Home-Credit Risk Prediction using Machine Learning

Pranavan Subendiran
dept of CSE
University of Moratuwa
Colombo, Sri Lanka
subendiran.21@cse.mrt.ac.lk

Ranasinghe K.S.
dept of CSE
University of Moratuwa
Colombo, Sri Lanka
kumudh.21@cse.mrt.ac.lk

Prabashwara D.G.H.
dept of CSE
University of Moratuwa
Colombo, Sri Lanka
hansana.21@cse.mrt.ac.lk

Perera I.T.M.
dept of CSE
University of Moratuwa
Colombo, Sri Lanka
tithira.21@cse.mrt.ac.lk

Senarathna L.P.S.U.K.
dept of CSE
University of Moratuwa
Colombo, Sri Lanka
udantha.21@cse.mrt.ac.lk

***Abstract*—Predicting the credit default risk of a customer is crucial for financial organizations when offering loans. Different techniques have been used over time to predict the credibility of a borrower by employing machine learning techniques.**

Using a dataset from Home Credit, a consumer finance provider, this study predicts consumer credit risk by employing various data preprocessing methods. Two classification models, LightGBM and CatBoost, are implemented, and their strengths are combined in an ensemble model.

Performance evaluation metrics including accuracy, precision, recall, and AUC score are used to gauge model effectiveness. Additionally, avenues for future improvement such as feature selection, cross-validation strategies, and addressing class imbalance are identified.

Combination of lightGBM and CatBoost overperform those individual models in terms of the metrics we are concerned about.

***Index Terms*—Credit Risk Assessment, Machine Learning, LightGBM, CatBoost, SMOTE, Ensemble, SHAP**

I. Introduction and Background

This paper focuses on the application of machine learning techniques in the prediction of credit risk within the context of "Home Credit - Credit Risk Model Stability" competition hosted in 'Kaggle' platform [5].

Credit risk is defined as "the probability of a financial loss resulting from a borrower's failure to repay a loan. Essentially, credit risk refers to the risk that a lender may not receive the owed principal and interest, which results in an interruption of cash flows and increased costs for collection" [1]. Thus, credit risk assessment is crucial for consumer finance providers, allowing them to determine the likelihood of loan repayment and make informed lending decisions.

There are different types of credit risks, some of which are fraud risks, default risks, counter-party risks, credit spread risks and concentration risks. The competition has given its main focus on default risks, where the borrower might not be able to meet their loan obligations and pay the agreed-upon amount in the loan contract.

Various techniques are used to assess a borrower's creditworthiness, including:

- Credit History Analysis: Reviewing past borrowing behavior, repayment history, and outstanding debts.
- Credit Scores: Numerical representations of creditworthiness based on credit history.
- Debt-to-Income Ratio (DTI): Comparing monthly debt obligations to gross monthly income.
- Financial Statements: Analyzing income statements and balance sheets for larger loans or businesses to assess profitability and financial health.

It is evident from the descriptions that these methods heavily rely on the credit history of the borrower. But not in all instances that a person has a long credit history. An individual may not have a credit history due to reasons including young age or a preference for cash. Under traditional approaches they are likely to be denied. This makes individuals that would highly benefited with credit access denied of the opportunity.

With the digitalization of processes, machine learning and artificial intelligence techniques have begun to be incorporated into credit risk analysis. Currently, consumer finance providers use various statistical and machine learning methods to predict loan risk. These models are generally called scorecards. A scorecard is essentially a statistical model that combines various factors about a borrower to predict the likelihood of them repaying a loan.

A scorecard system offers benefits like consistency and accuracy, but maintaining effectiveness over time poses a significant challenge. Client behavior and economic conditions are constantly changing, requiring regular updates to scorecards. Unfortunately, the process of model redevelopment, validation, and implementation is resource-intensive and time-consuming. Ensuring scorecard stability is crucial to prevent deterioration in performance, which can lead to inadvertent approval of loans to higher-risk borrowers and increased defaults. Striking

a balance between accurate loan risk prediction and sustained effectiveness in the future is key for lenders.

The competition focuses on the application of machine learning techniques to evaluate the default risk of potential customers, even those with limited or no credit history, while ensuring model stability over time.

II. Related Work

In recent years, several studies have focused on credit risk assessment using various data analysis and machine learning techniques. These studies have explored the basic information of personalities, tax information, credit history and financial activities (deposits, account information). Here, we present a summary of some relevant related work in this area:

1) **Credit Risk Assessment in P2P Lending Using LightGBM and Particle Swarm Optimization**: Yosza Dasril a, Much Aziz Muslim b, M. Faris Al Hakim c, Jumanto d, Budi Prasetiyo e published a journal on predicting credit risk using lightGBM model and Particle Swarm Optimization in peer to peer lending platform. The highest accuracy also presented satisfactory results with 98.094% accuracy, 90.514% Recall, and 97.754% NPV, respectively. The combination of LightGBM and PSO had resulted in better outcome. [2]

2) **Machine Learning for Credit Risk Prediction: A Systematic Literature Review**: Noriega, J.P., Rivera, L.A. and Herrera, J.A. published an article that provides an insightful overview of utilizing Machine Learning in credit risk prediction, emphasizing its critical importance for financial institutions. It identifies key areas such as algorithm selection, evaluation metrics, dataset usage, and prevailing challenges in credit risk assessment. Notable findings include the dominance of Boosted Category ML models and the widespread adoption of metrics like AUC and ACC for evaluation. [3]

3) **A study on predicting loan default based on the random forest algorithm**: L. Zhu , D. Qiu, D. Ergu, C. Ying, and K. Liu pulished a paper in IQTM 2019 discussing an approach of load default prediction using random forest algorithm. They followed usual machine learning processes and used SMOTE to handle the class imbalance. Finally, their random forest model achieved 98% accuracy. [4]

While the above studies have made significant contributions to the field of credit risk prediction, our study differs in terms of the specific approach of building models. We have built an Ensemble model using LightGBM and CatBoost models by averaging out the results obtained by them to get a generalized result that closely follows the actual trend in credit defaults.

III. Method

A. Data Collection

For our study we used the dataset provided in the competition, which was collected by competition hosts Home Credit. It is an international consumer finance provider focusing on responsible lending primarily to people with little or no credit history. The dataset includes a large number of files capturing different information which will aid in predicting the credit risk of a consumer.

B. Data Preprocessing

The dataset contains several instances of data quality deficiencies. Initially, data instances within certain dataset tables were fragmented across multiple files. Furthermore, certain data tables needed aggregation to consolidate relevant information. A significant proportion of columns contained a high number of missing values. Additionally, there were instances of attribute duplication with distinct column names. Moreover, numerous attributes featured masked original data. We approached the solutions for these problems via following stages:

1) **Data Integration**: The provided data is significantly large so that merging them into a one data frame and processing the data frame was impossible because of memory constraints. To address this issue, a preprocessing approach was adopted, where each table underwent preprocessing prior to merging with other tables.
   First, data was loaded from the files while specifying some additional strings to indicate null values. Next, tables with depth more than 0 were aggregated by taking max value, mean value and last value. Further, if the data is divided into separate files, those files were loaded, aggregated and merged together into a single data frame. Furthermore, attributes with a missing value percentage of more than 50% and imbalanced attributes that has a single value for more than 80% of its data instances were removed to handle the biases. Additionally, the attribute data types were transformed to have better memory efficiency. At last, this data frame is merged with the other data frames attribute-wise.
2) **Data Cleaning**: Although the data frames are merged into one, the number of data instances in each data frame was not always the same. As a result, a significant amount of missing values were caused in the combined data frame. To handle this issue, again attributes with a missing value percentage of more than 50% were removed and then all attributes were imputed.
3) **Data Reduction**: For attribute reduction correlation coefficients among attributes were observed and attributes with high correlation were removed from the combined data frame.
4) **Data Transformation**: Initial data transformations were handled during the data integration. Numerical attributes were converted into numerical data types. Date attributes

were transformed into timestamps. Further, data types of attributes were transformed into sub-data types to reduce memory usage. All columns containing strings or objects were transformed into categorical data types. Later these categorical data type attributes were converted into string data type to train the CatBoost Classifier. Further, the class imbalance was handled using random down-sampling.

C. Models

The credit risk of a consumer in the training dataset is given as a binary value. Hence we employed a classification model for the training process. To increase the accuracy of the model we built an ensemble model using two different classification models.

1) ***LightGBM Classifier***: LightGBM is a gradient-boosting framework that uses tree-based learning algorithms. It is designed for distributed and efficient training of large-scale datasets and offers high performance in terms of both faster convergence and improved accuracy.
2) ***CatBoost Classifier***: CatBoost is a gradient-boosting library specifically optimized for categorical features. It stands out for its ability to handle categorical variables directly without the need for one-hot encoding or pre-processing. This approach reduces overfitting and improves the model's generalization performance

D. Model Evaluation

To assess the performance of our regression models we calculated several evaluation metrics. These include Accuracy, Precision, Recall, F1-score, Confusion Matrix, ROC-AUC score, Mean Absolute Error (MAE), Root Mean Squared Error(RMSE) and R-squared(Coefficient of Determination). These metrics allowed us to compare the accuracy and predictive capabilities of models in different folds during cross-validation to select the best models.

E. Cross Validation

To ensure the robustness of our results and avoid overfitting we used StratifiedGroupKFold cross-validation. It is a cross-validation method that combines stratified sampling and group-based splitting. Stratified sampling ensures the same distribution of target classes as the whole dataset, which is useful for imbalanced datasets like this dataset. Group-based splitting helps in preventing data leakage in scenarios where samples are correlated within groups.

F. Model Selection

Through cross-validation, we obtain the best-performing model for each classifier through the different evaluation metrics employed. Instead of selecting one of these models, we will build an ensemble model using these two models. Under the ensemble model, the individual prediction probability of each model was taken and using the average of those probabilities, the final prediction was determined.

G. Interpretation of results

Based on the final results we can interpret the relationship between different measurement criteria and their implications on the credit risk. By further research through the findings we obtain here, we can improve the lives of people who have historically been denied due to a lack of credit history.

IV. Results

The performance of the models were evaluated using a variety of metrics as mentioned earlier. Given below are the results for the different metrics for the best-performing models.

TABLE I
PERFORMANCE METRICS FOR LIGHTGBM, CATBOOST AND ENSEMBLE MODELS

Metrics	***LightGBM***	***CatBoost***	***Ensemble***
Accuracy	0.7563	0.7495	0.7551
Precision	0.76	0.75	0.76
Recall	0.76	0.75	0.76
AUC Score	0.8353	0.8293	0.8323
MAE	0.3382	0.3649	0.3649
RMSE	0.4067	0.4133	0.4099
R-squared	0.3382	0.3164	0.3276

V. Discussion

In this segment, we aim to comprehensively analyze and interpret the findings derived from our investigation into predicting which clients are more likely to default on their loans through the analysis of their loan, bank interaction, tax and credit information. Furthermore, we will delve into potential avenues for enhancing the predictive models and mitigating any constraints encountered.

A. Discussion and Interpretation of Results

The results obtained revealed that both models perform consistently well across all the evaluation metrics with the LightGBM model slightly outperforming the CatBoost model. This is understandable as we employed the LightGBM model for its high performance and accuracy while the CatBoost model was there for reducing overfitting and generalizing the results. When we combine both these models into the Ensemble model we can see the performance being similarly consistent. With the individual capabilities of the LightGBM and CatBoost models the Ensemble model becomes far more superior with its performance and accuracy while also being generalized.

B. Possible Improvements

1) ***Feature Selection***: Using techniques like forward selection, backward elimination, or methods such as tree-based approaches and Lasso can help us pick out the most important features from our data. This means we focus on the key factors that actually matter for our prediction and ignore the rest. Doing this makes

our models work better, prevents them from getting too complex, and reduce the computational resource requirement of the model.

2) ***Cross validation***: Applying rolling cross-validation instead of stratified k-fold cross-validation can be a better approach when dealing with this dataset as it contains time attributes. Rolling cross-validation takes into account the temporal nature of the data, making the model evaluation more realistic and robust over time.
3) ***Class Imbalance***: Oversampling techniques can be performed instead of random undersampling. Even though using random oversampling is possible, using other techniques like SMOTE (Synthetic Minority Over-sampling Technique) is impossible as some attributes are not encoded. Oversampling prevents the loss of valuable information as it preserves the original distribution of the majority class. However, this can lead to higher model training times.
4) ***Ensemble Methods***: Ensemble models, which combine the strengths of multiple models, offer several advantages over individual models. They often result in better predictive performance, increased robustness to outliers and unseen data, and the ability to capture diverse aspects of the data by incorporating various learning algorithms. However, using ensemble models can also introduce complexity, making interpretation more challenging.
5) ***Interpretability***: Having the ability to interpret the model predictions is crucial in applications like credit risk management. Even though it is hard to interpret ensemble models, we can consider SHAP (SHapley Additive exPlanations) values or LIME (Local Interpretable Model-Agnostic Explanation) to explain the model. Also both models used for ensembleing provide feature importance scores and decision trees to understand the model. Other than that we can consider partial dependence plots to understand the dependency between the model predictions and considered features.

C. Ethical Considerations

When using predictive models to assess credit risk, it's important to think about fairness. Sometimes, these models can unintentionally treat certain groups unfairly, especially those who are less represented in the data. For example, if the model doesn't have enough information about people in a particular group, it might make decisions that aren't fair to them.

To make sure everyone is treated fairly, we need to be aware of these biases and take steps to fix them. This might mean using special techniques or tools to make sure the model doesn't unfairly disadvantage certain groups. It's also important to keep checking how the model is working to catch any problems early on.

By being careful and making sure our models are fair, we can help make sure everyone has equal access to financial opportunities and that nobody gets treated unfairly.

VI. Conclusion

In conclusion, our study has provided valuable insights into the domain of credit risk assessment using machine learning techniques within the context of "Home Credit - Credit Risk Model Stability" competition hosted on 'Kaggle' platform [5]. Through the utilization of machine learning algorithms such as LightGBM and CatBoost, we have demonstrated a robust methodology for evaluating the default risk of potential customers while ensuring model stability over time.

Our analysis showcased the effectiveness of ensemble modeling, combining the strengths of multiple classifiers to enhance predictive accuracy and generalization. By leveraging the individual capabilities of LightGBM and CatBoost, we achieved consistent performance across various evaluation metrics, demonstrating the utility of ensemble approaches in credit risk assessment.

Furthermore, we identified several avenues for potential improvements in future research endeavors. These include feature selection techniques to focus on key predictive factors, adopting rolling cross-validation to account for temporal data dynamics, addressing class imbalance through advanced oversampling techniques, and enhancing model interpretability through SHAP values, LIME, and partial dependence plots.

Ethical considerations were also emphasized, highlighting the importance of fairness in predictive modeling to ensure equal access to financial opportunities and mitigate unintended biases.

Overall, our study contributes to advancing the field of credit risk assessment by providing a comprehensive framework for building stable and accurate predictive models, with the potential to improve lending decisions and promote financial inclusivity. Through continued research and refinement of methodologies, we aim to further enhance the effectiveness and fairness of credit risk assessment models, ultimately benefiting individuals and financial institutions alike.

References

[1] T. Brock, "What is credit risk?," Investopedia, Aug. 15, 2023. https://www.investopedia.com/terms/c/creditrisk.asp (accessed May 10, 2024).

[2] Y. Dasril, M. A. Muslim, M. F. A. Hakim, J. Jumanto, and B. Prasetiyo, "Credit risk assessment in P2P lending using lightgbm and particle swarm optimization," Register, https://journal.unipdu.ac.id/index.php/register/article/view/3060 (accessed May 10, 2024).

[3] J. P. Noriega, L. A. Rivera, and J. A. Herrera, "Machine Learning for Credit Risk Prediction: A systematic literature review," MDPI, https://www.mdpi.com/2306-5729/8/11/169 (accessed May 10, 2024).

[4] L. Zhu , D. Qiu, D. Ergu, C. Ying, and K. Liu , "A study on predicting loan default based on the random forest algorithm," Procedia Computer Science, https://www.sciencedirect.com/science/article/pii/S1877050919320277 (accessed May 10, 2024).

[5] Daniel Herman, Tomas Jelinek, Walter Reade, Maggie Demkin, Addison Howard. (2024). Home Credit - Credit Risk Model Stability. Kaggle. https://kaggle.com/competitions/home-credit-credit-risk-model-stability