IJSRA

International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

Int. J. Sci. Res. Arch.
IJSRA
International Journal of Science and Research Archive
Research Journal Archive, INDIA

(Review Article)

Check for updates

Explainable AI for dynamic ensemble models in high-stakes decision-making

Nishant Gadde [1], Avaneesh Mohapatra [2, *], Dheeraj Tallapragada [3], Karan Mody [4], Navnit Vijay [5] And Amar Gottumukhala [6]

[1] *Jordan High School, Fulshear, Texas, United States.*
[2] *Georgia Institute of Technology, Atlanta, Georgia, United States.*
[3] *Dublin High School. Dublin, California, United States.*
[4] *Irvington High School. Fremont, California, United States.*
[5] *Acellus High School. Atlanta, Georgia, United States.*
[6] *Heritage High School. Frisco, Texas, United States.*

International Journal of Science and Research Archive, 2024, 13(02), 1170–1176

Publication history: Received on 21 September 2024; revised on 10 November 2024; accepted on 13 November 2024

Article DOI: https://doi.org/10.30574/ijsra.2024.13.2.2091

Abstract

AI is replacing human decisions with algorithmic ones in finance, healthcare, and justice, among other high-stake domains. We develop in this study the use of ensemble machine learning models in predictions related to credit default risks using two different datasets of credit records and application records. We achieve more robust predictions by using a combination of Random Forest, Gradient Boosting, and Decision Tree classifiers that use soft voting. Probably one of the most significant issues discussed in this research is class imbalance; defaults constitute only a really small fraction of cases compared to non-defaults in both datasets. In this respect, the Synthetic Minority Oversampling Technique was applied for balancing the classes by artificially creating synthetic samples of the minority class to have reasonably balanced training data. Ensemble model gave a respectable score of 0.88 on the ROC-AUC against the credit record dataset, where risk classification was dependent upon the 'STATUS' field. Applicants were divided into high-risk or low-risk candidates in case of an application record dataset, assuming income thresholds, and yielded an ideal score of 1.00 on the ROC-AUC. Precision-recall and ROC curves underlined the best class differentiation provided by models considering the imbalanced datasets. While these performances may have been quite strong for the ensemble models, the nature of the models is black-box, which can be an issue in financial domains where transparency is very much expected along with accountability. To this end, we suggest the inclusions of XAI techniques in future work: dynamically and in real time, explanations of model decisions. It will be certain that stakeholders have a belief not only in the accuracy of AI-driven decisions but also in the rationale behind them-developing trust and enhancing interpretability for credit risk predictions.

Keywords: AI; High-stakes decision-making; Human decisions; finance; Healthcare

1. Introduction

Artificial intelligence has emerged as an indispensable tool in many sectors, especially those sectors that have a high level of interest, such as finance, healthcare, and criminal justice. These sectors are determined by decisions made quickly and accurately, which places the adoption of machine learning models at the forefront for risk management and prediction tasks. In this respect, the ensemble models have become very powerful in predicting financial credit risk, combining multiple classifiers to strengthen the predictive power. This may include a combination of predictions from different algorithms; therefore, the models of ensembles are very prone to overfitting and generalize much better. The

[*] Corresponding author: Avaneesh Mohapatra

Copyright © 2024 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution Liscense 4.0.

examples include Random Forest, Gradient Boosting, and Decision Trees, which apply to complex outcomes such as those related to loan defaults.

However, one of the persistent challenges in machine learning is how to deal with imbalanced datasets, which are very frequent in real-world financial data. The credit risk dataset often contains far fewer cases of default as opposed to non-default cases, due to which predictive modeling suffers from biased outputs if appropriate considerations are not given. The Synthetic Minority Oversampling Technique solves the problem of one of the most usable solutions, that is, artificially increasing the representation by generating synthetic samples from the minority class. In many instances, SMOTE has been found to be quite effective in rebalancing the datasets and improving the performance of several models that effectively predict rare events such as loan defaults. It was also proved that in the case of an imbalanced classification problem, predictive accuracy can be more improved by using SMOTE with ensemble models. Wongvorachan, He, & Bulut, 2023

Despite having very good predictive power, an important weakness of these models is the lack of interpretability. These systems, sometimes called "black-box" models, do not shed much light on how decisions derive there. This becomes a problem when the stakes are high-for instance, in finance-people need to understand and have confidence in the logic behind AI-driven decisions. As a result, Explainable AI methods have been developed, which allow for clear and understandable descriptions of model predictions. XAI techniques create more transparent and interpretable machine learning models to help decision-makers understand underlying factors driving predictions with the aim of engendering trust in AI applications. This is evident in a study by Rudin (2018).

This paper will look into how ensemble models and SMOTE could be applied in credit risk prediction to also handle class imbalance issues and enhance transparency in the models developed with the implementation of XAI techniques.

2. Literature Review

In the last few years, ensemble models have become one of the best approaches to credit risk prediction, outperforming classic machine learning algorithms by combining the strengths of multiple classifiers. The RandomForest, GradientBoosting, and DecisionTree classifiers, due to their robustness and adaptiveness to complex datasets, have gained extensive applications in risk management. Liu et al. (2023) present that the usage of ensemble models in credit risk prediction can achieve higher accuracy and AUC scores, especially while working with an imbalanced dataset. According to them, while using traditional models, ensemble techniques like bagging and boosting substantially lower the bias and variance, which effectively enhances overall model performance.

One of the major problems in credit risk modeling is class imbalance, when high-risk classes are underrepresented within the dataset. Synthetic Minority Over-sampling Technique has recently emerged as the standard approach to deal with this issue. As Fernández et al. (2018) point out, the SMOTE technique creates artificial samples for the minority class, thereby enhancing the strength of a classifier in detecting high-risk cases. This approach enhances the precision and recall when combined with ensemble models. It has been proven in several works, such as Wei et al. (2022), that the integration of SMOTE within RandomForest and GradientBoosting classifiers is effective in detecting high-risk financial profiles.

Explainable Artificial Intelligence, XAI, will be another important development in the credit risk prediction area. Given the "black-box" nature of most machine learning models, there is an increasing need for transparent and interpretable models across many high-stakes domains, such as finance. Rudin (2018) states that dependence on black-box models on which vital decision-making processes rely is problematic since clear outputs are needed by stakeholders to instill confidence in model predictions. Recent efforts at integrating XAI into ensemble models, as explored by Kanaparthi 2013, have shown real-time explanations of model decisions. This holds more value in certain applications, such as loan approval and fraud detection.

Other hybrid models have also emerged that further the predictive power by fusing deep learning architectures such as CNN into traditional ensemble methods. Li et al. (2023) also present a hybrid model of CNN with LSTM, combined with mechanisms of attention; indeed, these have been among those that show improved performance in credit risk prediction, especially in temporal financial statement data analysis. This hybrid approach allows the model to grasp intricate patterns of time-series data for better accuracy in the evaluation of risk.

Therefore, to a large degree, literature supports the efficiency of ensemble models in credit risk prediction, especially in class imbalance with techniques such as SMOTE. Integration of XAI and hybrid deep learning models hence present milestones in both interpretability and predictive power of the models, and can thus be considered perfect tools in financial risk management.

3. Methodology

Two credit risk prediction datasets were employed in this research: Application Record Dataset, which includes various features of the loan applicants, like income, employment status, and family details. This is accompanied by the Credit Record Dataset, a dataset showing the credit record over time for every individual in relation to different states of repayment. These have been preprocessed to make them ready for machine learning. The target variable Class was created within the Application Record Dataset to classify applicants based on their AMT_INCOME_TOTAL: applicants with an income below $50,000 were classified as high-risk (Class 1) and applicants with a higher income were classified as low-risk (Class 0). In the case of the Credit Record Dataset, binary risk classes were mapped onto the STATUS field representing repayment history. Values between 0, 1, and 2 were considered low risk and formed Class 0, while values of 3, 4, and 5 came into the category of high risk, forming Class 1. Non-credit status records (C, X) were rejected. Missing target values cleaning involved removing the rows of missing data so that data was complete and consistent for model training.

Data Preprocessing: In order to prepare the data to be fed as input into machine learning models, encoding was done using pd.get_dummies(). All the datasets, after preprocessing, were then divided into training and test sets, respectively, in a 70-30 split. The training set consisted of 70% of the data, while 30% was used for testing. The function used was the train_test_split from Scikit-learn; stratified sampling was performed to ensure that the proportions of the high-risk and low-risk classes were maintained in both the training and test datasets. Feature scaling was done to bring the features to one standard scale, and for this, StandardScaler from Scikit-learn was used. That way, all features would be on the same scale, without any single feature dominating over the rest during model training.

Ensemble learning was performed to improve the predictive performance. The three classifiers included were RandomForestClassifier, GradientBoostingClassifier, and DecisionTreeClassifier. These classifiers were combined using VotingClassifier, along with the soft-voting strategy. In this approach, the forecasted probabilities of each classifier were averaged, and the class with the highest average probability was selected as the final prediction. This ensembling method enabled the study to capitalize on the strengths of each single model while mitigating their respective weaknesses, thereby improving overall model performance.

Another key challenge with these datasets was class imbalance. Both the Application Record and Credit Record datasets had a significant imbalance between high-risk and low-risk instances, where the low-risk class was overrepresented. This was dealt with using the Synthetic Minority Over-sampling Technique. SMOTE generates synthetic examples of the minority class for balancing. It allows the models to learn the features of high-risk cases more preferably by oversampling the minority class; hence, improving their capabilities in predicting the same with much better precision.

Model performance evaluation was carried out based on different metrics, considering different performance aspects for the various ensemble models developed. The classification report provides a metric for Precision, Recall, and F1-score while giving insight into performance by both high-risk and low-risk class. Finally, the capability of distinguishing between classes in the models was expressed with the ROC-AUC score, while precision-recall curves were plotted to evaluate the trade-off between precision and recall. These metrics were important in assessing how well these models performed in the minority class' prediction (high-risk class). ROC curves are also plotted to show the comparative performance of the models based on the false positive rate versus the true positive rate. In that respect, AUC is leveraged as a summary statistic.

After applying SMOTE to balance the datasets, the ensemble models were retrained on this resampled data. This is crucial for improving the performance of the model overall and most importantly improving its performance on identifying those cases that, if classified as high risk, had been underrepresented. Reevaluate the retrained models on test data and compare their results to the initial models that had been trained using imbalanced data.

Finally, different visualizations were done to interpret the results of the models better. First, precision-recall curves were plotted to show the relationship between precision and recall. ROC curves were plotted to provide the balance of the model's trade-offs between false positive and true positive rates. These plots gave a better view of models' performance, especially in differentiating high-risk people from low-risk ones.

The following Python libraries were used to implement this research: Pandas for data manipulation and NumPy for numerical computations, scikit-learn played a central role in the machine learning pipeline, providing tools to build, evaluate, and preprocess the models; the class imbalance was tackled using the Imbalanced-Learn library to apply SMOTE, while Matplotlib and Seaborn were used to create the precision-recall and ROC curves.

4. Results

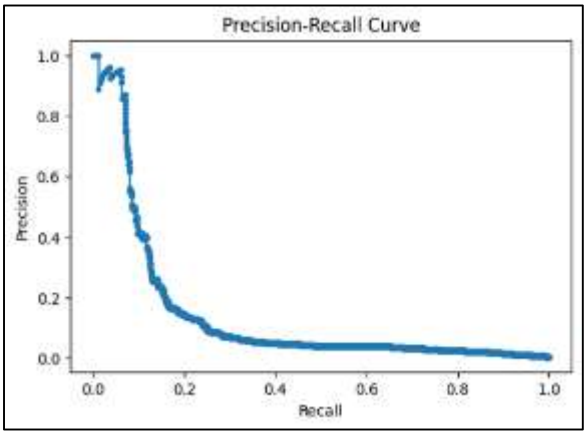

Figure 1 Credit Record Precision-Recall Curve

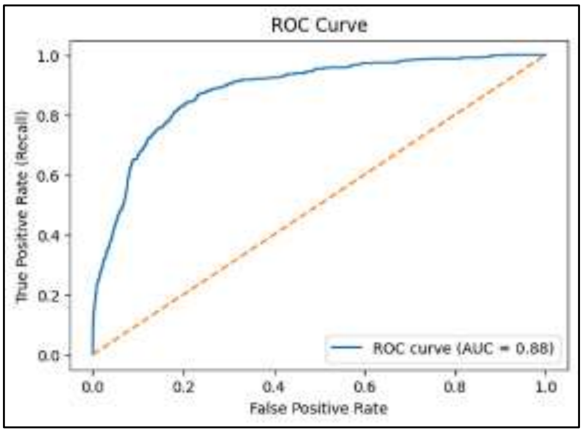

Figure 2 Credit Record ROC Curve

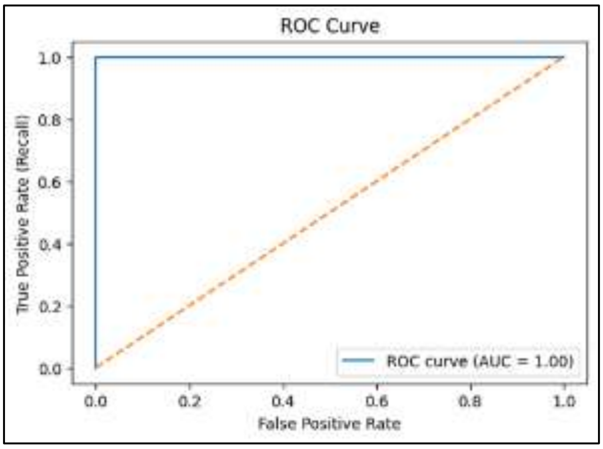

Figure 3 Application Record ROC Curve

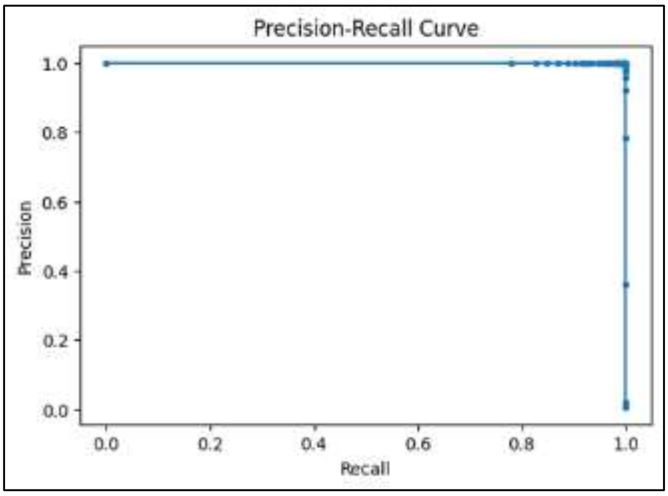

Figure 4 Credit Record Precision-Recall Curve

After the risk mapping process was done using the STATUS field, the dataset of the Credit Record Dataset contained 395,078 low-risk samples and 2,236 high-risk samples. Subsequently, after splitting the data into the training set and test set, an ensemble model containing Random Forest Classifier, Gradient Boosting Classifier, and Decision Tree Classifier was used to predict the credit risks. The classification report of the first run had an excellent performance, with overall accuracy being 1.00 for both low and high risks. However, after balancing the classes using SMOTE, the classification report on the resampled data showed poor performance for the high-risk class. The precision and recall of Class 1, which represents high-risk patients, significantly deteriorated, hence indicating the difficulties in classifying subjects in the high-risk category after resampling. More precisely, the pre-resampling classification results reported that in Class 0 (low risk), the precision, recall, and F1-score were all 1.00, while for Class 1, the precision was 0.91 and recall 0.62, with an F1-score of 0.74. These contributed to a macro-average F1-score of 0.87. Using SMOTE, Class 0 retained high precision at 1.00 and recall at 0.91, with a slight decrease in the F1-score to 0.95. However, the Class 1 precision decreased all the way to 0.04, while the recall increased to 0.65. The overall F1-score macro-average score turned out to be 0.51, while the overall accuracy was 0.91.

The score for the ROC-AUC on the credit record dataset was 0.88, therefore considered to reflect moderate performance of the model in classifying between low-risk and high-risk classes. From Figure 1 (Precision-Recall curve) down to Figure 2 (ROC curve), the performance is further highlighted, with the latter indicating an AUC score of 0.88, which basically underlines the ensemble model to predict credit risk but with many difficulties in identifying high-risk people.

4.1. Application Record Dataset Results

In this Application Record Dataset, there are 435,516 low-risk and 3,041 high-risk samples. In this, the target variable 'Class' was created based on an income threshold where applicants earning less than $50,000 were categorized as high risk. Similar to the credit record dataset, here also an ensemble model using RandomForest, GradientBoosting, and DecisionTree classifiers was applied in order to predict the likelihood of default. The model behaved very well, reaching an accuracy of 1.00 for both the classes: low risk and high risk. From the classification report, one can notice that both Class 0 (Low Risk) and Class 1 (High Risk) receive perfect precision, recall, and F1 scores. What is interesting is that SMOTE resampling didn't grossly alter the model performance, which may imply that the original model was doing quite a good job in identifying high-risk versus low-risk applicants even from an imbalanced dataset.

For the dataset of application records, the ROC-AUC had a perfect score of 1.00, demonstrating that the model distinguished between the high-risk and low-risk classes flawlessly. Figure 3 ROC curve showing the AUC of 1.00 is further reinforced by Figure 4 Precision-Recall curve, further strengthening the predictive power this model carries. This consistency of high performance across all the metrics besides suggesting the suitability of the ensemble model for risk predictions across the application record dataset.

5. Discussion of the results

The performance of the ensemble models over the two datasets was overall very good, although the correctness for the low-risk class was particularly high. While the model was working very well for the high-risk class in the application record dataset, it was somehow worse for the credit record dataset, especially after resampling with SMOTE. Resampling improved recall at the expense of far lower precision for the high-risk class in the credit record dataset, indicating a trade-off between the identification of more high-risk individuals against a possible increase in false positives. On the contrary, metrics in the case of the application record dataset were all quite high, even without resampling, thus showing the strength of the model when dealing with less challenging data.

It follows that the performance of the models with precision-recall and ROC curves visualizes the efficiency in the models developed to predict credit risk. The values of AUC indeed further support high accuracy in forecasting in low-risk and high-risk categories to 0.88 in the case of the credit record dataset and a perfect AUC of 1.00 in the application record dataset.

5.1. Discussion

The results of this study have proven that ensemble models can be applied in credit risk prediction, especially for situations requiring the differentiation of high-risk and low-risk individuals. An effective integration of RandomForest, GradientBoosting, and DecisionTree classifiers allowed for substantial overall performance, particularly in predictions of the low-risk class. However, the credit record dataset experienced typical problems with imbalanced data, whereby the high-risk class was significantly underrepresented. While SMOTE was helpful for balancing the dataset, the performance of the model on the high-risk group showed a typical tradeoff between precision and recall in imbalanced classification tasks. The high accuracy combined with very strong metrics concerning the low-risk class, together with the challenges present in the high-risk class, underlined careful handling of imbalanced datasets in financial contexts where false positives can lead to incorrect decisions. On the other hand, the application record dataset was very good. Conjointly, therefore, this means that the ensemble approach was usually much better when the data was less complexly structured or more evenly distributed.

5.2. Evaluation

The corresponding results reflect that the models are performing well for low-risk classes but are struggling to identify high-risk individuals, especially in imbalanced datasets like the credit record. The credit record model presented a very respectable ROC-AUC score of 0.88 but reflected a diminished capacity to separate high-risk and low-risk cases. SMOTE post-resampling greatly improved the recall of high-risk people but significantly lowered the precision. That is, even though more cases were found for high-risk ones, a larger number of low-risk cases was misclassified as high risk. This is an important trade-off in financial risk prediction since false positives might lead to incorrect financial decisions that involve, for example, denying credit to people who are not actually at risk. However, the application record dataset did not pose any issues like this, probably because the nature of the structure of that data was simpler and the class distribution was better. This again would reaffirm that an ensemble model will be quite adaptive with less complex data, or even better, more balanced data.

5.3. Future Directions

This is the limitation that was observed in handling imbalanced data, especially when trying to predict high-risk cases. A possible related emphasis in future research could also involve making use of more sophisticated resampling or ensemble techniques which can better balance the precision-recall tradeoff, like adaptive boosting or cost-sensitive learning approaches where higher penalties are placed on misclassified high-risk cases. Further, the integration of XAI techniques will provide better interpretability of the models and clarify for the stakeholders why the model would make certain predictions, particularly in high-stakes financial decisions. This is especially important because, while black-box models are often quite accurate, they lack transparency. This will also include hybrid models that capitalize on both the strengths of deep learning architectures, such as LSTM or CNN, and ensemble learning to capture even more complex patterns in credit data, especially in time-series contexts. Testing these models with various sets of financial datasets from different regions or sectors would enhance the generalizability and robustness of the models, hence making them more applicable to real financial systems.

6. Conclusion

This study has demonstrated how different ensemble models like RandomForest, GradientBoosting, and DecisionTree classifiers can serve as the best options in credit risk prediction within financial datasets. The models were performing very strongly in identifying low-risk individuals, supported by their high accuracy and AUC scores, especially when their performance in the application record dataset yielded a perfect ROC-AUC score of 1.00. However, imbalanced datasets pose a lot of challenges, especially in the credit record dataset, with volumes about how difficult it really is to get the exact number of high-risk people correctly. On one hand, although the application of SMOTE helped improve recall for the high-risk class, this came at the cost of precision, entailing more false positives. In financial risk prediction, this trade-off between precision and recall becomes very important because wrong classifications lead to major real-world repercussions.

These findings point out the treatment of class imbalance with care and model selection that will perform well in both low-risk and high-risk classes. Though the overall predictive performance of the ensemble models was strong, this study finds that further refinements are required, especially in treating high-risk cases for more reliable and actionable financial risk assessments. Future work should deal with advanced approaches for higher precision of high-risk individuals and model interpretability via Explainable AI to provide financial institutions with the ability to make truly informed and transparent decisions based on such predictive models. This is where the integration of strong ensembling methods with resampling techniques and interpretability holds immense promise for furthering credit risk prediction in high-stakes environments.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Bao, Y., & Yang, S. (2023). Two Novel SMOTE Methods for Solving Imbalanced Classification Problems. IEEE Access, 11, 5816-5823.

[2] Elreedy, D., Atiya, A., & Kamalov, F. (2023). A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. Machine Learning, 1-21.

[3] Wongvorachan, T., He, S., & Bulut, O. (2023). A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining. Inf., 14(54).

[4] Rudin, C. (2018). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215.

[5] Kaggle Datasets: https://www.kaggle.com/datasets/rikdifos/credxit-card-approval-prediction/data

https://www.kaggle.com/competitions/GiveMeSomeCredit/data

https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud?resource=download