 0 0 0 ...
$ ORGANIZATION_TYPE_Emergency : int 0 0 0 0 0 0 0 0 0 0 ...
$ ORGANIZATION_TYPE_Government : int 0 0 0 0 1 0 0 0 0 0 ...
$ ORGANIZATION_TYPE_Hotel : int 0 0 0 0 0 0 0 0 0 0 ...
$ ORGANIZATION_TYPE_Housing : int 0 0 0 0 0 0 0 0 0 0 ...
$ ORGANIZATION_TYPE_Industry: type 1 : int 0 0 0 0 0 0 0 0 0 0 ...
$ ORGANIZATION_TYPE_Industry: type 10 : int 0 0 0 0 0 0 0 0 0 0 ...
$ ORGANIZATION_TYPE_Industry: type 11 : int 0 0 0 0 0 0 0 0 0 0 ...
$ ORGANIZATION_TYPE_Industry: type 12 : int 0 0 0 0 0 0 0 0 0 0 ...
$ ORGANIZATION_TYPE_Industry: type 13 : int 0 0 0 0 0 0 0 0 0 0 ...
$ ORGANIZATION_TYPE_Industry: type 2 : int 0 0 0 0 0 0 0 0 0 0 ...
$ ORGANIZATION_TYPE_Industry: type 3 : int 0 0 0 0 0 0 0 0 0 0 ...
$ ORGANIZATION_TYPE_Industry: type 4 : int 0 0 0 0 0 0 0 0 0 0 ...
$ ORGANIZATION_TYPE_Industry: type 5 : int 0 0 0 0 0 0 0 0 0 0 ...
$ ORGANIZATION_TYPE_Industry: type 6 : int 0 0 0 0 0 0 0 0 0 0 ...

```
$ ORGANIZATION_TYPE_Industry: type 7          : int  0 0 0 0 0 0 0 0 0 0 ...
$ ORGANIZATION_TYPE_Industry: type 8          : int  0 0 0 0 0 0 0 0 0 0 ...
 [list output truncated]
```

Lasso Coefficients

```
> print(lasso_coefficients)
156 x 1 sparse Matrix of class "dgCMatrix"
                                                  s1
(Intercept)                                -4.505958e+00
CNT_CHILDREN                                .
AMT_INCOME_TOTAL                            2.578217e-08
CURRENT_AMT_CREDIT                          .
CURRENT_AMT_ANNUITY                         5.337984e-06
CURRENT_AMT_GOODS_PRICE                    -8.788019e-07
REGION_POPULATION_RELATIVE                  7.936978e-01
DAYS_BIRTH                                 -1.761800e-05
DAYS_EMPLOYED                              -6.003947e-05
CNT_FAM_MEMBERS                             5.594786e-03
EXT_SOURCE_2                               -2.002711e+00
EXT_SOURCE_3                               -2.770387e+00
ADDRESS_MISMATCH_COUNT                      2.193629e-02
WEIGHTED_CREDIT_BUREAU_INQUIRIES           -2.799677e-01
SOCIAL_CIRCLE_DEFAULT_RATE                  3.074771e-01
NUM_DOCUMENTS_PROVIDED                     -3.277745e-01
CONTACT_PROVIDED_SUM                        9.866644e-03
PREVIOUS_AMT_ANNUITY                        9.232987e-07
AMT_APPLICATION                            -4.496796e-07
PREVIOUS_AMT_CREDIT                         2.190880e-07
CNT_PAYMENT                                1.541852e-02
LOG_AMT_INCOME_TOTAL                        1.162876e-01
LOG_CURRENT_AMT_CREDIT                      1.105337e+00
LOG_CURRENT_AMT_GOODS_PRICE                -7.960900e-01
LOG_AMT_APPLICATION                         1.248001e-02
LOG_PREVIOUS_AMT_CREDIT                    -9.612397e-02
debt_to_income_ratio                        3.757942e-02
age_in_years                                .
NO_PREVIOUS_LOANS                          -8.705877e-01
CURRENT_NAME_CONTRACT_TYPE_Revolving loans -5.527599e-01
CODE_GENDER_M                               2.906195e-01
CODE_GENDER_XNA                            -2.716536e+00
FLAG_OWN_CAR_2                             -2.494344e-01
FLAG_OWN_REALTY_2                           3.703949e-02
NAME_INCOME_TYPE_Commercial associate       .
NAME_INCOME_TYPE_Maternity leave            3.950647e+00
NAME_INCOME_TYPE_Pensioner                 -1.386834e-01
NAME_INCOME_TYPE_State servant              .
NAME_INCOME_TYPE_Student                   -2.537951e+00
NAME_INCOME_TYPE_Unemployed                 1.874638e+00
```

NAME_INCOME_TYPE_Working 1.042714e-01
NAME_EDUCATION_TYPE_Higher education -5.206103e-02
NAME_EDUCATION_TYPE_Incomplete higher 1.670813e-02
NAME_EDUCATION_TYPE_Lower secondary 2.616547e-01
NAME_EDUCATION_TYPE_Secondary / secondary special 2.411616e-01
NAME_FAMILY_STATUS_Married -1.279185e-01
NAME_FAMILY_STATUS_Separated -1.017481e-02
NAME_FAMILY_STATUS_Single / not married -3.809858e-02
NAME_FAMILY_STATUS_Unknown -5.176151e-02
NAME_FAMILY_STATUS_Widow -1.623525e-01
NAME_HOUSING_TYPE_House / apartment -6.412196e-02
NAME_HOUSING_TYPE_Municipal apartment 1.098339e-01
NAME_HOUSING_TYPE_Office apartment -2.071893e-01
NAME_HOUSING_TYPE_Rented apartment 1.315797e-02
NAME_HOUSING_TYPE_With parents 3.010486e-02
OCCUPATION_TYPE_Accountants -1.281760e-01
OCCUPATION_TYPE_Cleaning staff -1.271886e-03
OCCUPATION_TYPE_Cooking staff 6.191960e-02
OCCUPATION_TYPE_Core staff -8.311128e-02
OCCUPATION_TYPE_Drivers 1.319645e-01
OCCUPATION_TYPE_High skill tech staff -1.001921e-01
OCCUPATION_TYPE_HR staff 2.676352e-02
OCCUPATION_TYPE_IT staff -6.119980e-02
OCCUPATION_TYPE_Laborers 7.585497e-02
OCCUPATION_TYPE_Low-skill Laborers 3.745936e-01
OCCUPATION_TYPE_Managers -2.122798e-02
OCCUPATION_TYPE_Medicine staff -6.663988e-02
OCCUPATION_TYPE_Private service staff -2.094451e-01
OCCUPATION_TYPE_Realty agents 1.185263e-02
OCCUPATION_TYPE_Sales staff .
OCCUPATION_TYPE_Secretaries 1.667037e-01
OCCUPATION_TYPE_Security staff 2.055162e-01
OCCUPATION_TYPE_Waiters/barmen staff 8.235708e-02
REGION_RATING_CLIENT_2 2.108772e-01
REGION_RATING_CLIENT_3 3.535015e-01
ORGANIZATION_TYPE_Agriculture 2.335942e-02
ORGANIZATION_TYPE_Bank -2.974086e-01
ORGANIZATION_TYPE_Business Entity Type 1 3.050427e-02
ORGANIZATION_TYPE_Business Entity Type 2 1.347225e-02
ORGANIZATION_TYPE_Business Entity Type 3 4.991144e-02
ORGANIZATION_TYPE_Cleaning -1.221559e-02
ORGANIZATION_TYPE_Construction 1.913624e-01
ORGANIZATION_TYPE_Culture .
ORGANIZATION_TYPE_Electricity -1.480533e-01
ORGANIZATION_TYPE_Emergency -2.417991e-01
ORGANIZATION_TYPE_Government -5.698411e-02
ORGANIZATION_TYPE_Hotel -2.339476e-01
ORGANIZATION_TYPE_Housing 3.956122e-02
ORGANIZATION_TYPE_Industry: type 1 1.732082e-01
ORGANIZATION_TYPE_Industry: type 10 -7.006763e-01
ORGANIZATION_TYPE_Industry: type 11 1.657209e-02
ORGANIZATION_TYPE_Industry: type 12 -9.347003e-01
ORGANIZATION_TYPE_Industry: type 13 -4.298382e-01
ORGANIZATION_TYPE_Industry: type 2 -1.983077e-01
ORGANIZATION_TYPE_Industry: type 3 1.014952e-01

ORGANIZATION_TYPE_Industry: type 4 -1.089545e-01
ORGANIZATION_TYPE_Industry: type 5 -2.291790e-01
ORGANIZATION_TYPE_Industry: type 6 -1.780376e-01
ORGANIZATION_TYPE_Industry: type 7 -1.646261e-02
ORGANIZATION_TYPE_Industry: type 8 3.126364e-01
ORGANIZATION_TYPE_Industry: type 9 -1.223222e-01
ORGANIZATION_TYPE_Insurance -2.377591e-01
ORGANIZATION_TYPE_Kindergarten -1.515694e-01
ORGANIZATION_TYPE_Legal Services 1.482123e-01
ORGANIZATION_TYPE_Medicine .
ORGANIZATION_TYPE_Military -2.915967e-01
ORGANIZATION_TYPE_Mobile .
ORGANIZATION_TYPE_Other 2.344747e-02
ORGANIZATION_TYPE_Police -2.125730e-01
ORGANIZATION_TYPE_Postal 3.299073e-02
ORGANIZATION_TYPE_Realtor 6.384490e-01
ORGANIZATION_TYPE_Religion 2.852798e-01
ORGANIZATION_TYPE_Restaurant 1.259690e-01
ORGANIZATION_TYPE_School -9.376114e-02
ORGANIZATION_TYPE_Security -1.710315e-01
ORGANIZATION_TYPE_Security Ministries -4.885734e-01
ORGANIZATION_TYPE_Self-employed 1.429144e-01
ORGANIZATION_TYPE_Services 6.700521e-02
ORGANIZATION_TYPE_Telecom 1.255070e-01
ORGANIZATION_TYPE_Trade: type 1 1.743050e-01
ORGANIZATION_TYPE_Trade: type 2 -4.157326e-01
ORGANIZATION_TYPE_Trade: type 3 1.122080e-01
ORGANIZATION_TYPE_Trade: type 4 -1.268898e+00
ORGANIZATION_TYPE_Trade: type 5 -4.424499e-01
ORGANIZATION_TYPE_Trade: type 6 -1.957662e-01
ORGANIZATION_TYPE_Trade: type 7 1.230320e-01
ORGANIZATION_TYPE_Transport: type 1 -6.846065e-01
ORGANIZATION_TYPE_Transport: type 2 .
ORGANIZATION_TYPE_Transport: type 3 5.714258e-01
ORGANIZATION_TYPE_Transport: type 4 -1.200717e-02
ORGANIZATION_TYPE_University -2.167237e-01
ORGANIZATION_TYPE_XNA .
PREVIOUS_NAME_CONTRACT_TYPE_Cash loans .
PREVIOUS_NAME_CONTRACT_TYPE_Consumer loans -1.458740e-01
PREVIOUS_NAME_CONTRACT_TYPE_Revolving loans 2.684191e-01
NAME_CONTRACT_STATUS_Approved .
NAME_CONTRACT_STATUS_Canceled -5.152981e-02
NAME_CONTRACT_STATUS_Refused 1.849536e-01
NAME_CONTRACT_STATUS_Unused offer -3.006617e-02
NAME_CLIENT_TYPE_New 1.858602e-01
NAME_CLIENT_TYPE_Refreshed .
NAME_CLIENT_TYPE_Repeater -2.215685e-03
NAME_CLIENT_TYPE_XNA -6.323516e-02
NAME_PORTFOLIO_Cards 9.154445e-02
NAME_PORTFOLIO_Cars 3.848947e-03
NAME_PORTFOLIO_Cash .
NAME_PORTFOLIO_POS .
NAME_PORTFOLIO_XNA 4.449399e-02
CHANNEL_TYPE_AP+ (Cash loan) 1.742730e-01
CHANNEL_TYPE_Car dealer -3.652101e-01

```
CHANNEL_TYPE_Channel of corporate sales        -5.941312e-01
CHANNEL_TYPE_Contact center                    1.110928e-02
CHANNEL_TYPE_Country-wide                      .
CHANNEL_TYPE_Credit and cash offices           1.533558e-02
CHANNEL_TYPE_Regional / Local                  -3.649909e-02
CHANNEL_TYPE_Stone                             -4.633298e-03
```

```
> selected_features
  [1] "AMT_INCOME_TOTAL"                          "CURRENT_AMT_ANNUITY"
  [3] "CURRENT_AMT_GOODS_PRICE"                   "DAYS_BIRTH"
  [5] "DAYS_EMPLOYED"                             "CNT_FAM_MEMBERS"
  [7] "EXT_SOURCE_2"                              "EXT_SOURCE_3"
  [9] "ADDRESS_MISMATCH_COUNT"                    "WEIGHTED_CREDIT_BUREAU_INQUIRIES"
 [11] "SOCIAL_CIRCLE_DEFAULT_RATE"                "NUM_DOCUMENTS_PROVIDED"
 [13] "PREVIOUS_AMT_ANNUITY"                      "AMT_APPLICATION"
 [15] "CNT_PAYMENT"                               "LOG_AMT_INCOME_TOTAL"
 [17] "LOG_CURRENT_AMT_CREDIT"                    "LOG_CURRENT_AMT_GOODS_PRICE"
 [19] "LOG_AMT_APPLICATION"                       "LOG_PREVIOUS_AMT_CREDIT"
 [21] "debt_to_income_ratio"                      "NO_PREVIOUS_LOANS"
 [23] "CURRENT_NAME_CONTRACT_TYPE_Revolving loans"  "CODE_GENDER_M"
 [25] "CODE_GENDER_XNA"                           "FLAG_OWN_CAR_2"
 [27] "FLAG_OWN_REALTY_2"                         "NAME_INCOME_TYPE_Maternity leave"
 [29] "NAME_INCOME_TYPE_Pensioner"                "NAME_INCOME_TYPE_State servant"
 [31] "NAME_INCOME_TYPE_Student"                  "NAME_INCOME_TYPE_Unemployed"
 [33] "NAME_INCOME_TYPE_Working"                  "NAME_EDUCATION_TYPE_Higher education"
 [35] "NAME_EDUCATION_TYPE_Lower secondary"       "NAME_EDUCATION_TYPE_Secondary / secondary special"
 [37] "NAME_FAMILY_STATUS_Married"                "NAME_FAMILY_STATUS_Widow"
 [39] "NAME_HOUSING_TYPE_House / apartment"       "NAME_HOUSING_TYPE_Municipal apartment"
 [41] "NAME_HOUSING_TYPE_Office apartment"        "OCCUPATION_TYPE_Accountants"
 [43] "OCCUPATION_TYPE_Cooking staff"             "OCCUPATION_TYPE_Core staff"
 [45] "OCCUPATION_TYPE_Drivers"                   "OCCUPATION_TYPE_High skill tech staff"
 [47] "OCCUPATION_TYPE_Laborers"                  "OCCUPATION_TYPE_Low-skill Laborers"
 [49] "OCCUPATION_TYPE_Medicine staff"            "OCCUPATION_TYPE_Private service staff"
 [51] "OCCUPATION_TYPE_Secretaries"               "OCCUPATION_TYPE_Security staff"
 [53] "OCCUPATION_TYPE_Waiters/barmen staff"      "REGION_RATING_CLIENT_2"
 [55] "REGION_RATING_CLIENT_3"                    "ORGANIZATION_TYPE_Bank"
 [57] "ORGANIZATION_TYPE_Business Entity Type 3"  "ORGANIZATION_TYPE_Construction"
 [59] "ORGANIZATION_TYPE_Electricity"             "ORGANIZATION_TYPE_Emergency"
 [61] "ORGANIZATION_TYPE_Government"              "ORGANIZATION_TYPE_Hotel"
 [63] "ORGANIZATION_TYPE_Industry: type 1"        "ORGANIZATION_TYPE_Industry: type 10"
 [65] "ORGANIZATION_TYPE_Industry: type 12"       "ORGANIZATION_TYPE_Industry: type 13"
 [67] "ORGANIZATION_TYPE_Industry: type 2"        "ORGANIZATION_TYPE_Industry: type 3"
 [69] "ORGANIZATION_TYPE_Industry: type 4"        "ORGANIZATION_TYPE_Industry: type 5"
 [71] "ORGANIZATION_TYPE_Industry: type 9"        "ORGANIZATION_TYPE_Insurance"
 [73] "ORGANIZATION_TYPE_Kindergarten"            "ORGANIZATION_TYPE_Military"
 [75] "ORGANIZATION_TYPE_Police"                  "ORGANIZATION_TYPE_Realtor"
 [77] "ORGANIZATION_TYPE_Restaurant"              "ORGANIZATION_TYPE_School"
 [79] "ORGANIZATION_TYPE_Security"                "ORGANIZATION_TYPE_Security Ministries"
 [81] "ORGANIZATION_TYPE_Self-employed"           "ORGANIZATION_TYPE_Trade: type 1"
 [83] "ORGANIZATION_TYPE_Trade: type 2"           "ORGANIZATION_TYPE_Trade: type 3"
 [85] "ORGANIZATION_TYPE_Trade: type 4"           "ORGANIZATION_TYPE_Trade: type 6"
 [87] "ORGANIZATION_TYPE_Trade: type 7"           "ORGANIZATION_TYPE_Transport: type 1"
 [89] "ORGANIZATION_TYPE_Transport: type 3"       "ORGANIZATION_TYPE_University"
 [91] "PREVIOUS_NAME_CONTRACT_TYPE_Consumer loans"  "PREVIOUS_NAME_CONTRACT_TYPE_Revolving loans"
 [93] "NAME_CONTRACT_STATUS_Refused"              "NAME_CLIENT_TYPE_New"
 [95] "NAME_PORTFOLIO_Cards"                      "NAME_PORTFOLIO_POS"
 [97] "CHANNEL_TYPE_AP+ (Cash loan)"              "CHANNEL_TYPE_Car dealer"
 [99] "CHANNEL_TYPE_Channel of corporate sales"   "CHANNEL_TYPE_Credit and cash offices"
[101] "CHANNEL_TYPE_Regional / Local"
```

Significant variables

```
> significant_vars
 [1] "CURRENT_AMT_ANNUITY"                       "CURRENT_AMT_GOODS_PRICE"
 [3] "DAYS_BIRTH"                                "DAYS_EMPLOYED"
 [5] "CNT_FAM_MEMBERS"                           "EXT_SOURCE_2"
 [7] "EXT_SOURCE_3"                              "ADDRESS_MISMATCH_COUNT"
 [9] "WEIGHTED_CREDIT_BUREAU_INQUIRIES"          "SOCIAL_CIRCLE_DEFAULT_RATE"
[11] "NUM_DOCUMENTS_PROVIDED"                    "PREVIOUS_AMT_ANNUITY"
[13] "AMT_APPLICATION"                           "CNT_PAYMENT"
[15] "LOG_AMT_INCOME_TOTAL"                      "LOG_CURRENT_AMT_CREDIT"
[17] "LOG_CURRENT_AMT_GOODS_PRICE"               "LOG_AMT_APPLICATION"
[19] "LOG_PREVIOUS_AMT_CREDIT"                   "debt_to_income_ratio"
[21] "NO_PREVIOUS_LOANS"                         "`CURRENT_NAME_CONTRACT_TYPE_Revolving loans`"
[23] "CODE_GENDER_M"                             "FLAG_OWN_CAR_2"
[25] "FLAG_OWN_REALTY_2"                         "`NAME_INCOME_TYPE_Maternity leave`"
[27] "NAME_INCOME_TYPE_Pensioner"                "NAME_INCOME_TYPE_Unemployed"
[29] "NAME_INCOME_TYPE_Working"                  "`NAME_EDUCATION_TYPE_Higher education`"
[31] "`NAME_EDUCATION_TYPE_Lower secondary`"     "`NAME_EDUCATION_TYPE_Secondary / secondary special`"
[33] "NAME_FAMILY_STATUS_Married"                "NAME_FAMILY_STATUS_Widow"
[35] "`NAME_HOUSING_TYPE_House / apartment`"     "`NAME_HOUSING_TYPE_Municipal apartment`"
[37] "`NAME_HOUSING_TYPE_Office apartment`"      "OCCUPATION_TYPE_Accountants"
[39] "`OCCUPATION_TYPE_Cooking staff`"           "`OCCUPATION_TYPE_Core staff`"
[41] "OCCUPATION_TYPE_Drivers"                   "`OCCUPATION_TYPE_High skill tech staff`"
[43] "OCCUPATION_TYPE_Laborers"                  "`OCCUPATION_TYPE_Low-skill Laborers`"
[45] "`OCCUPATION_TYPE_Medicine staff`"          "`OCCUPATION_TYPE_Private service staff`"
[47] "OCCUPATION_TYPE_Secretaries"               "`OCCUPATION_TYPE_Security staff`"
[49] "REGION_RATING_CLIENT_2"                    "REGION_RATING_CLIENT_3"
[51] "ORGANIZATION_TYPE_Bank"                    "`ORGANIZATION_TYPE_Business Entity Type 3`"
[53] "ORGANIZATION_TYPE_Construction"            "ORGANIZATION_TYPE_Electricity"
[55] "ORGANIZATION_TYPE_Government"              "ORGANIZATION_TYPE_Hotel"
[57] "`ORGANIZATION_TYPE_Industry: type 1`"      "`ORGANIZATION_TYPE_Industry: type 10`"
[59] "`ORGANIZATION_TYPE_Industry: type 12`"     "`ORGANIZATION_TYPE_Industry: type 13`"
[61] "`ORGANIZATION_TYPE_Industry: type 3`"      "`ORGANIZATION_TYPE_Industry: type 5`"
[63] "`ORGANIZATION_TYPE_Industry: type 9`"      "ORGANIZATION_TYPE_Insurance"
[65] "ORGANIZATION_TYPE_Kindergarten"            "ORGANIZATION_TYPE_Military"
[67] "ORGANIZATION_TYPE_Police"                  "ORGANIZATION_TYPE_Realtor"
[69] "ORGANIZATION_TYPE_Restaurant"              "ORGANIZATION_TYPE_School"
[71] "ORGANIZATION_TYPE_Security"                "`ORGANIZATION_TYPE_Security Ministries`"
[73] "`ORGANIZATION_TYPE_Self-employed`"         "`ORGANIZATION_TYPE_Trade: type 1`"
[75] "`ORGANIZATION_TYPE_Trade: type 2`"         "`ORGANIZATION_TYPE_Trade: type 3`"
[77] "`ORGANIZATION_TYPE_Trade: type 4`"         "`ORGANIZATION_TYPE_Trade: type 6`"
[79] "`ORGANIZATION_TYPE_Trade: type 7`"         "`ORGANIZATION_TYPE_Transport: type 1`"
[81] "`ORGANIZATION_TYPE_Transport: type 3`"     "ORGANIZATION_TYPE_University"
[83] "`PREVIOUS_NAME_CONTRACT_TYPE_Consumer loans`"  "NAME_CONTRACT_STATUS_Refused"
[85] "NAME_CLIENT_TYPE_New"                      "`CHANNEL_TYPE_AP+ (Cash loan)`"
[87] "`CHANNEL_TYPE_Car dealer`"                 "`CHANNEL_TYPE_Channel of corporate sales`"
[89] "`CHANNEL_TYPE_Regional / Local`"
```