Project CheckPoint 1

Targeted Risk Understanding & Scoring Technology (TRUST)

Arunbh Yashaswi, Swattik Maiti, Eniyan Ezhilan , Ajaykumar Balakannan , Ritik Pratap Singh

09.27.2024

Data 602 Principles of Data Science

Github Repo: TRUST Repo

INTRODUCTION

In today’s financial environment, which is pretty sophisticated, assessment of the borrower's ability to repay has become a key factor in the risk management of any lending institution. With the continuous growth in the credit market, the hectic process of arriving at informed and well-judged decisions on providing loans remains a constant challenge to the financial institutions to efficiently manage the risk of defaults. The project addresses that challenge by developing a predictive model based on data from the Home Credit Default Risk dataset, which includes detailed information on clients' applications, credit histories, and previous financial behavior, as well as various socio-economic factors.

Our Primary objective will be to elaborate a predictive model able to evaluate, with high precision, the probability of a given candidate defaulting on a loan. This would help lenders make more appropriate credit decisions that further optimize the ratio of loan approvals by minimizing losses due to defaults. We will therefore endeavor to enhance the precision of credit risk assessment for lenders to go further into the profiles of high-risk applicants through trend and pattern analyses.

Why Did We Choose This Project?

Loan default is one of the major problems a financial institution faces, as this has a huge impact on their profitability and the capability of giving credit to other customers.

Drawing inspiration from the mishap of 2008 where the whole financial market collapsed due to millions of dollars worth subprime home loans which were given out to individuals without proper background check on their credit worthiness – this project aims to create a structured framework of machine learning based data processing to prevent such instances in the future.

This Solution will help financial institutions in being more efficient while making proper decisions by enabling them to evaluate which applicants are most likely to default, hence mitigating risks and ensuring a long-term financial health for them. We chose this project to apply data science techniques to a real problem impacting both the financial industry as a whole and those individuals applying for credit.

The Home Credit Default Risk dataset is quite feature-rich, hence leaving a lot of scope for detailed exploratory analysis along many dimensions. We're really excited to dive into this data, extract meaningful patterns out of it, and contribute to developing a more robust loan assessment process.

What problem are we solving?

The major problem we are solving is to identify the applicants who hold a higher risk of defaulting on loans. Most loan applications consist of a huge volume of customer data; it's really tough for lenders to assess the risk manually. That is why we will develop a machine learning model that would assist in the automation of loan default risk prediction which helps financial institutions in determining defaulters and increasing their efficiency.

Specifically, our analysis will:

- Finding key factors that add to the probability of default.
- Creating a model to segregate applicants into different risk categories.
- Provide insight into which customer profiles are more likely to default and which are the customers who will repay.

What Is the Impact of the Project?

The project has immense potential for various financial institutions in improving credit risk assessment with the view to reducing the likelihood of bad loans, hence enabling companies to focus resources on customers who will most probably pay. Equally, it may also enable lenders to extend credit in a responsible way, offering loans to those in real need while minimizing the risk of financial loss.

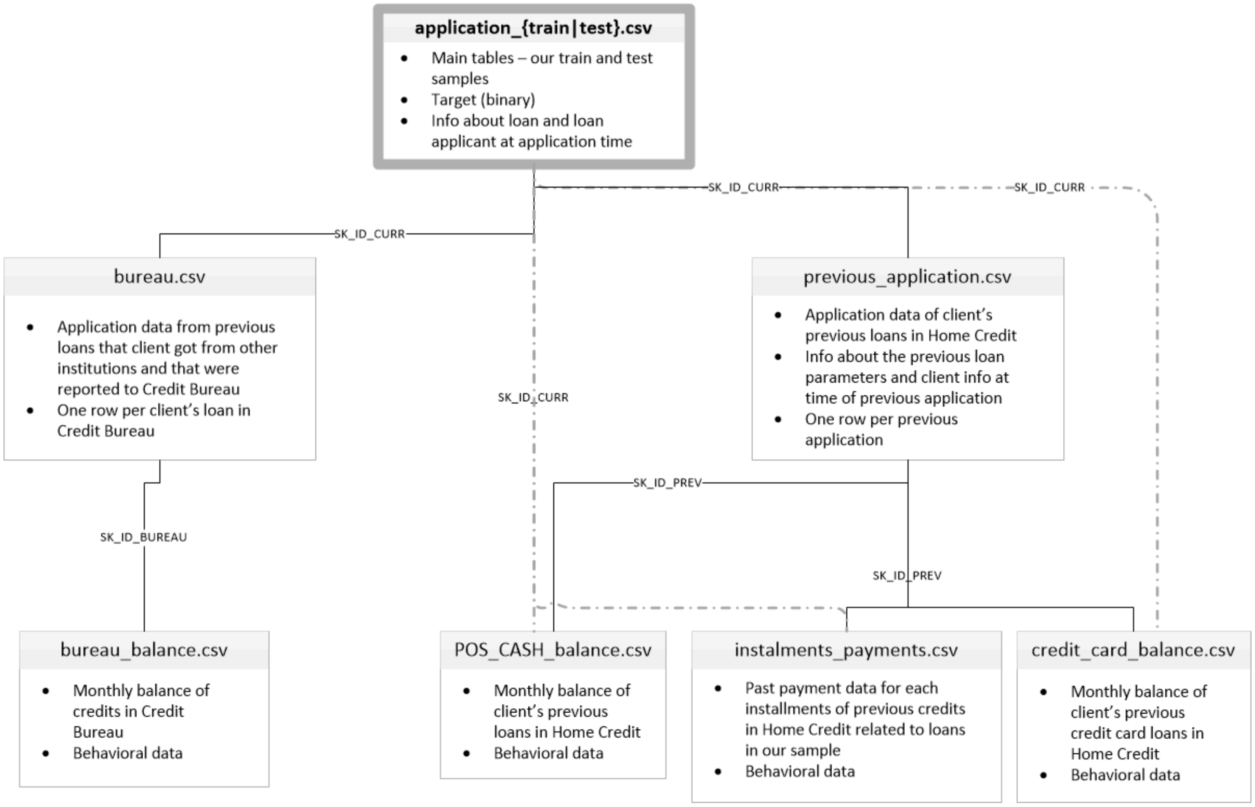
The model will:

- Help lenders make more informed decisions based on data.

- Reduce financial losses due to default.
- Support responsible lending by offering a loan to a borrower that is suitable for them.

Dataset Source:

Home Credit Default Risk: A dataset provided by Kaggle, containing customer and loan data, including demographics, credit history, and previous repayment behavior. Link



application_{train|test}.csv
- Main tables – our train and test samples
- Target (binary)
- Info about loan and loan applicant at application time

bureau.csv
- Application data from previous loans that client got from other institutions and that were reported to Credit Bureau
- One row per client's loan in Credit Bureau

previous_application.csv
- Application data of client's previous loans in Home Credit
- Info about the previous loan parameters and client info at time of previous application
- One row per previous application

bureau_balance.csv
- Monthly balance of credits in Credit Bureau
- Behavioral data

POS_CASH_balance.csv
- Monthly balance of client's previous loans in Home Credit
- Behavioral data

instalments_payments.csv
- Past payment data for each installments of previous credits in Home Credit related to loans in our sample
- Behavioral data

credit_card_balance.csv
- Monthly balance of client's previous credit card loans in Home Credit
- Behavioral data

Figure : The dataset schema

Why This Dataset?

The Home Credit Default Risk dataset contains a very diverse and elaborated range of data points, which will be ideal to predict the loan default risks. Therefore, we can build a complete model for multiple dimensions in financial history and behavior that bring us to a complete view about the borrower's creditworthiness.

Dataset Details:

- **application_{train|test}.csv:** This is the primary dataset containing all applications. The training set includes a target variable indicating whether a client defaulted, and the test set contains the same features without the target variable. This file provides static data for each loan application, helping us identify key patterns between applicant characteristics and default risk.
- **bureau.csv:** This dataset contains information on clients' previous credits from other financial institutions, reported to the Credit Bureau. By incorporating external credit histories, we can better evaluate an applicant's overall financial responsibility.
- **bureau_balance.csv:** A monthly record of balances for previous credits reported to the Credit Bureau. This file gives us temporal insights into a client's credit activity and trends, which could signal impending defaults or reliable repayment patterns.
- **POS_CASH_balance.csv:** This table provides monthly snapshots of balances for POS (point of sales) and cash loans. It shows us the loan dynamics and repayment behavior of applicants, which are crucial for understanding short-term financial stability.
- **credit_card_balance.csv:** Similar to POS and cash loans, this dataset tracks monthly balances for credit cards. Credit card repayment behavior is often a strong indicator of an applicant's financial health, adding another layer of precision to our analysis.
- **previous_application.csv:** This dataset includes details of previous applications for loans. Understanding the history of applications and how they were handled gives us more context on an applicant's loan-seeking behavior and risk of default.
- **installments_payments.csv:** This table captures the repayment history for previously disbursed credits, with one row for each payment (or missed payment). By analyzing the frequency and timeliness of repayments, we can gain deeper insights into the applicant's repayment habits, which are crucial for credit risk evaluation.
- **HomeCredit_columns_description.csv:** This file contains the descriptions of each column in the various datasets, ensuring we can interpret the variables correctly and use them effectively in our model.

Why This Dataset for Credit Risk Prediction?

If we integrate all the data together from different sources like past loan histories, credit card balances, repayments behavior, and so on, a very accurate prediction model can be developed. Because the dataset has such depth and breadth, we can analyze applicant profiles much more deeply than just basic credit scores to show nuanced risk factors that may otherwise have been missed in traditional sets of evaluations.

The richness of the monthly balance snapshots, repayment history, and external credit records captures both static and dynamic features, hence providing a complete view of the financial stability of every applicant. This will, therefore, enable us to come up with a model on which financial institutions can rely in making better data-driven decisions on lending.

What Methodologies Will You Use?

To solve the prediction problem of default risk of loans, we will cover the core phases of the Data Science Lifecycle:

1. **Data Collection:** For the project, the datasets used include application_train.csv, bureau.csv, and several other datasets taken from the Home Credit Default Risk challenge. This will give a wide feature range from clients' demographics to previous credit history and even repayment behavior that gives us a very sound foundation to work our analysis from.
2. **Data Processing:** After data collection, certain preprocessing will be done in order to clean and structure the data. This involves handling missing values, outlier detection, feature transformation-which can be encoding categorical variables or normalizing numeric data. Proper data processing is an important step in making the dataset ready for analysis and machine learning models.
3. **EDA & Data Visualization:** EDA will outline the pattern and trend within the data. Distribution, correlation, and relationships of major variables like loan amount, income, credit history are visualized with heatmaps, box plots, and pairwise correlation matrices. This shall help in narrowing down the features most influential in loan default cases.
4. **Analysis and Hypothesis Testing:** This will include an association of variables, then hypothesizing on what contributes to default. The hypotheses will be tested through statistical tests. For example, we may consider that the longer credit history decreases the risk of default and then verify our assumption through the use of statistical testing on data.
5. **Machine Learning Models:** In this regard, we will use machine learning

algorithms: Logistic Regression, Decision Trees, and Gradient Boosting to classify loan defaults. Model selection and tuning will involve key performance metrics including but not limited to accuracy, precision, recall, and ROC-AUC with the goal of optimizing said predictions to minimize financial risk.

6. **Insights & Policy Decisions:** The final stage involves the interpretation of the model's prediction and extraction of actionable insights. It will guide the lenders through proper risk assessment by pointing out specific credit behaviors and personal profile details from which an applicant has a high risk. These can then be developed into policy, enabling financial institutions to minimize potential losses while still serving a broader customer base.

What Is Your Expected Outcome?

Our goal is to train multiple models including both classical and neural network based models on the data and classify each user into different risk bands. Using the risk bands the underwriter can take actions about sanctioning a loan or not. We also aim to give a comparative performance analysis of the different models trained. The 2 best models are to be deployed in a champion-challenger fashion, which could be utilized by financial organizations on their respective data.

We would like to offer actionable insights by identifying crucial risk factors that would, in turn, allow lenders to intelligibly understand profiles of applicants who are rated highly and lowly risky. By the end of this project, it is our hope that we reduce uncertainty in loan approvals and develop the use of an overall decision-making process for financial institutions.

Credit:

Home Credit Group (link)

Image 0: (Link)