Home Credit Default Risk Analysis - Solution Competition

Fabian Esteban Perilla Lugo,

Julián Alejandro Pinzon Torres

Daniel Alejandro Chavez Bustos

[1] Ingeniería de Sistemas, Facultad de Ingeniería

Universidad Distrital Francisco José de Caldas

fperillal@udistrital.edu.co

dachavezb@udistrital.edu.co

japinzont@udistrital.edu.co

ABSTRACT

This document will address the solution we gave as a team to the kaggle competition "Home Credit Default Risk Analysis", understanding the problem to be evaluated and whose purpose is to create a predictive model able to determine whether a person is able to pay a loan, for the development to perform a respective work of data preprocessing, classification of features, exploratory data analysis, decision making and training of predictive models was done. The results of this work include a jupyter notebook with all the process performed.

INTRODUCTION

For banks, these types of loans can be highly risky due to the probability of defaults. Therefore, it is crucial to accurately analyze the information provided by each loan applicant. This includes data such as age, loan type, employment history, current household and family type, and other relevant information that can offer a clear perspective when making lending decisions.

To provide a solution to this problem, the competition utilizes data from various sources, including external ones, to better assess applicants' creditworthiness and improve the performance of the decision-making model. It is essential to analyze and clean this data to ensure that the model training is as precise as possible.

Implementing a credit system based on alternative data sources presents several challenges for banks. One major difficulty is ensuring the accuracy and reliability of non-traditional data, as these sources may not be as standardized or regulated as traditional credit data. Additionally, integrating various data types into a cohesive and functional model requires significant technical expertise and resources. A notable part of the population, particularly young adults, may have limited traditional financial records, making it essential to develop a robust system that accurately assesses their creditworthiness.

To address these challenges, banks can leverage advanced machine learning algorithms and data processing techniques. These tools can help in accurately predicting credit risk by analyzing large and diverse datasets. Implementing robust data cleaning and validation processes ensures the reliability of the data used. Additionally, utilizing secure database management systems and encryption technologies can safeguard sensitive information, ensuring compliance with data privacy regulations.

In this context, this article aims to explore the development of an efficient credit evaluation system for housing loans, using advanced computing techniques to optimize its functionality and efficiency. Through this work, we aim to lay the groundwork for more effective information management in the dynamic environment of the housing finance industry.

METHOD AND MATERIALS

In this section we will talk about the technical development carried out for the data science process, decisions made, and elements used in Python to reach the final solution.

- Libraries: For the complete development of the competition, essential libraries were used for data processing such as numpy, pandas and some scikit-learn modules, for the visualization of data in graphs we used matplotlib and seaborn, for the classification of features we use scikit-learn modules and to train predictive models we use XGBoost and Tensorflow

- Selection of .csv files: The competition included a variety of CSV files with which the data science process can be carried out but taking into account that 'application_train.csv' contains the general information of the loans, it has the schematized information of once with the test CSV 'application_test.csv' and presents a significant volume of cases to make the respective predictive models.

- Data preprocessing: In this section, our goal is to make the data as processable as possible for visualization and training predictive models. As a first step after loading the dataset, we check for the presence of null values in each column. Given the large size of the dataframe (122 columns), we create a function that takes a dataframe as input and returns another dataframe. The index of this new dataframe consists of the column names from the initial dataframe, and it has two additional columns: the first column shows the number of null values, and the second column displays the percentage of null values relative to the total number of records in the dataset.

When performing this step, we noticed that there are a large number of columns with null values that share a common trait: all these columns have suffixes such as _AVG, _MODE, and _MEDI. According to the competition documentation, these suffixes indicate that the columns contain normalized data and refer to statistical expressions of other columns. However, due to the high volume of null values, we can discard these columns. As a result, we reduced the dataframe from 122 columns to 75 columns.

As the next step in data preprocessing, we examined the data types of the columns. For columns with the data type 'object', we reviewed the number of unique values. We found columns with 2, 3, or more unique values. To handle these, we applied either label encoding or one-hot encoding as appropriate.

To avoid introducing any bias or weighting when converting string values to numeric values, we used the following approach:

- For columns with 2 unique values, we applied label encoding.
- For columns with 3 unique values, we reviewed them to determine if any could be imputed; for the remaining columns, we applied one-hot encoding.

This process aimed to ensure that converting categorical variables to numeric form did not skew the data, even though it might increase the number of columns in the dataframe.

As the final step in data preprocessing, we performed the detection and removal of outliers in key columns. One significant case was the 'DAYS_EMPLOYED' column, which contained the highest number of outliers. To address this, we applied a z-score with a threshold of 2.

- Exploratory Data Analysis (EDA): In this section, we aim to examine data behaviors, correlations, and the relevance of various factors that influence whether a person will repay a loan or not.

Realizing that examining too many columns, even after applying label encoding/one-hot encoding, would be a very complex and time-consuming task, we decided to focus on identifying the relevance of specific columns. To achieve this, we opted to apply the Random Forest model to obtain an output with feature importance. Once applied, this provided us with a list, which we then sorted according to the importance of each feature.

We decided to create a heatmap matrix of the columns that had an importance score between 0.01 and 0.06. We noticed that there are no significant correlations directly with the target column, but according to the Random Forest model, these features still hold substantial weight (such as EXT_SOURCE_2, EXT_SOURCE_3, and DAYS_BIRTH, DAYS_EMPLOYED). We want to examine how these features relate to the binary values of the target column, in this way, we discovered interesting findings such as younger populations being more prone to loan repayment difficulties, and that external credit scores have a significant influence. Scores closer to 0 indicate a higher likelihood of payment issues.

- **Model Building:** We selected two models that we believed could yield good results, moving beyond traditional linear and logistic regression models. Therefore, we chose to apply XGBoost and neural networks using TensorFlow. This approach resulted in high effectiveness when comparing the training and test CSV files.

RESULTS

After conducting exploratory analysis and data cleaning, it was possible to identify the features most influential during model training and their relationships crucial for effective classification. The following results were found:

- External Evaluation:

Within the competition dataset, external data sources assigning ratings to participants were identified. These exhibited strong correlations in the final model, significantly enhancing prediction accuracy. This external rating captures aspects not fully reflected in traditional data, based on independent evaluations, thereby bolstering robustness by considering multiple perspectives on participants' repayment capacity.

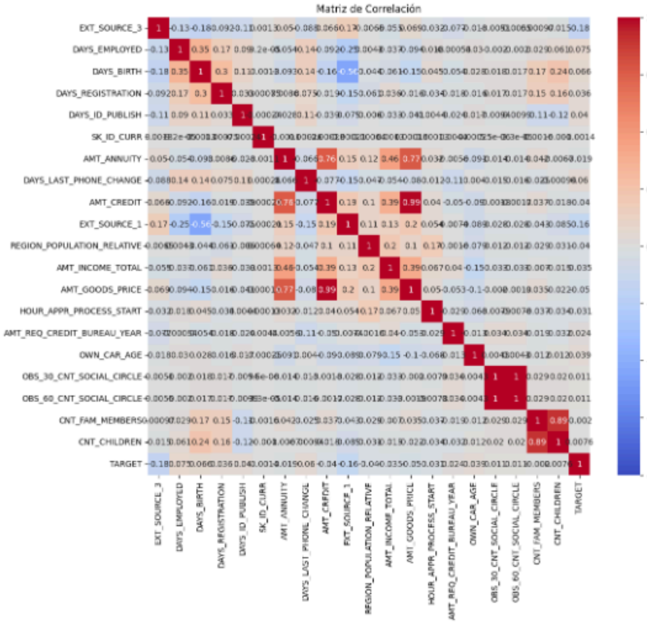


Figure 1: correlation hratmap

- Age:

Age emerged as a significant variable influencing credit repayment capability within the dataset. Younger applicants tended to pose higher default risks, likely due to lower financial stability. Conversely, older applicants demonstrated higher compliance tendencies, possibly attributable to greater job stability, as linked to the following item.

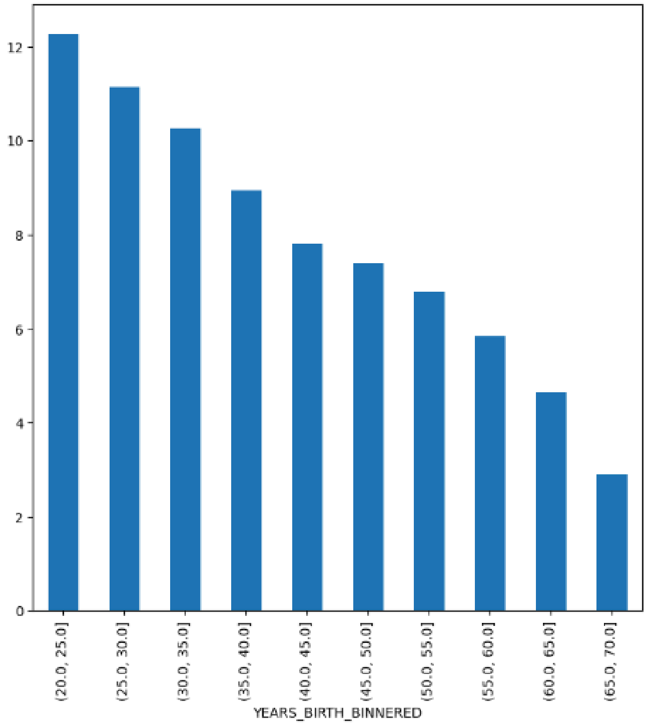
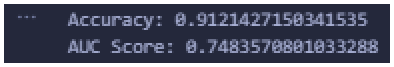


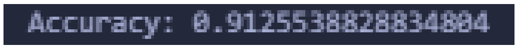
Figure 2: probability of no repayment by age range

The combination of these features, along with discarding some that had various characteristics such as many empty data points or correlations that did not contribute to the model, allowed us to train the models we used, including XGBoost and a neural network model using TensorFlow for training. This enables a comparison between these models.



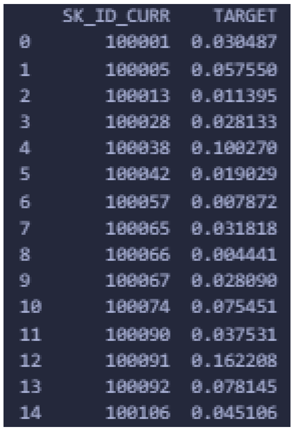
```
Accuracy: 0.9121427150341535
AUC Score: 0.7483570801033288
```

Figure 5: Accuracy XGBoost



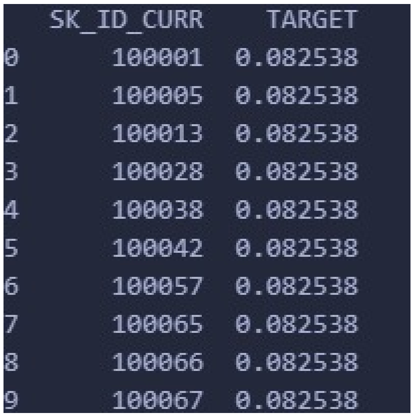
```
Accuracy: 0.9125538828834804
```

Figure 6: Accuracy Neural Network whit Tensor Flow



```
    SK_ID_CURR    TARGET
0       100001  0.030487
1       100005  0.057550
2       100013  0.011395
3       100028  0.028133
4       100038  0.100270
5       100042  0.019029
6       100057  0.007872
7       100065  0.031818
8       100066  0.004441
9       100067  0.028090
10      100074  0.075451
11      100090  0.037531
12      100091  0.162208
13      100092  0.078145
14      100106  0.045106
```

Figure 7: Output XGBoost



```
   SK_ID_CURR    TARGET
0      100001  0.082538
1      100005  0.082538
2      100013  0.082538
3      100028  0.082538
4      100038  0.082538
5      100042  0.082538
6      100057  0.082538
7      100065  0.082538
8      100066  0.082538
9      100067  0.082538
```

Figure 8: Output Neural Network

"With these images, we can see that training for both models yielded good results, but when generating the submission file, the neural network shows overfitting, whereas the XGBoost model performed better between the two."

DISCUSSION

In this section, we analyze the results obtained from the loan assessment model and discuss potential areas for improvement and future developments. Reflections on challenges encountered during the model development process are provided, along with suggestions for alternative solutions, considering the appropriateness of our approach.

-Alignment with Loan Assessment Objectives:

A thorough analysis is conducted to evaluate how the obtained results align with the initial objectives of the loan assessment project. Specific goals set at the outset are examined in relation to the achievements made during the implementation of the model. Areas where significant progress was achieved in predicting creditworthiness are highlighted, while potential gaps or deviations from the original objectives are identified for future refinement.

-Model Performance:

We conducted a detailed evaluation of the performance of our loan assessment models post-implementation. Metrics such as accuracy, precision, recall, and F1-score were scrutinized to assess how well the models predict credit risk. Results were benchmarked against industry standards and discussed in terms of their operational effectiveness. Potential avenues for optimizing model performance, such as fine-tuning hyperparameters or incorporating additional data sources, are considered to enhance predictive capabilities and decision-making accuracy.

-Data Integrity and Reliability:

The reliability and integrity of the data used in training and testing the loan assessment models are crucial. We examined the robustness of data cleaning and preprocessing techniques to ensure that the models are trained on accurate and representative data. Discussion includes strategies employed to handle missing data, outlier detection, and ensuring the consistency and quality of input variables to minimize model bias and improve reliability.

-Challenges Overcome:

We reflect on the challenges encountered during the development and deployment of the loan assessment models. These include technical hurdles such as model overfitting, data heterogeneity, and the interpretability of complex machine learning algorithms in a regulatory environment. Strategies employed to overcome these challenges, such as iterative model refinement and stakeholder engagement, are outlined with insights gained and lessons learned for future model iterations.

-Future Directions and Improvements:

Looking ahead, we identify specific areas where the loan assessment models can be further enhanced. This includes exploring advanced machine learning techniques, such as ensemble methods or neural networks, to improve predictive accuracy and robustness. Enhancements in data enrichment through alternative data sources and continuous monitoring of model performance are proposed to adapt to dynamic market conditions and borrower behaviors.

-Exploration of Alternative Approaches:

Alternative approaches that could have been explored during the model development phase are considered. This involves evaluating the feasibility of different model architectures or algorithmic frameworks and their potential impact on loan assessment outcomes. Discussions on the trade-offs between model complexity, interpretability, and computational efficiency are provided, reflecting on the decision-making process and the rationale behind adopting specific methodologies.

CONCLUSIONS

"The key findings of this study are summarized, emphasizing the project's contributions to improving loan management efficiency through robust data analysis.

Model Performance:

The analysis demonstrated that the XGBoost model outperformed the neural network in overall performance for creditworthiness evaluation. This indicates that, for our dataset and specific loan evaluation objective, XGBoost provides superior predictive capability and generalization.

Overfitting Challenges:

It was identified that the neural network experienced overfitting during the submission file generation, limiting its ability to generalize beyond the training dataset. This finding underscores the importance of regularization techniques and ongoing performance monitoring to mitigate overfitting in future iterations.

Significance of Attributes: Age and employment stability, as measured by the number of days worked, emerged as critical variables in assessing loan repayment capacity. These findings highlight the need for robust and representative data to enhance the accuracy of credit assessment.

Continuous Optimization:

Continuous optimization of the XGBoost model through hyperparameter tuning and integration of new data sources, such as external evaluations, is proposed to further improve the precision and reliability of loan decisions.

Future Outlook:

For future research, exploring advanced machine learning methods and model interpretation techniques to enhance transparency and explainability of loan decisions is suggested. Additionally, assessing risks in volatile economic environments and adapting the model to changes in market conditions are crucial for ongoing development.

In conclusion, this study has explored various machine learning models, including XGBoost and neural networks, for assessing credit risk. The findings underscore the effectiveness of XGBoost in enhancing predictive accuracy and generalization in loan evaluation. Moving forward, continuous refinement of these models through advanced techniques and incorporation of diverse data sources will be crucial for further improving the reliability and efficiency of credit assessment processes in financial sectors.

REFERENCIAS

[1] Kaggle, "Home Credit Default Risk Competition". Disponible en: https://www.kaggle.com/c/home-credit-default-risk

[2] Home Credit Group, "About Home Credit Group". Disponible en: https://www.homecredit.net/

[3] TensorFlow, "TensorFlow Documentation". Disponible en: https://www.tensorflow.org/