PROJECT REPORT
PREDICTING CREDIT RISK

University of Missouri – Kansas City

Lecturer: Syed Jawad Hussain Shah

3rd May 2024

Group Members
Abhinav Reddy Ayyadapu
Yogitha Mekala
Vedanth Goud Nagapola
Prajwal Eswar Chejarla

INTRODUCTION:

The possibility that borrowers would stop making loan payments is known as credit risk, and controlling it is a continuous issue for the financial sector. For loan approvals, interest rate setting, and portfolio risk assessment, accurate loan status prediction is essential. The purpose of this study is to investigate how supervised learning methods may be used to this important issue. Using a real-world dataset that includes loan parameters including loan size, interest rate, borrower income, debt_to_income, number of accounts, derogatory marks, total debt, and loan status, we assessed 10 different algorithms. Finding the best model to anticipate loan defaults (as shown by loan status) and comprehending the fundamentals of its performance are our main objectives.

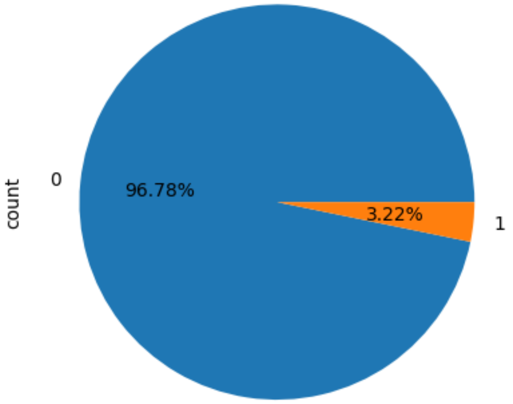
RELATED WORK:

The prediction of credit risk has been the subject of much machine learning research. Numerous techniques, such as random forest and logistic regression, have been used in previous research. New developments in technique such as Gradient Boosting Classifier have demonstrated encouraging outcomes in terms of obtaining good precision and high accuracy. Machine learning techniques are becoming more and more popular for predicting credit risk, especially for complicated and unbalanced datasets. Building on this basis, our study compares the performance of sophisticated and well-known algorithms and investigates the particular model that produces the most accurate results.

METHODOLOGY:

We implemented processes including understanding the dataset, performing basic preprocessing, visualizing the data to help understand it better, separating the features and labels based on the desired project goal, using classification techniques, and calculating evaluation metrics for each model to determine which is the best fit for the project.

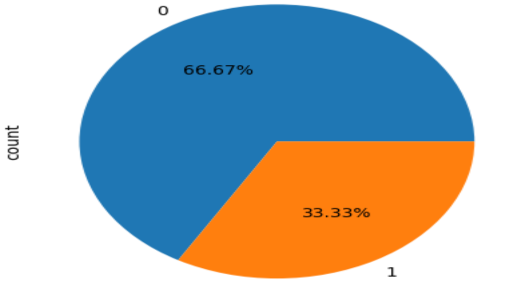
Due to imbalanced dataset, we will be using the SMOTE as an oversampling technique. As we can see the Data analysis the data predicted value is imbalanced. The below pie chart shows the imbalanced data.



SMOTE (Synthetic Minority Over-sampling Technique):

In datasets with imbalanced class distributions, such as the loan status column where approximately 97% of instances are labeled as 0 (indicating not approved) and only 3% as 1 (indicating approved), addressing this skew is crucial for building robust machine learning models. In this scenario, employing techniques like SMOTE (Synthetic Minority Over-sampling Technique) becomes imperative. SMOTE generates synthetic samples for the minority class (class 1 in this case) by interpolating between existing instances, thereby alleviating class imbalance. By augmenting the dataset with synthetic instances of approved loans, SMOTE ensures a more balanced representation of both classes, enhancing the model's ability to learn from minority class instances effectively. This approach mitigates the risk of the model being biased towards the majority class, leading to more accurate predictions and better overall performance.

Following the application of the oversampling technique, the distribution of loan statuses shifted, with approximately 66.67% of instances labeled as 0 and 33.33% as 1. This transformation is visually represented in the chart below. The oversampling method has effectively balanced the class distribution, resulting in a more equitable representation of both loan approval outcomes.



In a dataset where 66.67% of instances are labeled as "no" and 33.33% as "yes," we can consider it as balanced since there isn't a significant class imbalance. This balanced distribution allows machine learning models to learn from both classes effectively without being biased towards one class over the other. As a result, the models can make predictions that are more representative and reliable across both classes, leading to better overall performance and generalization.

Some supervised learning methods were put into practice. Classification methods include K-Nearest Neighbors, Support Vector Classifier (SVC), Random Forest, Gradient Boosting, and Logistic Regression. To separate the data into training and testing sets for each model, Used a train-test split. The model was fitted to the training set, and then its performance was assessed using performance metrics such as accuracy (classification) on the testing set.

RESULTS AND DISCUSSION:

1. Data Collection:

Obtaining information on credit risk or loan status forecast is the focus of this step. The datasets are sourced from internet repositories that focus on financial statistics.

2. Dataset:

The dataset includes data that is essential for evaluating credit risk:
loan_size: The total requested loan amount.
interest_rate: The interest rate associated with the loan.
borrower_income: The borrower's income.
debt_to_income: Ratio of debt to income.
num_of_accounts: Number of accounts the borrower has.
derogatory_marks: The quantity of negative marks included in the credit report of the debtor.
total_debt: The total amount that the borrower owes.
loan_status: The target variable that represents the status of loan approval (0 for not approved, 1 for approved).

3. Data Preparation:

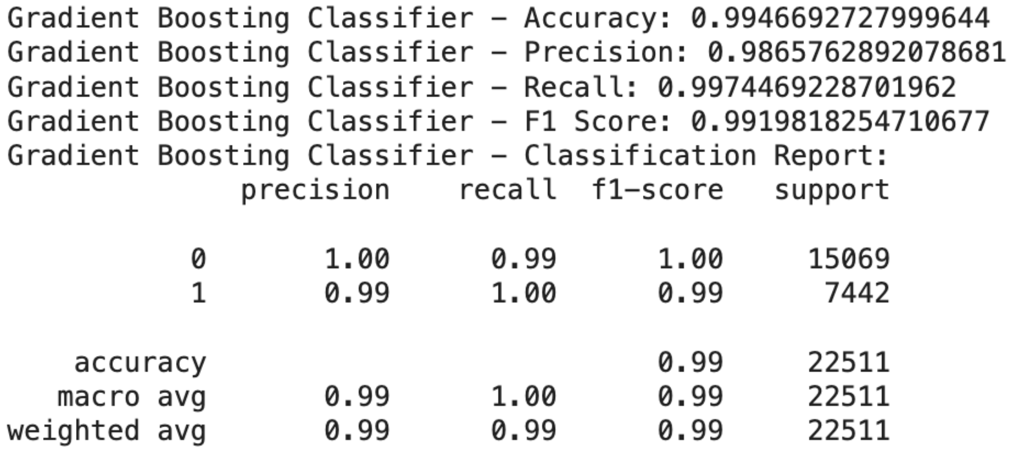
Data Visualization: Investigate relationships between count of unique labels, uncover class imbalances, and correlation between all characteristics. Data Cleaning: Find Null, find association between label with feature columns.

4. Model Selection:

Based on the dataset and problem statement, choose appropriate models for credit risk assessment or loan status prediction. The techniques used for this like:
Random Forest Classifier, Gradient Boosting Classifier, Support Vector Classifier, Logistic Regression, K-Nearest Neighbors Classifier.

Selected Model Gradient Boosting Classifier: The basic idea behind Gradient Boosting Classifier (GBC) is a popular machine learning technique used for classification tasks. It's an ensemble learning method that combines the predictions of several base estimators, typically decision trees, to improve accuracy.

The most accurate classification model was found to be the Gradient Boosting Classifier (GBC), which achieved an amazing 99.47% accuracy on the testing data. This implies that it has a remarkable capacity to identify the patterns that separate loans that have defaulted from those that have not. Understanding the particular decision limits that the Gradient Boosting Classifier would be possible by analyzing with n_estimators = 100.

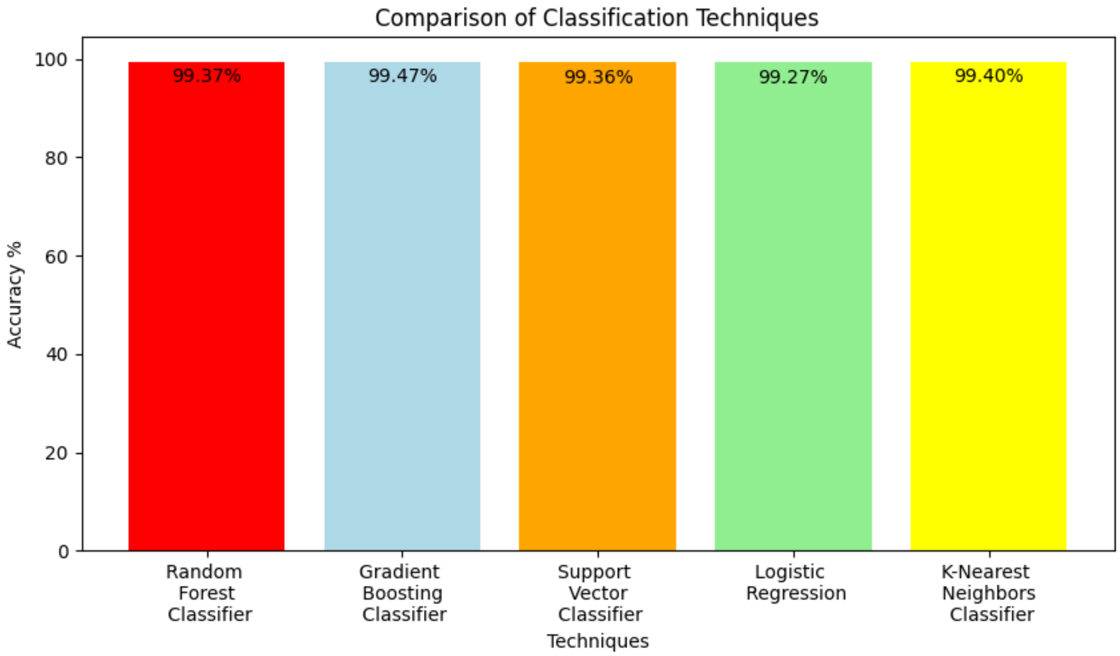


```
Gradient Boosting Classifier - Accuracy: 0.9946692727999644
Gradient Boosting Classifier - Precision: 0.9865762892078681
Gradient Boosting Classifier - Recall: 0.9974469228701962
Gradient Boosting Classifier - F1 Score: 0.9919818254710677
Gradient Boosting Classifier - Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.99      1.00     15069
           1       0.99      1.00      0.99      7442

    accuracy                           0.99     22511
   macro avg       0.99      1.00      0.99     22511
weighted avg       0.99      0.99      0.99     22511
```

However, it's critical to consider limitations. Overfitting, a condition in which the model performs well on training data but may find it difficult to generalize to new data, might be indicated by the high accuracy. Cross-validation techniques can be applied to assess generalizability and lower this risk. Moreover, focusing just on accuracy might not be sufficient for evaluating credit risk. The model has previously been evaluated for its ability to discriminate between true and false positives and negatives, which is crucial for decision-making, using metrics like accuracy, recall, and F1-score.

CONCLUSION AND FUTURE WORK:

This research delved into assessing the effectiveness of various supervised learning methods in predicting credit risk, with a particular focus on the Gradient Boosting Classifier (GBC). Among the array of classification models examined, the GBC emerged as the standout performer, showcasing remarkable accuracy and predictive power. The GBC accuracy is 99.47% which is more compared with other classifiers.



The GBC operates by sequentially combining weak learners, typically decision trees, to form a robust predictive model. Through an iterative process, it identifies and corrects errors made by preceding models, gradually refining its predictions with each iteration. This inherent adaptability allows the GBC to excel in capturing intricate patterns within complex datasets, making it an ideal candidate for tasks such as credit risk assessment.

Future projects may entail enhancing overall prediction accuracy by amalgamating multiple models, such as the Gradient Boosting Classifier (GBC), with other potent algorithms like Random Forest. By leveraging the strengths of each individual model, this approach aims to capitalize on the diverse perspectives and methodologies they offer.

Additionally, the project may explore the efficacy of deep learning architectures, including convolutional neural networks (CNNs) or recurrent neural networks (RNNs). These advanced techniques could prove invaluable in tackling complex datasets, such as sequential loan data, should they be available. The ability of CNNs and RNNs to capture intricate patterns and temporal dependencies within sequential data aligns well with the dynamic nature of credit risk assessment, potentially yielding significant improvements in predictive performance.

REFERENCES

[1]. Jomark Pablo Noriega; Luis Antonio Rivera , Jose Alfredo Herrera. Machine Learning for Credit Risk Prediction: A Systematic Literature Review, 2023.

[2]. Norshakirah Aziz; Emelia Akashah Patah Akhir; Izzatdin Abdul Aziz. A Study on Gradient Boosting Algorithms for Development of AI Monitoring and Prediction Systems, 2020.

[3]. Aized Amin Soofi; Classification Techniques in Machine Learning: Applications and Issues, 2017.

[4]. Swastik Satpathy; SMOTE for Imbalanced Classification with Python, 2023.