Alexia

Analysis of the Polish Bankruptcy Dataset

Introduction

This report details the analysis of the Polish Bankruptcy dataset. Our primary objective is to develop a predictive model to identify potential bankruptcy cases based on financial indicators. This task has significant implications in the financial sector, providing insights into risk management.

Data Preparation and Preprocessing

- Loading Data: The dataset, available in ARFF format, was loaded into a pandas DataFrame.

- Data Type Conversion: The 'class' column, representing bankruptcy status, was converted from byte strings to strings and then to integers for binary classification. Other features were converted to floats to facilitate numerical analysis.

- Handling Missing Values: A significant portion of the dataset contained missing values. The columns `Attr37` and `Attr21` had the highest percentages of missing data, approximately 39% and 23%, respectively. Decisions regarding the handling of these missing values were important in order to preserve the integrity of the dataset.

- Renaming Features: Features were renamed with more descriptive titles to enhance interpretability. This step was crucial for understanding the financial indicators used in plots.

- Handling Imbalanced Data: We have noticed that the dataset was highly imbalanced, which can false results on classification models. So, we decided to under sample the majority class.

Exploratory Data Analysis

- Visualizing Missing Data: Heatmaps were generated to visualize the distribution of missing values across different years and correlation between variables. This step helped in understanding the extent and pattern of missing data in the dataset and their balance.

- Statistical Summary: Basic statistical analyses were conducted to understand the distribution, variance, and outliers in the data. This included examining measures like mean, median, and standard deviation for each feature.

Model Selection and Evaluation

- We have tested 4 different models for Knn and Mean imputation to decide which is the best : Knn Classifier, Logistic Regression, Random Forest and Decision Tree.

- Random Forest Classifier: A RandomForestClassifier was selected for its effectiveness in handling the dataset. GridSearchCV was used for hyperparameter tuning, optimizing parameters.

- Model Evaluation Metrics: The models were evaluated using accuracy and F1 score. The F1 score was particularly important for balancing the precision.

Results and Interpretation

- Model Performance: The best-performing model parameters were identified through GridSearchCV. The RandomForest model showed promising results in predicting bankruptcy with high accuracy and F1 score.

Conclusion

The analysis of the Polish Bankruptcy dataset demonstrated the potential of machine learning in predicting bankruptcy. The RandomForest model, with optimized hyperparameters, proved effective in this task. The findings from this analysis could inform risk assessment strategies in the financial sector.

To go further…

Future analysis could explore other modeling techniques, like neural networks or ensemble methods, to potentially improve prediction accuracy. Additionally, deeper analysis into the impact of missing data and more sophisticated imputation methods could further refine the model's performance.