Final Report

Course: 16:958:588:01
2024SP - DATA MINING

Stability Selection

Variable Selection

Team 4

Aayush Manish Pradhan	ap2527
Himani Hooda	hh660
Manas Tanaji Maskar	mm3660
Vanshika Ram Gurbani	vg460

Abstract

This study explores the utilization of Stability Selection alongside L2 regularization and bootstrapping for feature selection within machine learning applications. Through comprehensive experimentation, we assess the efficacy of Stability Selection in bolstering model stability and generalization across diverse classification tasks. Our investigation primarily focuses on evaluating the influence of Stability Selection on key model performance metrics, including accuracy, precision, recall, and F1-score. Results indicate that the integration of Stability Selection with regularization (implemented L2) and bootstrapping facilitates the identification of durable features across varied data subsets, ultimately enhancing model interpretability and mitigating overfitting risks. These findings emphasize the significance of Stability Selection as a valuable approach for optimizing machine learning model performance in practical scenarios.

Introduction

In the realm of machine learning and data analysis, the evolution of feature selection methodologies is vital for tackling the challenges posed by complex datasets. Traditional techniques often struggle with high-dimensional data, leading to overfitting and diminished model performance, particularly in noisy or heterogeneous datasets. Stability selection addresses these concerns by leveraging resampling and aggregation, enhancing both stability and reliability in feature selection.

In the context of regularization, while performing bootstrapping, stability selection emerges as a powerful technique. It operates through repeated subsampling and aggregation, creating multiple data subsets, and applying a base feature selection method to each. Through this process, stability selection identifies features consistently selected across subsets, ensuring robustness in model performance and aiding generalization. Its versatility spans across various domains, from genomics to finance, making it adaptable to regression, classification, and clustering tasks. In our project, we delve into stability selection mechanism by performing it on Financial Data.

Exploring the dataset

The dataset’s inclusion of various financial indicators—like ROA, liquidity ratios, and debt ratios—is essential because these factors collectively paint a picture of a company's financial stability. Metrics such as ROA indicate operational efficiency, while liquidity ratios help assess a company's ability to meet short-term obligations without raising additional capital. Debt ratios and financial leverage indicators are crucial for understanding long-term solvency and a company's risk of defaulting.

Regarding data distribution, the presence of high kurtosis in many features signifies that the data have heavy tails or outliers. This is typical in financial data, which can be skewed by a few companies performing exceptionally well or poorly compared to the average. Thus, applying Standard Scaler. Apart from this also maintaining the class values in target values by applying

SMOTE and Undersampling and deciding which one to keep by implementing a base model. We also perform Correlation Plots and draw the network of the correlation of each feature with other features and target distance, here for this network, the distance is the correlation between them.

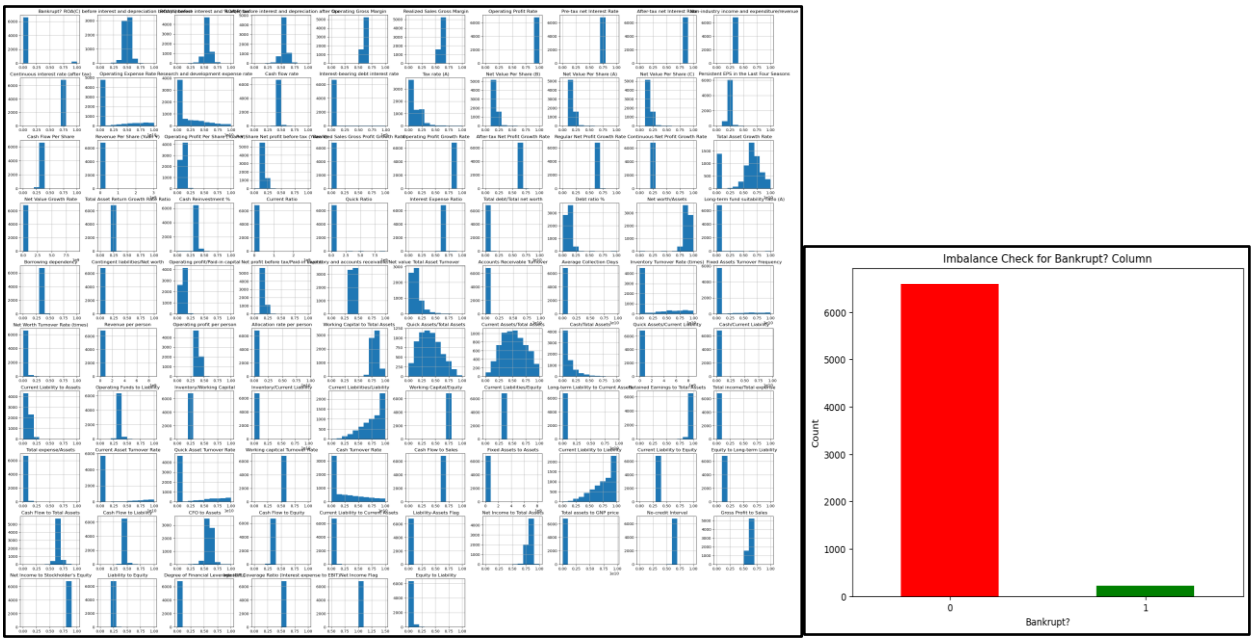

Figure 1: Checking imbalance in Bankrupt column.

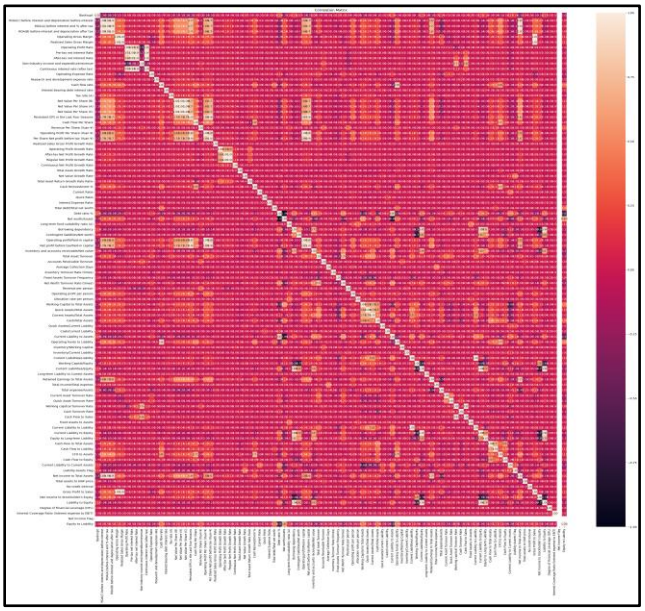

Figure 2: Correlation Matrix

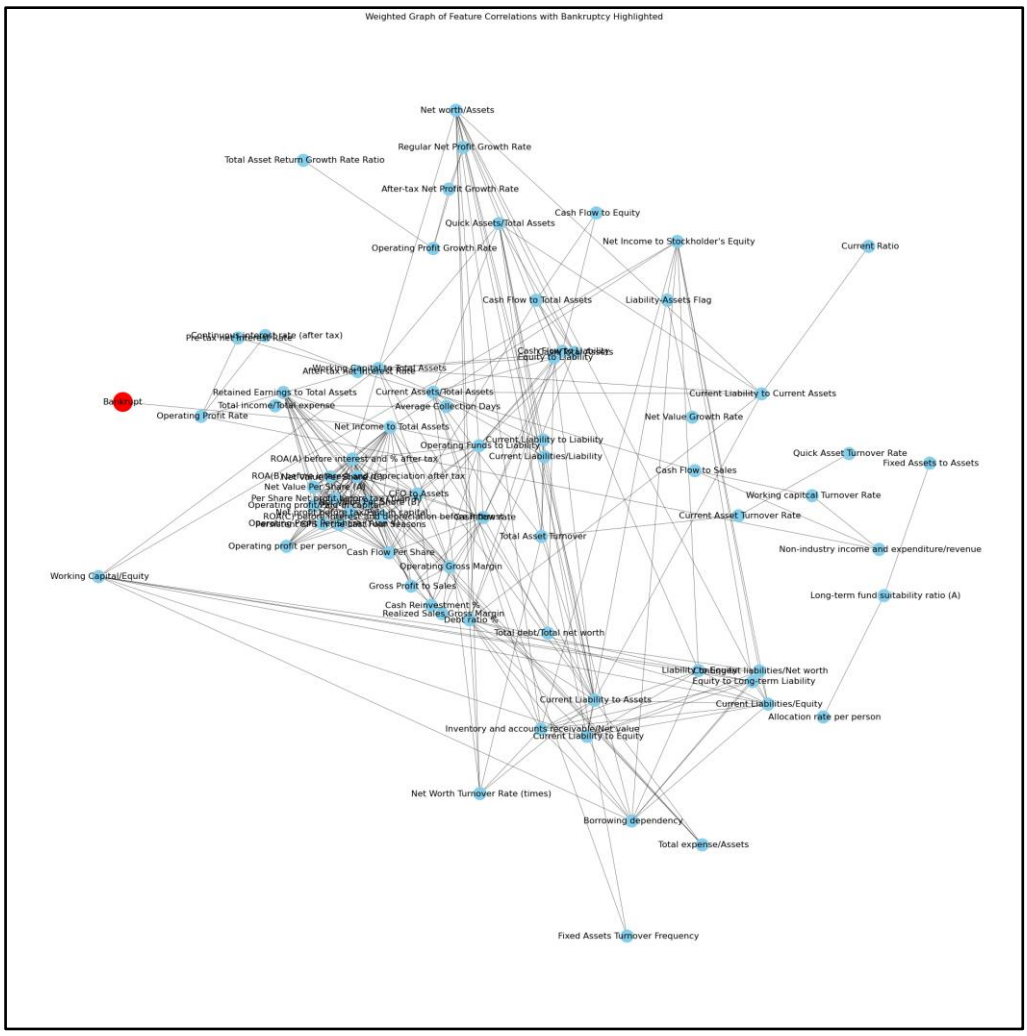

Figure 3: Network Graph

Implementing PCA

While implementing Stability Selection for feature selection in our classification model, we faced a significant challenge due to the dataset's high dimensionality, which strained computational resources. To address this, we decided to implement dimensionality reduction as a preliminary step.

Choosing Principal Component Analysis (PCA) over other dimension reduction techniques was strategic for several reasons. Firstly, PCA operates in an unsupervised manner, selecting features without relying on output labels and avoiding human bias in the selection process. Additionally, PCA is generally less computationally demanding compared to alternatives like t-SNE and AutoEncoders. While t-SNE and AutoEncoders have their strengths, such as visualizing high-dimensional data and capturing nonlinear relationships, respectively, they require significant computational resources. By reducing the feature set from 95 highly correlated features to 17 principal components, PCA not only alleviated the computational burden but also enhanced the

performance of the Stability Selection algorithm, resulting in a more efficient analysis and improved identification of stable and predictive features. This strategic use of dimensionality reduction ensured our model remained computationally viable while enhancing its robustness and accuracy in predicting classifications.

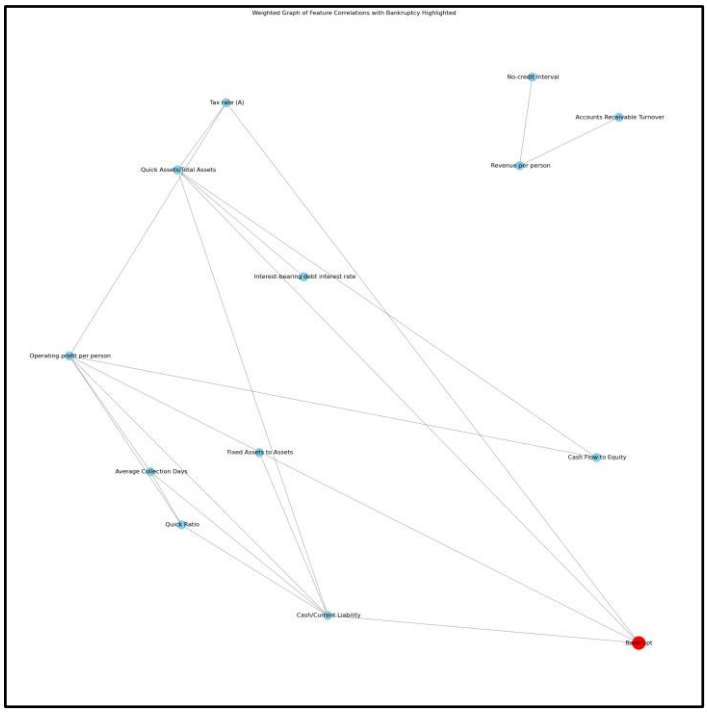

Figure 4: Correlation Network graph after PCA

```
Selected Features by PCA:
Index([' Operating Expense Rate', ' Interest-bearing debt interest rate',
       ' Tax rate (A)', ' Net Value Growth Rate', ' Quick Ratio',
       ' Interest Expense Ratio', ' Accounts Receivable Turnover',
       ' Average Collection Days', ' Revenue per person',
       ' Operating profit per person', ' Quick Assets/Total Assets',
       ' Cash/Current Liability', ' Cash Turnover Rate',
       ' Fixed Assets to Assets', ' Cash Flow to Equity',
       ' Total assets to GNP price', ' No-credit Interval'],
      dtype='object')
17
```

Figure 5: Features selected by PCA.

Implementing Stability Selection →

A. What is Stability Selection

Stability Selection involves evaluating the stability of chosen features across various data subsets to enhance their reliability. This is achieved through resampling techniques such as bootstrapping and applying a base feature selection algorithm to each subset. Incorporating regularization methods like L1 and L2 within Stability Selection can bolster the robustness of the model. By controlling parameters like alpha, the technique aims to identify variable features and those strongly correlated with the target variable, thereby mitigating overfitting risks, and enhancing model interpretability.

B. How to apply Stability Selection

Due to huge multicollinearity proceeding with L2 Regularization was a necessity as,

In our methodology, we chose to incorporate only L2 regularization within Stability Selection, omitting the use of L1 regularization altogether. Unlike L1 regularization, which encourages sparsity by driving specific coefficients to zero, L2 regularization penalizes the square of the coefficients' magnitudes, controlling their overall size. We made this decision to solely utilize L2 regularization to address potential multicollinearity issues and foster a smoother optimization landscape, thereby enhancing both model stability and convergence. By focusing solely on L2 regularization, our aim was to effectively balance feature importance and coefficient magnitude, ultimately improving the model's overall robustness and interpretability.

Regularization via bootstrapping entails generating several dataset subsets through resampling, followed by the application of the technique, L2 regularization to each subset. This process fortifies model stability and adaptability, rendering the model more resilient to overfitting. With this model we can specify the importance of features as those demonstrating stability over numerous subsamples of the data and the selection threshold and alpha can be controlled which is used to control the thresholding and the regularization strength.

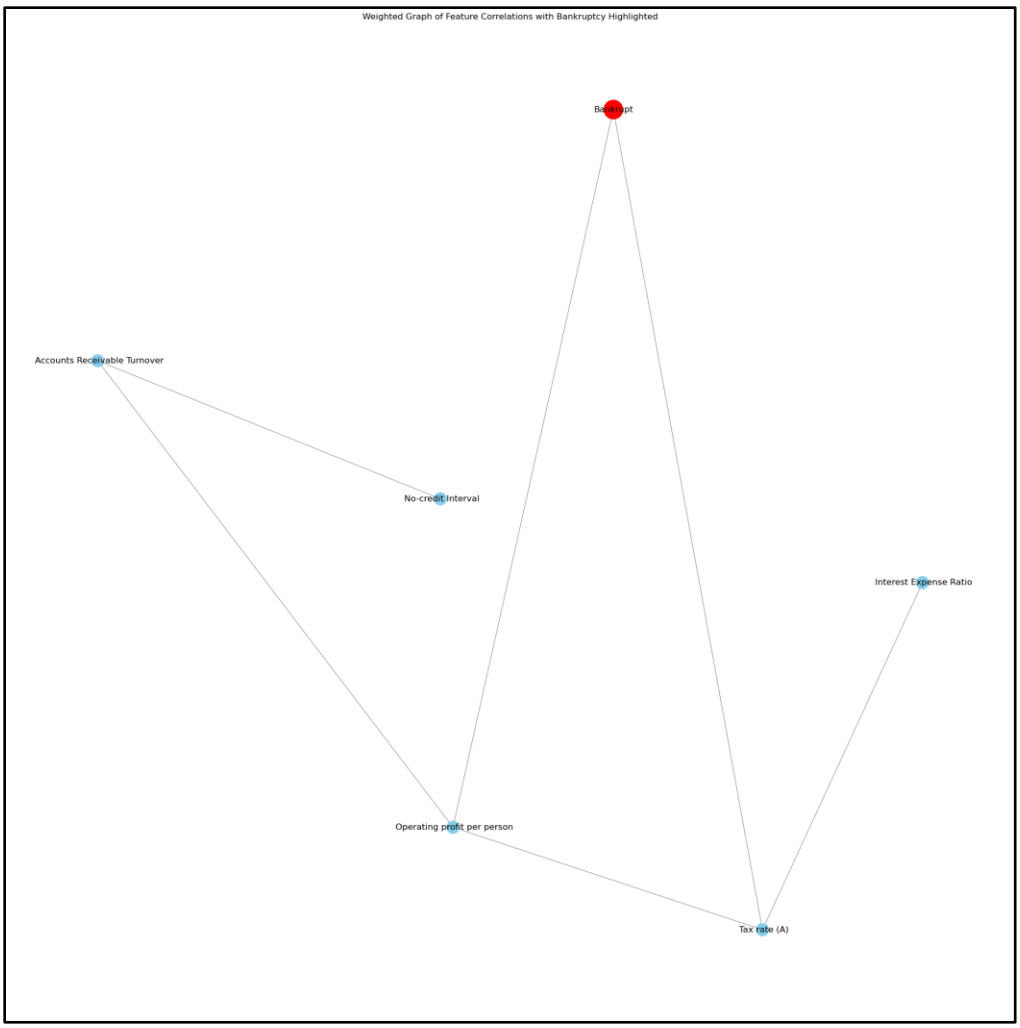

Figure 6: Network Graph after Stability Selection

Comparing the models →

In evaluating the performance of various classification models, we found notable strengths across Logistic Regression, Random Forest Classifier, and Gradient Boosting Machines (GBM). Initially, Logistic Regression served as a dependable baseline, delivering an accuracy of 0.76 alongside well-balanced precision, recall, and F1-score metrics hovering around 0.74-0.75. Subsequent implementation of Grid Search CV yielded marginal enhancements, showcasing refined parameter optimization without altering the accuracy significantly.
Transitioning to Random Forest Classifier, we observed a marked improvement in performance, surpassing Logistic Regression with an accuracy of 0.795. Following Grid Search CV, its accuracy further ascended to 0.818, accompanied by a well-balanced F1-score of 0.82 and improved precision and recall metrics. Notably, Random Forests demonstrated resilience against overfitting and proved adept at navigating non-linear data relationships, cementing their position as formidable contenders in classification tasks.

In contrast, Gradient Boosting Machines (GBM) exhibited exceptional performance, boasting an accuracy of 0.875 and an impressive AUC-PR of 0.92. With both precision and recall at 0.87 and an F1-score of 0.88, GBM excelled in capturing intricate data patterns and interactions. Despite its computational demands, GBM's outstanding predictive power and balanced performance metrics make it the optimal choice when maximizing model efficacy is paramount, provided ample computational resources are available.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression (Baseline Model)	0.76	0.73	0.75	0.74
Logistic Regression (Grid Search CV)	0.76	0.77	0.77	0.76
Random Forest Classifier	0.795	0.79	0.79	0.8
Random Forest Classifier (Grid Search CV)	0.818	0.82	0.81	0.82
Gradient Boosting Machines	0.875	0.87	0.87	0.88

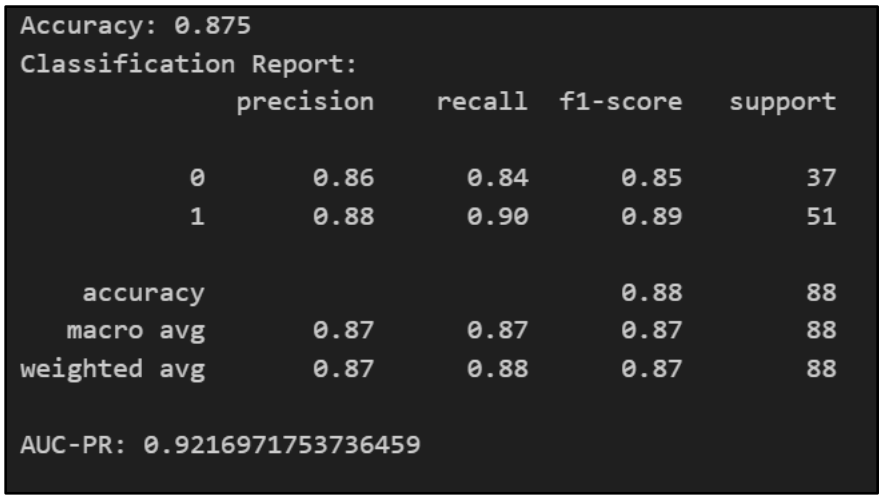

Figure 7: Classification Report after Gradient Boosting Machine

Conclusion

This project has successfully demonstrated the efficacy of Stability Selection integrated with L2 regularization and bootstrapping in feature selection for machine learning models dealing with financial data. The use of these methods has notably enhanced model stability and interpretability, crucial in handling high-dimensional and multicollinear datasets typically found in financial contexts. Our comparative analysis with various classification models underscored the strengths of Gradient Boosting Machines for optimal performance, supported by substantial computational resources. Both Random Forest and Logistic Regression models also exhibited strong performance, marking them as viable options depending on the specific requirements of the task. The practical applications of this project pave the way for more resilient and effective predictive models in financial data analysis, highlighting the importance of robust feature selection techniques in improving model accuracy and generalization across diverse data scenarios. This project contributes to the broader understanding of feature selection in data science and sets a foundation for future explorations into advanced data mining techniques in the financial industry.