

CS 477

Spring 2024

Jishnuvardhan Karpuram

Computer Science

**Introduction**

Credit scoring is a pivotal process in the banking sector, essential for evaluating the creditworthiness of potential borrowers. Traditional approaches to credit scoring typically involve the application of classical machine learning models, such as logistic regression and gradient boosting, on tabular data containing various features related to the borrower's credit history. These features encompass a wide range of financial and personal information, providing insights into the borrower's credit behavior and repayment capabilities.

In recent years, there has been growing interest in exploring alternative methodologies to enhance credit scoring accuracy and predictive performance. One such approach involves leveraging sequential data, such as transactional customer histories, in its raw form. By doing so, the model can capture temporal dependencies and patterns inherent in the sequential data, potentially leading to improved predictive capabilities.

In this project, we aim to delve into the realm of credit scoring using machine learning techniques including logistic regression and ensemble methods like Random Forest and XGBoost, specifically focusing on the application of traditional models to a comprehensive dataset. This dataset comprises a wide array of features pertinent to credit products and borrower characteristics, providing valuable insights into the credit lifecycle and repayment behavior. Features include credit opening dates, planned and actual closing dates, outstanding loan amounts, delinquency records, and various indicators of credit utilization and overdue debt.

In this report, We implemented and evaluate the performance of different machine learning models, starting with logistic regression as a baseline, followed by Random Forest with parameter tuning. Finally, we delve into the implementation of XGBoost to improve the performance with that of the other machine learning models.

**2. Methods**

**Baseline mode: Logistic Regression:**

Logistic regression operates on the principle of fitting a linear relationship between the input features and the log-odds of the binary response variable. In essence, it models the probability of an event occurring given the input features. During training, logistic regression estimates the coefficients for each input feature, which represent the strength and direction of the association with the target variable. These coefficients, when multiplied by the corresponding feature values and summed together, form the linear predictor. This linear predictor is then transformed using the logistic function, mapping it to a probability between 0 and 1. The decision boundary, often set at 0.5, separates the two classes based on their predicted probabilities. The coefficients obtained from logistic regression provide direct insights into the impact of each feature on the predicted probability, making it easy to understand and interpret the model's predictions.

The accuracy of the logistic regression is 96.96%. The score of this model is biased towards class-0, and all the examples are classified as class-0. This is happening due to class imbalance in the data. By improving the model by handling the class imbalance problem using under sampling approach, the accuracy falls down to 49.88%.

**Random Forest:**

Random Forest is an ensemble learning method that aggregates the predictions of multiple decision trees to make final predictions. Each decision tree in the forest is trained on a bootstrap sample of the training data, and at each split, only a random subset of features is considered. This introduces randomness and diversity among the trees, reducing overfitting and improving generalization. During prediction, the final prediction is determined by aggregating the predictions of all individual trees, typically using majority voting for classification tasks. Random Forest is highly robust and can capture complex nonlinear relationships in the data due to its ensemble nature and feature randomization.

Random Forest is highly robust to overfitting and noise in the data, making it suitable for a wide range of tasks. It can handle high-dimensional datasets with ease and provides measures of feature importance, allowing analysts to identify the most influential features in the prediction process. However, Random Forest may suffer from high computational complexity and memory requirements, especially when dealing with large datasets or a large number of trees in the forest. The accuracy of the model is 91.77%

**XGBoost (Extreme Gradient Boosting):**

XGBoost is an optimized implementation of gradient boosting, a powerful ensemble learning technique that builds a strong predictive model by sequentially adding weak learners to the ensemble. XGBoost is designed to be highly scalable and efficient, making it suitable for large-scale machine learning tasks. During training, XGBoost fits decision trees to the residuals of the previous predictions, optimizing a specified loss function at each step. It incorporates regularization techniques to prevent overfitting and tree pruning to control the complexity of individual trees. XGBoost is known for its state-of-the-art performance on various machine learning tasks and competitions, often outperforming other algorithms.

XGBoost is highly flexible and customizable, supporting a wide range of loss functions and regularization techniques. It provides measures of feature importance, allowing analysts to understand the contribution of each feature to the model's predictions. XGBoost is optimized for parallel processing, enabling efficient training on multi-core CPUs and distributed computing environments. However, XGBoost may require careful tuning of hyperparameters to achieve optimal performance, and its interpretability may be lower compared to simpler models like logistic regression. The accuracy of the model is 63.14%.

**Experiments and Results**

|  |  |
| --- | --- |
| Method | Accuracy |
| Logistic regression | 96.96% |
| Random forest | 91.77% |
| Extreme gradient boosting | 60.29% |

Conclusion

Logistic regression is used as a baseline model to is showing the better accuracy, but the data is found to have the class imbalance problem biasing towards the class label “0”. The accuracy of the model is 96% without handling the class imbalance problem. But by using the under sampling method the class imbalance problem is solved but the accuracy is reduced to 49.88%. The accuracy by the other methods also replicated the same issues. So, in future work different approaches can be investigated to address the issue of class imbalance and to improve the accuracy of the model.

References

1. Jones, S., et al , An empirical evaluation of the performance of binary classifiers in the prediction of credit ratings changes. Journal of Banking & Finance, Vol 56, 72-85, 2015.
2. Khandani, et al,. Consumer credit-risk models via machine learning algorithms. Journal of Banking & Finance, 34(11), 2767-2787, 2010.
3. Lai, L, Loan default prediction with machine learning techniques. In: 2020 International Conference on Computer Communication and Network Security (CCNS). published by IEEE pp. 5–9. 2020

.