Developed Ensemble Model Based on Multiple Machine Learning Models

Weishan Li

Information System, University of International Relations, Beijing, 100091, China
liweishan@uir.edu.cn

Abstract. In recent years, the practical application of machine learning and natural language processing models, such as the random forest model and decision tree model, has become widespread. These models have been successfully employed in various domains, including economic forecasting and sentiment analysis, leading to enhanced convenience in people's lives and improved work efficiency. However, the current predominant use of individual machine learning models has exhibited diminishing performance, making it challenging to enhance their accuracy in certain cases. The objective of this study is to identify several machine learning models with superior performance and subsequently employ their output results as independent variables. A new model is then selected and trained using these variables to improve accuracy. By utilizing ensemble models, the risk of overfitting is mitigated, and the robustness of the models is increased, enabling effective handling of complex problems. In the conducted experiment, the integration of models resulted in a 0.5% improvement over the original best-performing single model, achieving an overall accuracy of 85.6%. Notably, this enhancement successfully predicted the correct outcomes for 12 challenging data points that were difficult to improve using the original single machine learning models.

Keywords: Machine Learning, Natural Language Processing, Ensemble Model.

1 Introduction

The advancements in data science and artificial intelligence have ushered in a flourishing era across various domains, such as finance [1], medical treatment [2], retail industry, manufacturing industry, natural language processing and computer vision. Currently, machine learning plays an important role in financial business forecasting. With the increasing amount of data, traditional financial business forecasting methods have been unable to handle large-scale data processing tasks. By training models, machine learning can extract features from massive amounts of data and make predictions regarding future trends.

Specifically, machine learning contributes to financial business forecasting in the following ways. Firstly, it aids investment decision-making processes by employing sophisticated algorithms to analyze diverse investment options such as stocks, bonds, and commodities, thereby identifying profitable investment opportunities. For

P. Kar et al. (eds.), *Proceedings of the 2023 International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2023)*, Advances in Computer Science Research 108,
https://doi.org/10.2991/978-94-6463-300-9_69

instance, the ability to predict stock price trends based on historical data enables investors to make more informed and accurate buying and selling decisions. Secondly, machine learning contributes to credit risk assessment by scrutinizing extensive historical data encompassing a customer's credit history, financial transaction records, and other relevant information, facilitating an accurate evaluation of their creditworthiness. These evaluation results can provide more accurate risk control and customer credit evaluation services for banks and insurance companies. Lastly, machine learning finds application in fraud detection scenarios, particularly in payment and credit card transactions. For example, based on the historical data of consumers, it can identify possible fraudulent transactions and conduct timely warning and prevention. In short, the application of machine learning in financial business forecasting can help improve forecasting accuracy, reduce risks, improve efficiency, and provide more reliable and accurate support for financial and business decisions [3-5].

In short, the application of machine learning in financial business forecasting can help improve forecasting accuracy, reduce risks, improve efficiency, and provide more reliable and accurate support for financial and business decisions.

In recent years, with the explosive growth of data and the continuous improvement of computing capability, the field of artificial intelligence has obtained unprecedented development opportunities. In this field, machine learning, as one of the most representative sub-fields, is widely used in data analysis, predictive modeling and other tasks. However, although the single-model machine learning method such as sales forecasting, stock price forecasting, customer churn forecasting, etc. can achieve satisfactory performance in many tasks, it performs poorly in some scenarios, such as sparse sample data and overlapping feature distributions, and the generalization performance of the model will be greatly affected. In [6], it used the decision tree and logistic regression to predict whether banks will lose customers or not. Although it used some other methods to improve the model performance like use the K-fold cross-validation and the data set fits well with these models, it finally the results were barely satisfactory but not perfect.

In order to overcome the limitations of a single model, this article considered the application of ensemble model, that is, when each single model finishes their prediction and output the result as the new variable x then use these variables to train and study by machine learning models. And evaluate each model to find the best model of these new variables. Finally get the best output of the integrating model and compare it with the single model used before.

2 Method

2.1 Dataset Description and Preprocessing

In this study, the dataset is sourced from the Kaggle [7], consisting of 10,000 items and 14 kinds of features about customers such as credit score, gender, age, etc. The label of this dataset is whether the customer exited and the 0 stands for the did not exit and the 1 stands for the exited. In the first step of the preprocessing, some useless

features like Row Number, Customer Id and Surname were removed when analyzing the whole data.

2.2 Ensemble model

Logistic Regression. Logistic regression is a statistical modeling technique used for predicting binary outcomes or probabilities [8]. It models the relationship between the independent variables (predictors) and the probability of an event occurring. The logistic function, or sigmoid function, is used to transform the linear relationship into probabilities. It has been widely used in various fields for binary classification tasks. It provides interpretable coefficients that represent the influence of each variable on the outcome's probability.

Support Vector Machine. Support vector machine (SVM) is a popular supervised machine learning algorithm used for classification and regression tasks [9]. It has gained significant popularity due to its effectiveness in processing high-dimensional data and its ability to handle complex decision boundaries. In support vector machines, the goal of the algorithm is to find an optimal hyperplane to separate different categories in the feature space. The hyperplane is determined by the subset of the training sample closest to the decision boundary (called the support vector). These support vectors play a crucial role in defining decision boundaries and maximizing the margin between classes. For classification tasks, the goal of SVM is to find a hyperplane that maximizes the margin between support vectors of different classes. This enables better generalization and reduces the risk of overfitting.

Random Forest Classifier. Random Forest Classifier is a popular machine learning algorithm used for classification tasks [10]. It is an ensemble method that combines multiple decision trees to make predictions. In a Random Forest Classifier, a collection of decision trees is created, where each tree is constructed using a random subset of the original features and training data. During training, the algorithm generates different subsets of the data through a process known as bootstrapping. To make predictions, each decision tree in the ensemble independently classifies the input data, and the final prediction is determined by a majority vote or averaging among the individual tree predictions.

To make predictions, each decision tree in the ensemble independently classifies the input data, and the final prediction is determined by a majority vote or averaging among the individual tree predictions.

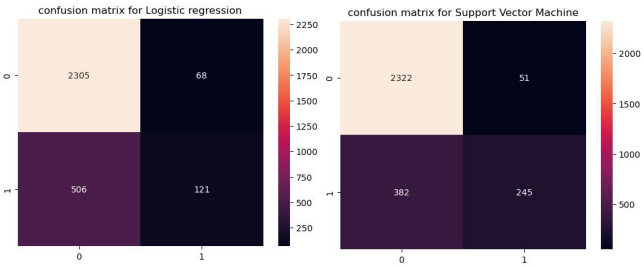
Decision tree classifier. Decision Tree Classifier is a machine learning algorithm used for classification tasks. It creates a tree-like model of decisions based on features and their respective thresholds. In a Decision Tree Classifier, the algorithm starts with the entire dataset and selects the best feature to split the data based on certain criteria. This process is repeated recursively for each node until a stopping condition is met, such as reaching a maximum depth or having a minimum number of samples at a node. Decision Tree Classifier is advantageous due to its interpretability, as the decision rules can be easily understood by visualizing the tree structure. It can handle both numerical and categorical features and is robust against missing data. Decision trees can also capture non-linear relationships between features.

Strategy of Ensemble. In order to determine the most optimal weights for each model's outcome, a process is undertaken whereby the output, denoted as y, from each individual model is utilized as a new variable x. Then using the machine learning model to train this integrated dataset to form a better prediction. The advantage of this integrated model over a single model are as follows: Firstly, when combining the predictions of multiple models into a new feature can combine the strengths of different models and improve the predictive power and accuracy of the models. Secondly, in order to deal with the problem that a single model cannot meet the needs in some cases, combining the prediction results of multiple models can make up for the shortcomings of a single model and improve the robustness and generalization ability of the model. Additionally, the incorporation of prediction outcomes from multiple models into the training set as new features enhances interpretability, providing insight into the rationale and foundations underpinning the model's predictions.

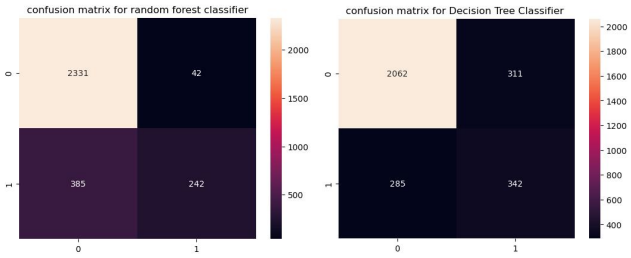
3 Results and Discussion

3.1 The Performance of Various Models

Upon meticulous examination of the confusion matrix shown in Fig. 1 for each of the four individual machine learning models, it becomes evident that the random forest classifier emerges as the most adept, attaining an accuracy of 85.7%. From the detailed analysis from the confusion matrix, the number of predictions as exit which the label is 1 is bit lower than the support vector machine. This disparity in performance is attributed to the nonlinear nature of the dataset and the presence of inherent noise within it. As the random forest classifier is not as robust as the support vector machine does, these noises may exert some interference on the it.



(a) (b)



(c) (d)

Fig. 1. The confusion matrix of machine learning models (Photo/Picture credit: Original).

3.2 The Feature Importance of the Model

As Fig. 2 shows, after each model finishes their predictions, the importance of each feature can be clearly shown. The three most important characteristics are age, number of products and balance, they add up to more than sixty percent.

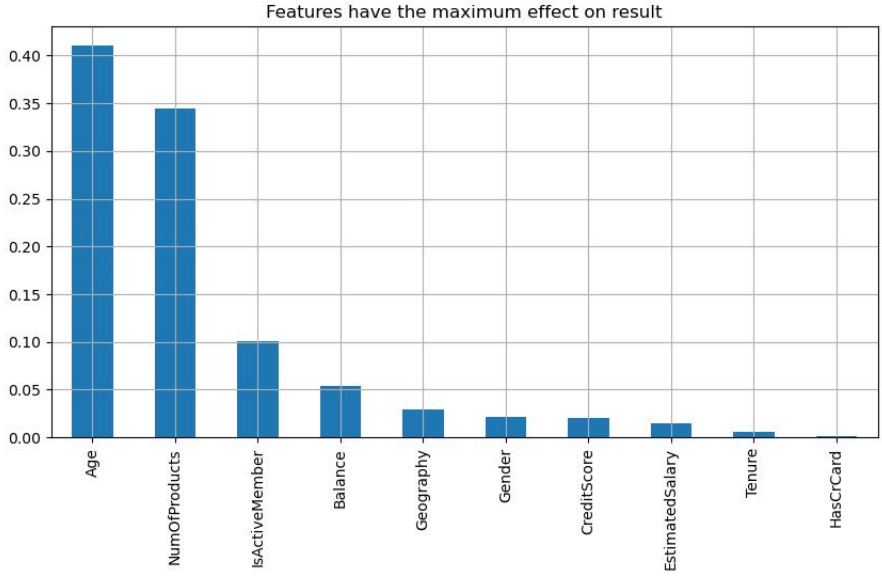


Fig. 2. The feature importance based on the random forest model (Photo/Picture credit: Original).

3.3 The Performance of the Ensemble Model

As shown in Fig. 3, after adopting the integrated model, the accuracy of the model improved by 0.5 percent, and 12 more data were successfully predicted. When compare the ensemble model with the best single model made prediction before which is random forest classifier model, the ensemble model predicts an additional 30 correct data predictions of user who will churn but an additional 18 wrong predictions of user who will not churn. This may because after model integration, the performance difference between each original model is too large such as the decision

tree model were too bad at predicting at true positives in confusion matrix, which is the user who will not exit, and the dataset imbalances resulted in the integration method negatively affecting a few category predictions.

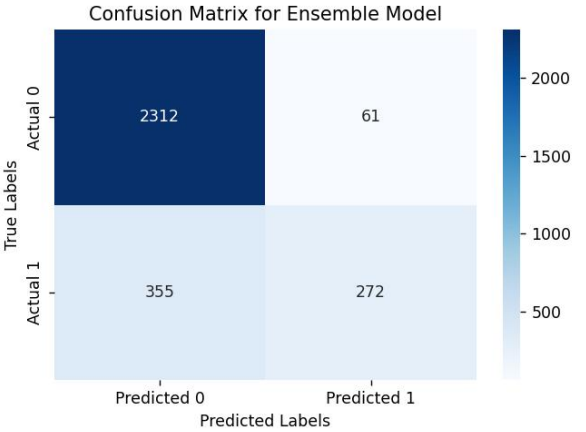


Fig. 3. The confusion matrix based on the ensemble model (Photo/Picture credit: Original).

4 Conclusion

This study aims to explore the effect of adopting an ensemble model instead of a single machine learning model in silent customer prediction tasks. The results show that the proposed ensemble model has achieved better performance in each evaluation index, which is obviously better than the single model. It improved the model's robustness because the ensemble model uses the set prediction results of multiple base models, it has a certain fault tolerance for the wrong prediction of individual base models. This makes the ensemble model more robust in the face of noise and outliers, and the prediction results more reliable. However, the ensemble also has some limitations: it reduced the model's interpretability that ensemble model is less interpretable than single models. Because the predicted results of the integrated model are the synthesis of multiple models, it is challenging to directly explain the contribution of each base model to the results. Therefore, in application scenarios where model interpretability is emphasized, the use of integrated models needs to be balanced. Moreover, since the integrated model is essentially an improvement based on the original single model, the improvement effect will be limited if the original models have similar performance.

References

1. Lappas, P. Z., Yannacopoulos, A. N.: A machine learning approach combining expert knowledge with genetic algorithms in feature selection for credit risk assessment. Applied Soft Computing, 107: 107391 (2021).
2. Belhaouari, S. B., and Islam, A.: Deep Learning in Healthcare, (2021).

3. Yu, Q., Chen, P., Lin, Z., et al.: Clustering Analysis for Silent Telecom Customers Based on K-means++, 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). IEEE, 1: 1023-1027 (2020).
4. Goodell, J. W., Kumar, S., Lim, W. M., et al.: Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis. Journal of Behavioral and Experimental Finance, 32: 100577 (2021).
5. Rundo, F., Trenta, F., di Stallo, A. L., et al.: Machine learning for quantitative finance applications: A survey. Applied Sciences, 9(24): 5574 (2019).
6. Alisa, B. Z.: Predicting Customer Churn in Banking Industry using Neural Networks. Interdisciplinary Description of Complex Systems - scientific journal 14.2, 116-124 (2016).
7. Kaggle: Bank Customer Churn Prediction https://www.kaggle.com/code/kmalit/bank-customer-churn-prediction (2018)
8. Imran, K.: Prediction of stock performance by using logistic regression model: evidence from Pakistan stock exchange (PSX). Asian Journal of Empirical Research 8.7 247-258 (2018).
9. Zhang, R., Zhang, X.: Support Vector Machine Prediction Modeling for Automobile Ownership. Journal of Computer and Communications, 10(06): 37-43 (2022).
10. Santos, E. E., Sena, N. C., Balestrin, D., et al.: Prediction of burned areas using the random forest classifier in the Minas Gerais state. Floresta e Ambiente, 27 (2020).