Home Credit Default Risk Analysis

Chirag Shah, Sachin Patel, Sneha Yadav, Yash Modi
Department of Computing, Dublin City University, Ireland
Chirag.shah2@mail.dcu.ie
sachin.patel3@mail.dcu.ie
sneha.yadav4@mail.dcu.ie
yash.modi2@mail.dcu.ie

Abstract – Home Credit Group provides personal and consumer loans to help the unbanked population gain financial inclusion. However, providing credit facilities puts financial institutions and businesses at risk. The purpose of this research is to examine the default risk associated with Home Credit's lending activities. To accomplish this, the dataset provided by Home Credit is to develop a machine learning-based approach. Methodologies followed in this project include data preprocessing, feature engineering, and model development. This project aims to predict the likelihood of loan default, the methods used were two classification models: Random Forest and XGBoost. Various performance metrics, such as accuracy, precision, recall and Area under the ROC Curve score, were utilized to evaluate the models. The comparison analysis demonstrates that XGboost outperformed the Random forest model when oversampling was used, whereas the Random forest model and XGboost model performed poorly when oversampling was not used. To determine the most important predictor influencing the probability of a loan default, a feature importance analysis was also conducted. According to the research, factors including income, the number of prior loans, and credit history length are excellent indicators of default risk. Overall, this study offers insightful information about the default risk connected to Home Credit's lending activities and suggests a machine learning-based method for estimating the likelihood of a loan default.

***Keywords*: Home Credit, Default Risk, Machine Learning, XGBoost, Random Forest, Feature Engineering, Performance Metrics, binary classification.**

1. Related Work

The study of Home Credit Default Risk has gained significant attention in recent years due to the increasing prevalence of non-traditional lending models. Many studies have used machine learning techniques to develop predictive models for credit risk assessment. In this literature review, we discuss studies that have employed gradient boosting, SVM, decision trees, and other machine learning algorithms to predict the likelihood of home credit default. Mahmudi [1] evaluated different gradient-boosting algorithms for home credit risk prediction and found that LightGBM outperformed other algorithms. Similarly, [2] Shinde and Pawar used decision tree-based algorithms to predict the default risk of home credit and achieved an accuracy of 78.2%. [3]Jin and Zhu applied a data-driven approach to predict default risk for peer-to-peer lending and found that the model improved as more data was included. [4] Qiu used machine learning models to develop a credit risk scoring analysis for personal loans and found that Gradient Boosting Decision Tree (GBDT) performed the best. Li [5] proposed a GBDT-SVM model that considered audit information to assess credit risk for peer-to-peer borrowers. Pandey[7] employed machine learning classifiers to predict credit risk and found that Support Vector Machine (SVM) outperformed other algorithms. Several studies have compared the performance of different machine learning algorithms for credit risk prediction. For example, Daoud [8] compared XGBoost, LightGBM, and CatBoost algorithms and found that LightGBM outperformed the others. Similarly, [9] Davis and Freeman used evolutionary computation techniques to develop a credit-scoring model and found that it performed better than traditional neural network models. [10] Fayyad and Fisher used classification and regression trees and Multivariate Adaptive Regression Splines (MARS) to mine customer credit data and found that the model performed well in predicting credit risk. Fayyad [11] proposed the Knowledge Discovery in Databases (KDD) process for extracting useful knowledge from data, which has since become a foundation for data mining and machine learning research. These studies have shown how well machine learning algorithms can foretell the risk of home credit default. However, these models have drawbacks, like the requirement for high-quality data, potential biases in the data, and the challenge of interpreting the outcomes. Additionally, these studies mainly concentrate on how well the models predict outcomes rather than necessarily offering any insights into the underlying causes of credit risk.

The studies covered in this literature review, in summary, offer a helpful framework for creating predictive models for the risk of home credit default. Even though these models have demonstrated high accuracy, there are still some problems to consider. Future research should focus on developing clearer, simpler models that can explain the factors that affect credit risk.

2. Data Mining Methodology

Dataset - The Home Credit Application Default Risk dataset is available on the Kaggle open-source repository and it was made available as part of the Feature Prediction Competition in August 2018 with the aim to “Can you predict how capable each applicant is of repaying a loan? ”. The dataset contains application_train.csv with 122 columns and 307,511 rows and the application_test.csv file contains 121 columns and 48,745 rows. The overall size of the dataset is 2.68 GB.

Methodology - CRISP-DM methodology is a well-structured approach with the following six phases: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment. It focuses on the significance of understanding the business context and the goal of the project in the initial phase. It also recognizes that the data mining process is iterative and adaptive which enables it to revisit the earlier stages as the project progresses. As new information and facts are found, it permits adjustments and enhancements to be made. The evaluation step of CRISP-DM is crucial for determining the effectiveness and validity of the produced models. Since it is widely utilised across many different businesses and fields, including the banking sector, finance and management of risk, it is well-known and useful in a wide range of real-world situations, including the "Home Credit Default Risk Analysis".

Business Understanding - In order to appropriately categorize loan applicants as having a low or high risk of default, the Home Credit Default Risk Analysis project intends to create a predictive model. The model will help Home Credit make better lending decisions, reduce the amount of money lost due to loan defaults, and improve lending opportunities for people who might have trouble obtaining traditional bank loans by identifying the key factors that affect loan default risk. In order to accomplish this goal, the project will collect, examine, and clean data on previous loan applicants' repayment patterns, create and train a predictive model, assess the model's performance, and then implement the model into Home Credit's loan application process.

Data Understanding - The training dataset contains 122 predictors, of which 67 have

missing data because of the following reasons :
(a) Merging of source datasets
(b) Random events
(c) Failure of measurements

While going through the columns of the missing data, it was observed that the type of missing data was of the following types:
(a) Structural deficiencies
(b) Random occurrences - Missing Completely at Random (MCAR, Missing at Random (MAR)
(c) Specific causes - Not Missing at Random (MNAR)

The outcome variable comprises of binary value ie. accepted or rejected. No missing data was detected in this column. The preliminary analysis of the response indicated an imbalance dataset.

Data Preparation - The data preparation phase was the lengthiest among all phases as it involved tasks such as data cleaning, data integration, data transformation, and feature engineering. The exploratory data analysis discovered 67 columns with missing data including both the categorical and numerical columns. To prepare the data for training the machine learning model, several imputation techniques were applied to the missing predictors. While performing the imputation, the limits of computational capacity were taken into consideration. The following approach was followed to impute missing data in categorical and numerical columns.

Categorical predictors with very few missing values (< 1 percent) were imputed with mode of that feature. The nature of missing values in most of the categorical columns (eg. house type, walls material, occupation type etc) were missing at random so a new category ‘Unknown’ was introduced to impute in order to avoid bias.
One feature “emergency state” contained binary values, therefore a decision was made to prevent creation of a third ‘Unknown’ category, and instead mode imputation was used.

All the categorical predictors were checked for ordered data and only one column “weekday application process start date” was found. The ordinal column was encoded with label encoding technique.
A decision was made to encode categorical features with less than 10 categories with dummy encoding. Predictors with more than 10 categories were encoded with hash encoding.

There were 61 numerical columns with missing values. The first step was to understand the distribution of all of them. Predictors with skewed data and minimal missing values were filled with median imputation. On the other hand, attributes (eg. external data source) in which the data was already normalised and missingness was low, mean imputation was employed.
Some 40 missing predictors were found having good correlations with others. A machine learning imputation approach would be the most accurate. However, to overcome computational limitations, only top 5 correlated predictors of a predictor where the correlation was more than 0.90 were used. The **K-nearest neighbors (KNN)** inferential method was utilised to impute missing values in these columns.
There were a few columns that were not found to correlate with any other columns. So, missing values in these columns were filled with simple (mean, mode, median) imputations.
Around 8-10 predictors with very large missing values and correlation of more than 95 percent with other predictors were dropped even before imputation. Anyway, feeding heavily correlated variables to a model can do more harm than good. The predictor 'price of goods' has only 0.09% missing values with almost 99% correlation with column 'credit amount'. So, it was dropped to simplify the modelling process and to improve the performance of the predictive model. High correlation between variables can sometimes lead to

multicollinearity, which can affect the performance and interpretability of the model. By dropping one of the highly correlated variables, the model may be more stable and produce more reliable results.

After imputation, a second round of exploratory data analysis was performed to check the underlying distribution and other statistical characteristics. Skewed continuous predictors like income, loan amount, etc. can be harmful to a model and need to be dealt appropriately. All the skewed predictors were visualised again after transforming them with Box-cox transformation and log transformation. These transformations gave close to normal distribution and therefore transformed.

On analysing the data, it was also found that there were few features [age of client in days, 'current employment days', 'client registration change days', 'clients ID change in days', 'client phone number change in days'] which possess all negative values in it so these columns were signed transformed to positive values. Some of the continuous predictors were already Min-Max scaled in the dataset so the same technique was used to scale the others. These transformations are typically needed when the model requires the predictors to be in common units.

Modelling - At the end of data preparation stage, there were 157 predictors in the dataset, but all the predictors may not be relevant for modelling. Models were selected primarily on two reasons: 1) Computational constraints 2) Model interpretability.

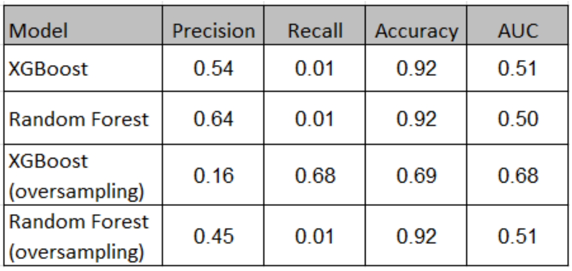
In Feature selection methodologies, intrinsic methods have feature selection naturally incorporated with the modeling process. Tree and rule-based models search for the best predictor and split point such that the outcomes are more homogeneous within each new partition. **XGBoost** and **Random forest** are ensemble methods which can handle high-dimensional data and they are appropriate for this imbalanced dataset. Initially, XGBoost and Random Forest models with no hyperparameter tuning were used.

The output reflected the disadvantages of having an imbalanced dataset. An approach was made to oversample the minority class using a data augmentation technique named **'Synthetic Minority Oversampling Technique'**. Both the models were run again but this time on a balanced dataset. After which, XGBoost model's hyperparameters (learning rate, max depth, and number of estimators) were fine tuned manually.

At last, a greedy wrapper class feature selection technique called **Recursive Feature Elimination** (RFE) was used to find the optimal subset of predictors. Permutation importance of all the predictors in the XGBoost model were analysed to give an understanding of the most and the least important predictors. The least important features were then removed from each iteration of RFE.

Evaluation/Results

Understanding precision, recall, and Area Under ROC (Receiver Operating Characteristic) curve are some of the best evaluation metrics for a binary classification problem. Following tables show the comparative performance of the XGBoost model and the Random Forest model with and without oversampling of the minority class.



Model	Precision	Recall	Accuracy	AUC
XGBoost	0.54	0.01	0.92	0.51
Random Forest	0.64	0.01	0.92	0.50
XGBoost (oversampling)	0.16	0.68	0.69	0.68
Random Forest (oversampling)	0.45	0.01	0.92	0.51

As it can be seen from the above table, both models performed poorly without oversampling with an AUC score of 0.51 (XGBoost) and 0.50 (Random Forest). The low recall rate suggests that an imbalance dataset makes it harder for models to identify True Positives (TP). Whereas, in case of oversampling the AUC score of XGBoost increased to 0.68.

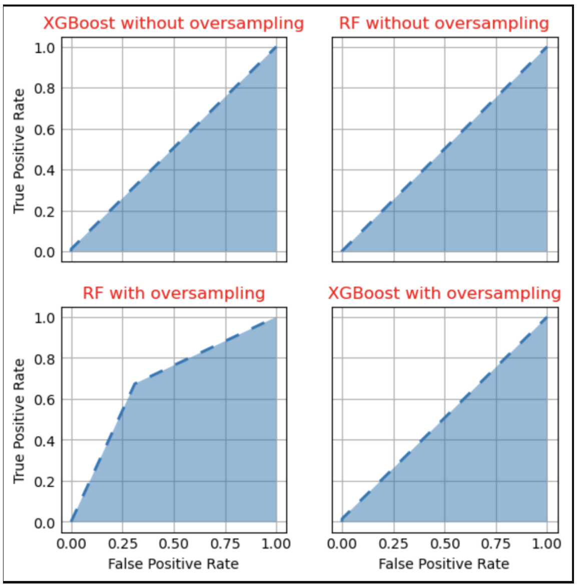


Fig 1. ROC Curve

This indicates that oversampling of the minority class increased the performance of the model. An interesting observation was the trade-off between precision and recall metrics in the XGBoost model in case of with and without oversampled datasets. No significant improvements were observed in Random Forest with oversampling. XGBoost outperformed Random Forest because of the built-in regularisation feature that prevents it from overfitting during training.

The RFE technique found 2 predictor subsets of near equal AUC score with the original list of predictors. However, extensive computations were clearly required to reach a global optimum subset of predictors.

3. Conclusions

In this project, we created a machine learning methodology to look into the default risk associated with Home Credit's lending activities. In our study, the probabilities of loan default were predicted using two classification models: Random Forest and XGBoost. Accuracy, precision, recall, and AUC score were some of the performance metrics that were used to evaluate the models. Our results showed that XGboost outperformed the Random forest model when oversampling was used, whereas when oversampling was not used, the Random forest model and XGboost both did not perform well. By performing a feature importance analysis, it was discovered that income, the volume of prior loans, and the length of credit history were all significant predictors of default risk. As a whole, our study offers insightful information about the default risk connected to Home Credit's lending activities, and we recommend a machine learning-based approach for calculating the probability of loan default. Our findings could assist financial institutions in reducing the risks involved in lending to unbanked groups of people and enabling their financial inclusion.

4. Future work

Despite producing positive results, our study has some limitations. To begin, the dataset used in this study is unbalanced, and more advanced techniques such as anomaly detection and clustering can be used to improve the models' performance. Second, our models were only trained on one dataset; it will be interesting to see how they perform on different datasets. Third, our models were unable to provide advanced insights into the underlying causes of credit risk, and more research is needed to explain the factors that influence credit risk. Finally, our models only considered demographic and financial variables of loan applicants; future studies could include social and behavioral variables to improve prediction accuracy.

Project links

1) Presentation Video youtube link
   Home Credit Defau…
2) Github link for this project
   https://github.com/chiragshah-16/Projects

5. References

1. H. Mahmudi, R. Bhargava and R. Das, "Evaluation of Gradient Boosting Algorithms on Balanced Home Credit Default Risk," 2022 International Conference on Trends in Quantum Computing and Emerging Business Technologies (TQCEBT), Pune, India, 2022, pp. 1-6, doi: 10.1109/TQCEBT54229.2022.10041584.

2. Shinde, G. and Pawar, S. (2022) International Journal for Research in Applied Science & Engineering Technology (IJRASET). Available at: https://issuu.com/ijraset/docs/home-credit_risk_analysis_and_prediction_modelling

3. Y. Jin and Y. Zhu, "A Data-Driven Approach to Predict Default Risk of Loan for Online Peer-to-Peer (P2P) Lending," 2015 Fifth International Conference on Communication Systems and Network Technologies, Gwalior, India, 2015, pp. 609-613, doi: 10.1109/CSNT.2015.25.

4. “Bank lending policy, credit scoring and value at risk.” [Online]. Available: https://pages.ucsd.edu/~aronatas/project/academic/Bank%20lending%20policy,%20credit%20scoring%20and%20value%20at%20risk.pdf.

5. Z. Qiu, Y. Li, P. Ni, and G. Li, "Credit Risk Scoring Analysis Based on Machine Learning Models," 2019 6th International Conference on Information Science and Control Engineering (ICISCE), Shanghai, China, 2019, pp. 220-224, doi: 10.1109/ICISCE48695.2019.00052.

6. Li, Z. (2018) GBDT-SVM Credit Risk Assessment Model and Empirical Analysis of Peer-to-Peer Borrowers under Consideration of Audit Information. Open Journal of Business and Management, 6, 362-372. https://doi.org/10.4236/ojbm.2018.62026

7. T. N. Pandey, A. K. Jagadev, S. K. Mohapatra and S. Dehuri, "Credit risk analysis using machine learning classifiers," 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), Chennai, India, 2017, pp. 1850-1854, doi: 10.1109/ICECDS.2017.8389769.

8. “Comparison between XGBoost, lightgbm, and CatBoost using a home credit ...” [Online]. Available: https://www.semanticscholar.org/paper/Comparison-between-XGBoost%2C-LightGBM-and-CatBoost-a-Daoud/b992fdb71b4b78d7b81dc3761402f4eb446077c2.

9. R. H. Davis and J. A. Freeman, “Credit scoring and rejected instances reassigning through evolutionary computation techniques,” Expert Systems with Applications, 20-Jan-2003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417402001914?casa_token=bzd8pbZrYA0AAAAA%3AuzYlGhPqlrIm_N0SrxooxWUfahUGUDQCD-3SwstTNkk0zakUofRJA07i_yK2s-vP7GspdDqy4tM

10. U. Fayyad and R. A. Fisher, “Mining the customer credit using classification and regression tree and multivariate adaptive regression splines,” *Computational Statistics & Data Analysis*, 07-Dec-2004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016794730400355X?casa_token=brg6v5VF8UcAAAAA%3AmyRNYjt8yJWP2z-ZSJZhz5t6uRbCcQ611OwWgtYczS16bzzcQRvT2EW2yRwOEJBni1pFMmsW824. [Accessed: 11-Apr-2023].

11. U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “The KDD process for extracting useful knowledge from volumes of data,” *Communications of the ACM*, 01-Nov-1996. [Online]. Available: https://dl.acm.org/doi/10.1145/240455.240464.

12. Data Mining. Available - https://en.wikipedia.org/wiki/Data_mining

13. What is Crisp dm? Available - https://www.datascience- pm.com/crisp-dm-2/

14. U. Fayyad and R. A. Fisher, “Mining the customer credit using classification and regression tree and multivariate adaptive regression splines,” Computational Statistics & Data Analysis, 07-Dec-2004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016794730400355X?casa_token=brg6v5VF8UcAAAAA%3AmyRNYjt8yJWP2z-ZSJZhz5t6uRbCcQ611OwWgtYczS16bzzcQRvT2EW2yRwOEJBni1pFMmsW824. [Accessed: 11-Apr-2023].

15. Data Mining. Available - https://en.wikipedia.org/wiki/Data_mining

16. Feature Engineering and Selection: A Practical Approach for Predictive Models - 10.2 Classes of Feature Selection Methodologies | Feature Engineering and Selection: A Practical Approach for Predictive Models