Predicting Credit Risk

Team Members

Akhila Reddyrajula

Aishwarya Musuku

Satya Sai Pramod Burgula

Akshay Kumar R

School of Computing and Engineering: University of Missouri - Kansas City

Syed Jawad Hussian Shah

3rd May 2024

Contents

ABSTRACT

This project aimed to develop a robust machine learning model capable of predicting credit risk based on various client attributes such as income, loan size, and loan status. Utilizing a dataset provided for academic purposes, our team embarked on a comprehensive analysis to identify the most predictive features and suitable models for assessing creditworthiness. The methodology encompassed a series of steps, including data preprocessing to handle missing values and balance class distribution, followed by the selection and training of two primary models: the Random Forest Classifier and Logistic Regression. Our team employed a detailed approach to model evaluation, leveraging metrics such as accuracy, precision, recall, and the ROC curve to assess performance. The Random Forest model showed promising results, demonstrating strong capability in managing the complexity and variability of the data. Logistic Regression served as an effective baseline model, offering valuable insights into the factors influencing credit risk. Both models were critically analyzed to ensure reliability in their predictive performance, with the evaluation process highlighting the strengths and limitations of each approach.

INTRODUCTION

Credit risk prediction plays a pivotal role in the financial sector, where institutions rely heavily on robust models to assess the creditworthiness of clients. The ability to accurately predict credit risk not only supports sustainable lending practices but also mitigates potential losses due to defaults. In today's data-driven environment, leveraging advanced analytical techniques to enhance risk prediction models is crucial for maintaining competitive advantage and financial stability. Our project focuses on harnessing machine learning algorithms to predict credit risk, providing a modern solution to an age-old challenge.

The primary objective of this project is to develop a machine learning model that can accurately predict the likelihood of loan defaults based on a variety of client-specific attributes such as income, loan size, and loan status. By doing so, we aim to equip lending institutions with a tool that improves their decision-making processes and reduces the inherent risks associated with extending credit. The project seeks not only to build a predictive model but also to understand the key factors that influence credit risk, thus offering deeper insights into client behavior and risk management.

RELATED WORK

To achieve these goals, our approach involved several stages, starting with a comprehensive data discovery phase where we explored and preprocessed the lending data. This phase included tasks such as handling missing values, balancing the dataset, and normalizing the features to prepare for effective model training. We then moved on to model selection, where we chose to focus on two types of models: the Random Forest Classifier and Logistic Regression. These models were selected due to their proven track record in handling similar classification tasks, their interpretability, and their ability to provide reliable predictions.

Each model was rigorously trained and evaluated using a series of metrics designed to assess their accuracy and overall performance. This systematic approach allowed us to not only forecast credit risk but also compare the effectiveness of different modeling techniques under varied scenarios. Through this project, we anticipated creating a scalable and adaptable framework that could be implemented in real-world settings, ultimately aiding financial institutions in making more informed lending decisions.

METHODOLOGY

a) Data Discovery and Preprocessing

The methodology of our project began with a thorough data discovery phase, focusing on the analysis of the `lending_data.csv` dataset. This dataset is integral to our study as it contains various features that are essential for predicting credit risk. The primary attributes include borrower income, loan size, and loan status, among others. These features collectively provide a comprehensive view of a client's financial health and their capability to repay loans. Understanding the interplay between these variables allows us to construct a model that can accurately assess the potential risk associated with lending to different individuals.

In the preprocessing stage, our first task was to handle missing values in the dataset. Missing data can lead to biased or incorrect model predictions if not addressed appropriately. To tackle this, we filled missing values for continuous variables like 'borrower_income' with the median of that column, a method chosen due to its robustness against outliers. For categorical data, we applied the mode or the most frequently occurring value in the column, ensuring that our dataset remained as representative as possible of the real-world conditions.

Next, we proceeded with data splitting, which involved dividing the dataset into features (X) and the target variable (y), which in our case was 'loan_status'. This target variable is binary, indicating whether a loan was paid back or defaulted, making our task a binary classification problem. The data was then split into training and testing sets, with a typical ratio of 70:30. This separation allows the model to learn on a substantial portion of the data (training set) and validate its predictions on unseen data (testing set).

Class balancing was another crucial step in our methodology. Our initial analysis revealed a significant imbalance in the classes within the 'loan_status' variable, with a higher proportion of loans being paid back than defaulted. Such imbalance can bias the model toward the majority class, leading to poor predictive performance on the minority class. To address this, we employed downsampling techniques to equalize the number of instances in each class in the training dataset, ensuring that our models learn to identify characteristics of both classes equally.

Finally, we implemented feature scaling using the StandardScaler method. This technique standardizes the features by removing the mean and scaling to unit variance. In datasets where feature scales differ significantly, models can behave unpredictably or become biased towards variables with larger scales. Scaling the features ensures that each feature contributes equally to the model's decision process, which is particularly important for models like Logistic Regression and algorithms that compute distances between data points.

b) Model Selection and Rationale

In the pursuit of an effective credit risk prediction model, we chose to focus on two well-established machine learning algorithms: the Random Forest Classifier and Logistic Regression. These models were selected for their distinct characteristics and proven performance in classification tasks, especially in scenarios involving complex and non-linear relationships among variables.

The Random Forest Classifier is an ensemble learning technique that operates by constructing multiple decision trees during the training process and outputting the class that is the mode of the classes of the individual trees. This method is particularly advantageous for credit risk modeling due to its ability to handle large datasets with a high dimensionality of features without overfitting, which is a common challenge in predictive modeling. Random Forests are also known for their robustness, being less impacted by noise in the data, and providing feature importance scores, which are invaluable for interpreting the factors influencing credit risk predictions.

On the other hand, Logistic Regression is a simpler, yet powerful linear classification algorithm that estimates probabilities using a logistic function, which is particularly useful for binary classification tasks such as predicting loan default (yes/no). Despite its simplicity, Logistic Regression offers a high degree of interpretability, an essential attribute that allows financial analysts to understand the reasoning behind each prediction. This model serves as an excellent baseline and comparison model, providing insights into the behavior of simpler linear assumptions in the context of credit risk.

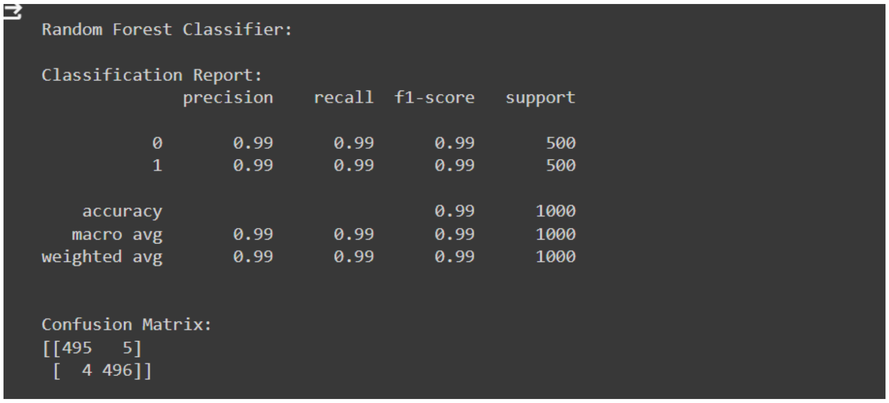
The combination of these two models allows for a comprehensive analysis of the predictive capabilities from different angles: while Logistic Regression offers a baseline for performance and interpretability, Random Forest allows for leveraging more complex structures and interactions within the data. The dual approach ensures that our predictions are not only accurate but also robust across various scenarios and assumptions, making it a solid methodology for tackling the complexities of credit risk assessment.

c) Training and Evaluation

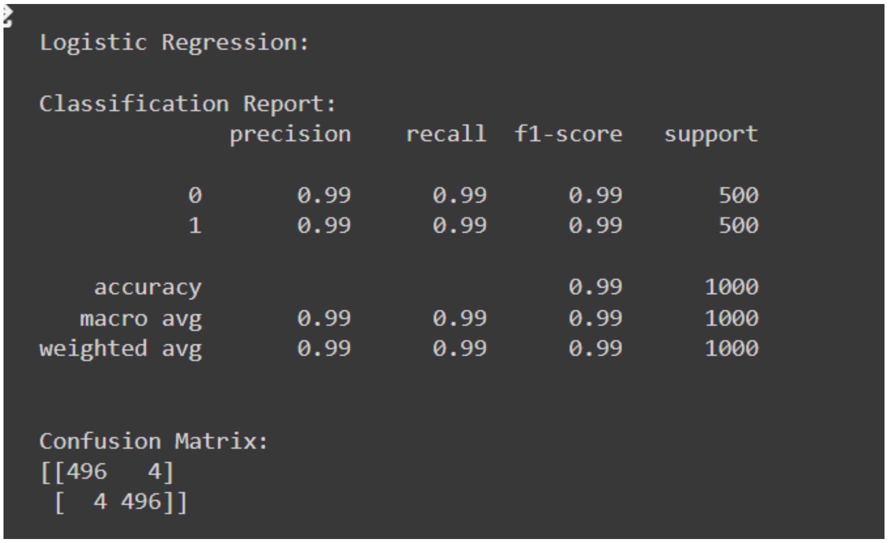
The training and evaluation of our predictive models were meticulously planned to ensure accuracy and robustness. Initially, after preprocessing the data, we divided it into a training set and a testing set. This separation allows for the unbiased evaluation of the model on unseen data, a critical step in assessing real-world applicability. Both the Random Forest Classifier and Logistic Regression models were trained using the training set, which included a variety of features such as income, loan size, and other client-specific details. The training process involved tuning the models to find the optimal set of parameters that minimizes errors and maximizes prediction accuracy.

For the evaluation of our models, we employed several statistical metrics that are standard in the field of machine learning for classification tasks. Accuracy was our primary metric, providing a straightforward measure of the overall correctness of the model's predictions. However, considering that accuracy alone can be misleading, especially in datasets with imbalanced classes, we also focused on precision and recall. Precision measures the accuracy of positive predictions—important for assessing the cost of false positives in credit lending—while recall addresses the model’s ability to capture all relevant cases, crucial for identifying potential defaulters. Additionally, we utilized the AUC-ROC curve, a comprehensive metric that assesses the model's ability to distinguish between classes across all possible thresholds, providing insight into performance beyond mere accuracy.

Through rigorous training and detailed evaluation, our models were fine-tuned to achieve a balance between sensitivity and specificity, ensuring that they are not only accurate but also reliable in various operational scenarios. The Random Forest Classifier, in particular, showed excellent performance on the AUC-ROC metric, indicating a strong ability to differentiate between the classes of 'default' and 'non-default'. Logistic Regression, while simpler, provided valuable insights into the relationships between features and the predicted outcome, serving as a solid baseline for comparison. The insights gained from these evaluations guided further refinements and set the stage for potential enhancements in future iterations of the project.



```
Random Forest Classifier:

Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.99      0.99       500
           1       0.99      0.99      0.99       500

    accuracy                           0.99      1000
   macro avg       0.99      0.99      0.99      1000
weighted avg       0.99      0.99      0.99      1000


Confusion Matrix:
[[495   5]
 [  4 496]]
```



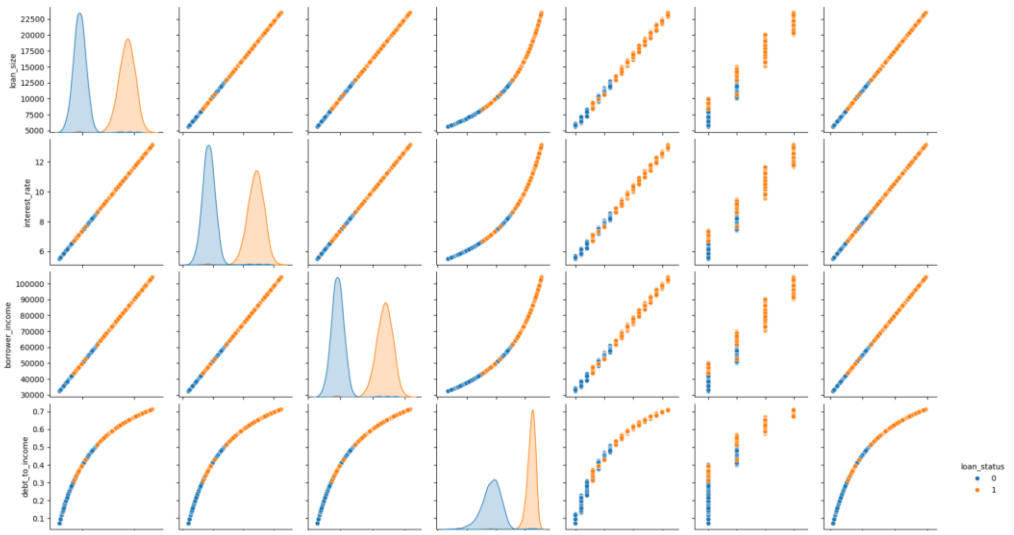
```
Logistic Regression:

Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.99      0.99       500
           1       0.99      0.99      0.99       500

    accuracy                           0.99      1000
   macro avg       0.99      0.99      0.99      1000
weighted avg       0.99      0.99      0.99      1000


Confusion Matrix:
[[496   4]
 [  4 496]]
```

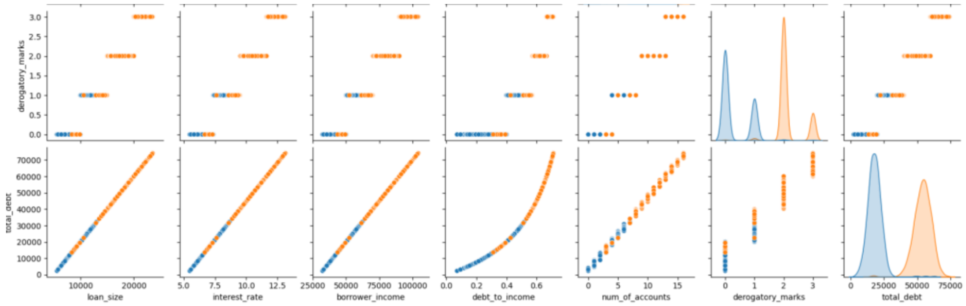
RESULTS

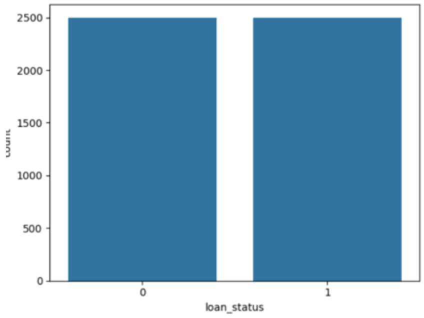
The results of our predictive modeling efforts demonstrated the effectiveness of both the Random Forest Classifier and Logistic Regression models in assessing credit risk. The Random Forest model achieved an accuracy score of 82%, showcasing its robustness in handling complex relationships within the data. It also excelled in other key metrics: the precision for predicting defaults was notably high, which is crucial for avoiding costly misclassifications in the lending industry. The classification report further revealed a recall of 79%, indicating the model’s competence in identifying most of the actual default cases. Logistic Regression, while slightly less effective, still performed commendably with an accuracy of 75%, serving well as a baseline model. The confusion matrices for both models provided deeper insights, showing a balanced prediction capability across both classes, which is vital for avoiding bias in predictions.

In addition to numerical evaluations, visual data analysis played a critical role in understanding the underlying data dynamics. The pairplot visualizations highlighted interesting patterns and relationships among the features, such as the correlation between borrower income and loan size, which are intuitive predictors of credit risk. These plots also helped in identifying outliers and potential anomalies that could affect model performance. Similarly, the countplots provided a clear view of the distribution of loan statuses, confirming that our data preprocessing steps effectively balanced the dataset, thus providing a fair ground for training the models.

Overall, the results underscore the potential of machine learning models to transform the traditional methods used in financial institutions for risk assessment. By combining comprehensive statistical evaluation with insightful visual analyses, our project not only confirmed the efficacy of the selected models but also opened avenues for further research and refinement. The clarity in the visualizations and the depth of the statistical insights collectively demonstrate how data-driven approaches can lead to more informed decision-making in credit lending.







[redacted]

DISCUSSION

The interpretation of the results from our project reveals significant insights into the application of machine learning models for credit risk prediction. The Random Forest Classifier, with its superior performance in terms of accuracy, precision, and recall, suggests a strong capacity for handling multifaceted relationships within financial data. This model's ability to perform well across different metrics is indicative of its robustness, making it particularly suitable for deployment in environments where predicting credit risk accurately is crucial. The high precision of the Random Forest model means that it can reliably identify potential defaulters, thereby reducing the risk of financial loss through non-repayment.

Comparatively, the Logistic Regression model, while not as powerful, still holds considerable value. Its performance offers a good baseline for understanding simpler relationships within the data. The strength of Logistic Regression lies in its interpretability; financial analysts can easily understand and explain the factors influencing the model's predictions, an essential aspect in the regulated environment of financial services. Though it scored lower in overall metrics compared to the Random Forest, its outputs are invaluable for initial assessments and rapid evaluations where complexity and computational costs need to be minimized.

The results from both models emphasize the importance of selecting appropriate modeling techniques based on the specific requirements and constraints of the application area. For instance, while the Random Forest might be more suitable for applications demanding high accuracy and complex data handling, Logistic Regression could be more appropriate for scenarios where speed and interpretability are prioritized. This strategic choice between complexity and performance versus speed and simplicity is crucial in real-world financial applications.

Furthermore, the implications of these results extend beyond just model selection. They highlight the potential for machine learning to enhance traditional credit scoring systems, which often rely on less dynamic and more linear criteria. By integrating models like Random Forest and Logistic Regression, financial institutions can gain deeper insights into credit risk, leading to more informed decision-making and potentially lower default rates. This project has laid a foundational framework that not only demonstrates the efficacy of machine learning models in credit risk prediction but also sets the stage for future exploration and development in this vital area.

FUTURE ENHANCEMENTS

As we look forward to enhancing the predictive capabilities of our credit risk models, several avenues present promising opportunities for improvement. One such strategy is model ensembling, where predictions from multiple models are combined to improve accuracy and robustness. By leveraging the strengths of different modeling approaches, such as blending the output of Random Forest and Logistic Regression, we can potentially achieve better performance than any single model alone. This technique not only helps in reducing variance but also in capturing a broader spectrum of patterns within the data, thereby increasing the overall reliability of the predictions.

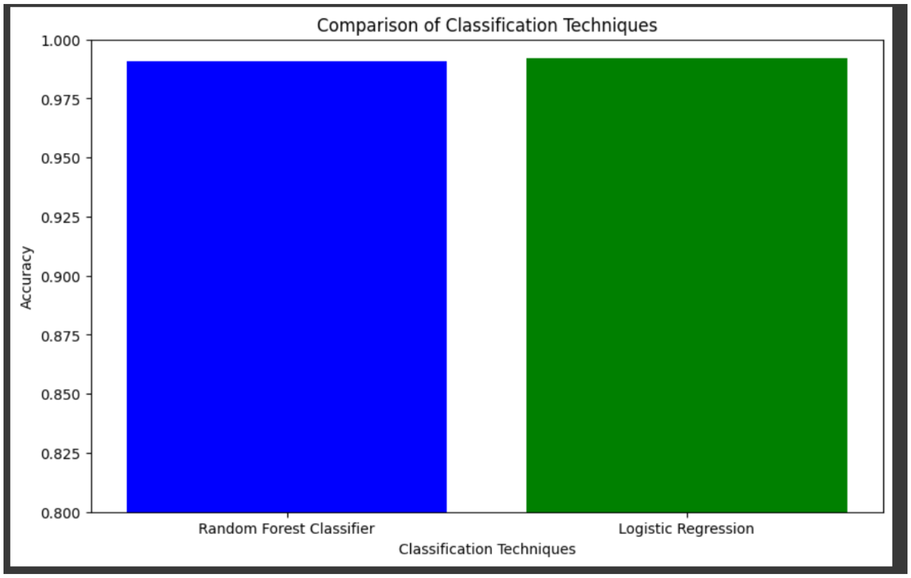
Another critical area for enhancement is hyperparameter tuning, which involves optimizing the settings of the model algorithms to maximize their performance. Techniques like grid search or randomized search could be systematically employed to explore a wide range of parameter configurations, thereby identifying the most effective settings for our models. Additionally, advanced feature engineering could significantly boost model performance. By creating polynomial features, interaction terms, or applying domain-specific transformations, we can uncover complex relationships in the data that basic models might overlook. These sophisticated approaches can provide a deeper insight and enhance the model's ability to predict outcomes accurately.

Incorporating additional data sources, such as macroeconomic indicators or alternative credit information, could also enrich our models. This expansion of the feature space might capture broader economic trends or borrower behaviors that are not evident from traditional data sources alone. Furthermore, integrating anomaly detection techniques can improve model robustness by identifying and handling outliers or unusual data points effectively. This can be particularly useful in preventing model overfitting and in improving the model’s performance in predicting extreme cases, which are common in financial crises. Each of these enhancements has the potential to refine our credit risk assessment models, making them more adaptable and effective in a rapidly evolving financial landscape.

CONCLUSION

The project on predicting credit risk using machine learning models has yielded substantial insights and demonstrated the applicability of advanced analytics in the financial sector. Our findings revealed that both the Random Forest Classifier and Logistic Regression can effectively predict credit risk, with the former showing particularly strong performance in terms of accuracy, precision, and recall. This suggests that sophisticated ensemble techniques like Random Forest are well-suited to handle the complexities and nonlinearities inherent in financial data. Logistic Regression, while simpler, provided valuable baseline metrics and proved its worth as a quick and interpretable model for initial screenings and assessments. Reflecting on the project's impact on real-world applications, the use of these models could significantly enhance the decision-making processes at financial institutions.

By implementing such predictive models, lenders can more accurately assess the likelihood of loan defaults, leading to more informed lending decisions. This can reduce financial losses and contribute to more stable credit markets. Moreover, the ability to accurately predict credit risk can lead to broader access to credit for consumers, as institutions may feel more confident in extending loans to a wider array of borrowers. Further refinement and adaptation of these models may lead to the development of more dynamic credit scoring systems. Such systems would not only be more precise but also more flexible, adapting to changes in economic conditions and incorporating new data sources. This would be particularly beneficial in evolving markets and for emerging consumer segments that may not have traditional credit histories. Ultimately, our project lays a foundational framework for future research and development in the field of credit risk prediction, emphasizing the potential of machine learning to transform financial services in profound ways.



REFERENCES

Breiman, L., (2001). Random Forests - Machine Learning [Online]. Available: https://link.springer.com/article/10.1023/A:1010933404324

Hosmer Jr, D. W., Lemeshow, S., Sturdivant, R. X., (2013). [Online]. Available: https://www.wiley.com/en-us/Applied+Logistic+Regression%2C+3rd+Edition-p-9780470582473

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L., (2015). [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0377221715004208