Assignment 1: Supervised learning – Classification Report

Styliani Kalaitzaki, aid24009

MLCV 2023

Introduction

The analysis and prediction of corporate bankruptcy represent a crucial field within the economic sector, enabling the anticipation of impacts on economies and investors alike. This report constitutes a comprehensive assessment of the capability of various classification techniques to address the complex issue of identifying companies at risk of bankruptcy.

The framework for this analysis is provided through pre-existing data supplied by a recognized organization, with the objective of comparing different classification techniques in this intricate domain.

Classification Models Used

a. Linear Discriminant Analysis (LDA)

- Description: LDA is a statistical method that finds the linear combination of features that characterizes or separates two or more classes.

b. Logistic Regression (LogReg)

- Description: Logistic Regression models the probability of a binary outcome and is widely used for classification tasks.

c. Decision Trees (DTrees)

- Description: Decision Trees use a tree-like model of decisions to predict the target value.

d. Random Forests (RForest)

- Description: Random Forests are an ensemble learning method that constructs a multitude of decision trees and merges them together.

e. k-Nearest Neighbors (KNN)

- Description: k-NN classifies data points based on the majority class of their k-nearest neighbors.

f. Naïve Bayes

- Description: Naïve Bayes is a probabilistic classifier based on applying Bayes' theorem with strong independence assumptions.

g. Support Vector Machines (SVM)

- Description: SVM constructs a hyperplane in a high-dimensional space to separate classes.

h. Neural Network

- Description: Neural Networks are composed of layers of interconnected nodes, mimicking the structure of the human brain.

Results

The Impact of the Unbalanced Dataset on Classifier Performance

An unbalanced dataset can significantly impact the performance of a classifier, introducing challenges that may compromise its ability to generalize effectively. In this dataset, the distribution of classes is skewed, with one class being underrepresented (bankrupt) compared to the other (healthy). This imbalance has lead the classifiers to exhibit biased behavior, favoring the majority class and neglecting the minority class. As a result, the model may achieve high accuracy on the majority class but struggle to accurately predict instances from the minority class, which is the primary objective. The consequence is a decreased overall performance and compromised predictive power, especially for the less prevalent class. Classifiers trained on the unbalanced dataset exhibit a tendency to classify instances into the majority class, as it minimizes the training error.

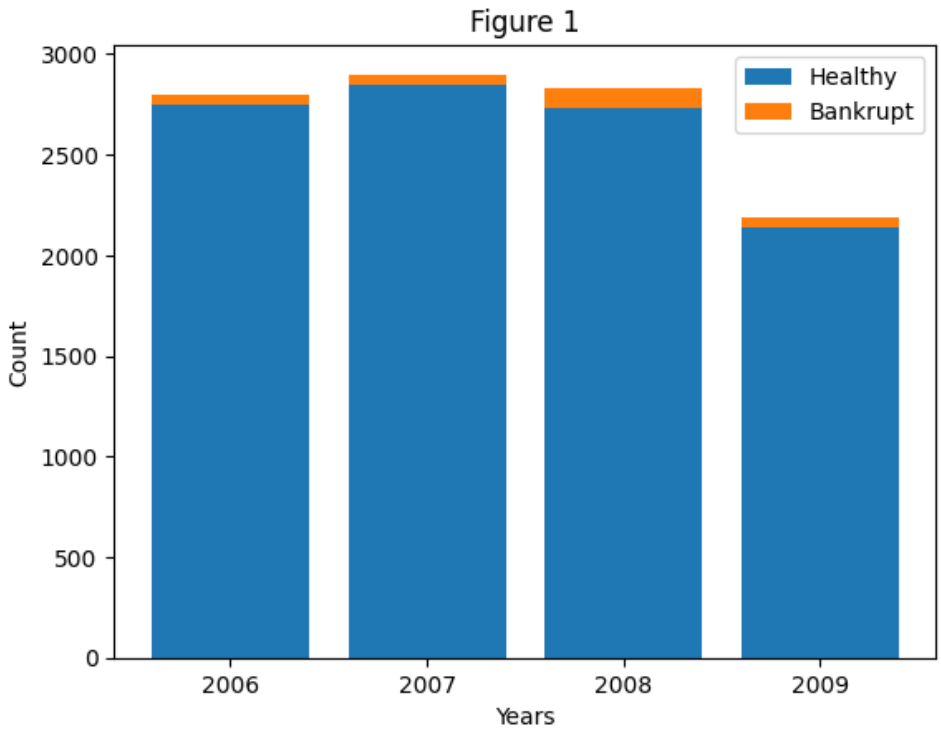

Figure 1: Number of healthy and Bamkrupt Companies per year

As illustrated in the figure below, both recall and specificity exhibit decline when applied to the unbalanced dataset, in contrast to the balanced dataset.

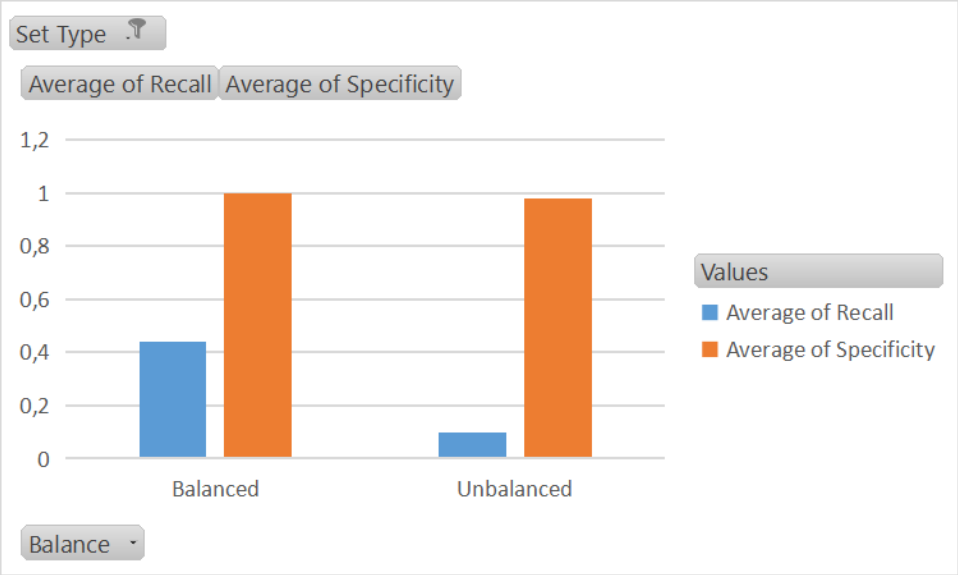

Figure 2: Average Recall and Specificity based on dataset balance

Difference in Performance between Train and Test set

The observed phenomenon of better performance on the training set compared to the test set is indicative of overfitting. This occurs when a model learns the training data too well, capturing noise and specificities that may not generalize well to unseen data. In the context of bankruptcy classification, this discrepancy suggests that the classifiers are memorizing the patterns present in the training set rather than learning the underlying relationships within the data. Consequently, when presented with new, unseen instances in the test set, the classifiers may struggle to generalize effectively, leading to a drop in performance.

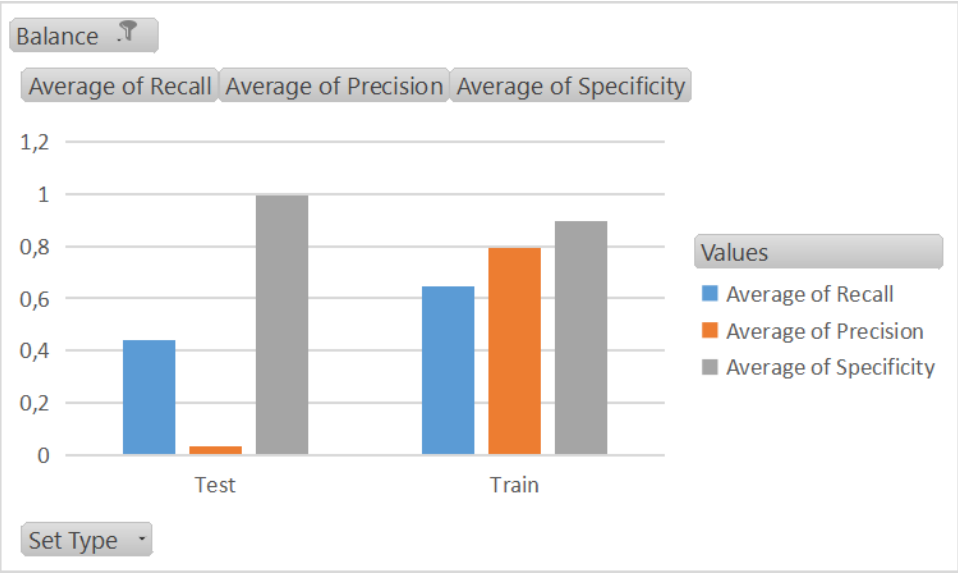

Figure 3: Train-Test Performance (balanced)

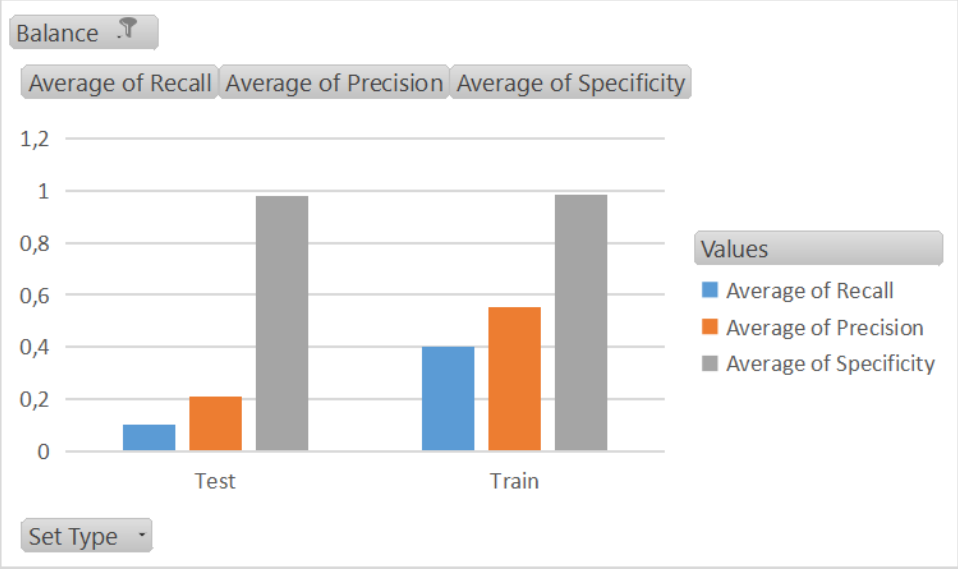

Figure 4: Train-Test Performance (unbalanced)

Below, is illustrated that despite the high recall values achieved by Dtrees and Random Forests on the training set, they fail to do so on the test set.

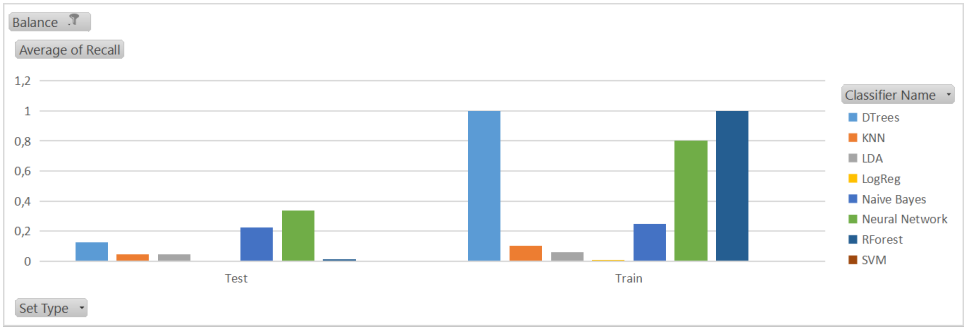

Figure 5: Average Recall Train vs Test per classifier (unbalanced)

The Decision Trees classifier demonstrated a moderate ROC-AUC score of 0.6019 on the test set, suggesting a reasonable ability to distinguish between healthy and bankrupt companies. However, the perfect score of 1 on the training set raises concerns about potential overfitting.

Decision trees are susceptible to overfitting, meaning they can memorize the training set, producing an ideal ROC-AUC score. This scenario may not reflect the model's true predictive capabilities, as it struggles when presented with new data. Alternatively, the perfect score on the training set could stem from perfect separation. In such cases, decision trees can create splits that precisely discriminate between classes in the training set, leading to an artificially high ROC-AUC that might not be indicative of the model's performance on real-world data.

Although, Decision trees are more prone to overfitting than Random forests that mitigate this issue by building multiple trees with different subsets of data and features, their results in this case are almost identical.

KNN exhibited a decent ROC-AUC of 0.5918 on the test set, demonstrating a fair discriminatory ability. The relatively close scores between the training and test sets suggest a balanced model, but the scores are not exceptionally high, indicating room for improvement.

LDA and Logistic Regression produced similar and modest ROC-AUC scores on both the test and training sets, suggesting a consistent but not highly discriminative performance.

The Neural Network classifier demonstrated the highest ROC-AUC score on the test set (0.7071), indicating a strong ability to discriminate between healthy and bankrupt companies. The notably higher score on the training set (0.9267) suggests the potential for overfitting.

Random Forests and SVM showed relatively comparable ROC-AUC scores on the test set. However, the perfect score of 0.9997 on the training set for

Random Forests indicates a high likelihood of overfitting.

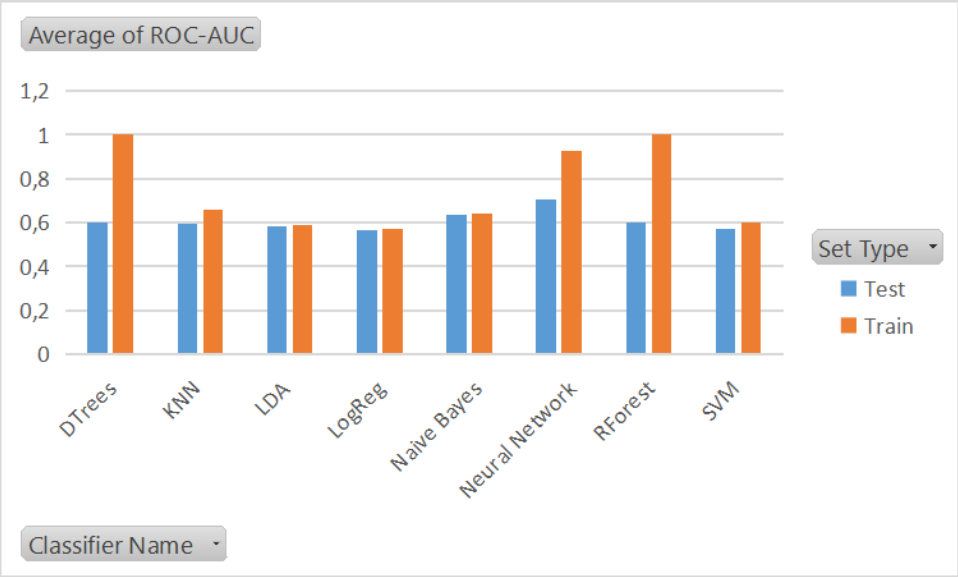

Figure 6: Average ROC-AUC

Accuracy

Accuracy alone is not a reliable metric for assessing the overall performance of a classifier in bankruptcy classification due to class imbalance. In this dataset, the occurrence of bankruptcies is rare compared to non-bankruptcies, resulting in a skewed dataset. A highly accurate classifier might simply predict the majority class (non-bankruptcy) for every instance, yielding a deceptively high accuracy. However, this would be of little practical use, as the primary goal is to identify and correctly predict the minority class (bankruptcy). Metrics such as precision, recall, and F1 score provide a more nuanced evaluation by considering true positive rates and false positive rates, providing a clearer understanding of a classifier's ability to identify instances of bankruptcy.

As depicted in the figure below, the accuracy values for all models are consistently high and nearly identical, limiting the insightful information we can derive from this metric.

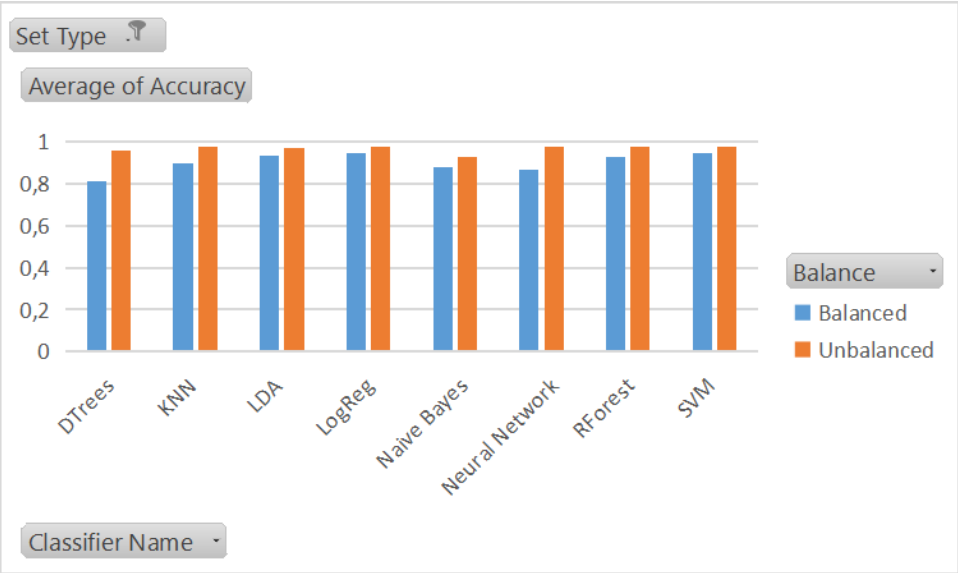

Figure 7: Average Accuracy

Precision

Upon investigating the average precision it becomes intriguing to note that the highest values are consistently attained on unbalanced data. Among the models examined, the noteworthy standout is the Random Forests followed by Neural Network and KNN.

It's not uncommon for models to achieve higher average precision on unbalanced datasets. Since, in the unbalanced dataset, there are more instances of one class than the other, making it easier for the model to achieve high precision by correctly identifying the dominant class. However, this doesn't necessarily mean the model is performing well overall, as it may not be effectively capturing the minority class, which, in this context, represents the crucial class we aim to predict.

The fact that Random Forests is performing well on unbalanced data may be attributed to its ensemble nature, which combines multiple decision trees to mitigate overfitting and capture complex relationships in the data. Neural Networks, being powerful, might excel with more data or intricate patterns or different architecture. KNN's performance could vary based on the dataset's characteristics, such as dimensionality and density.

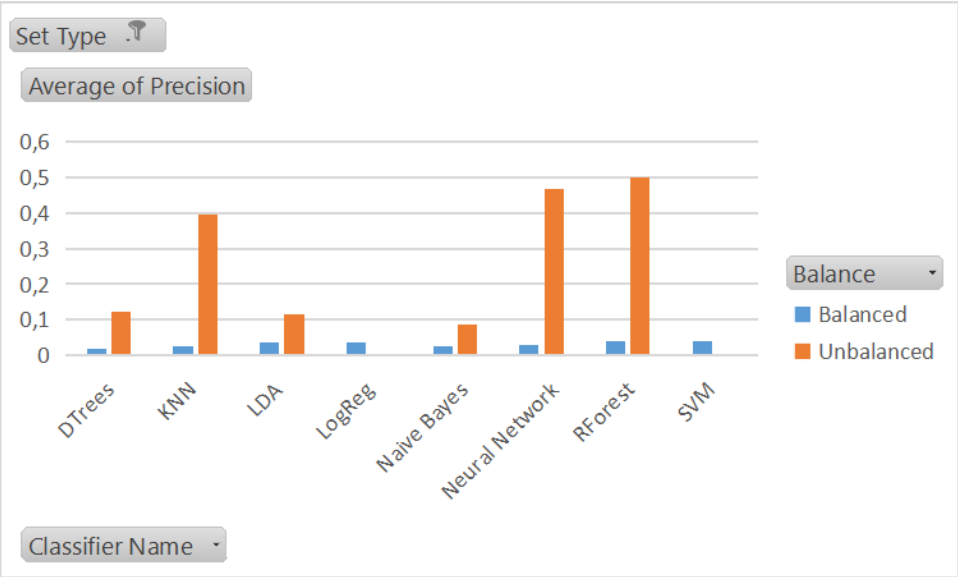

Figure 8: Average Precision

Roc-Auc

In essence, ROC-AUC summarizes the trade-off between sensitivity (recall) and specificity across different threshold values, providing a single value to assess the overall performance of a binary classification model. It measures how well the model can distinguish between two classes.

Notably, the Neural Networks and Random Forests models outperformed others on balanced data, scoring 0.75 and 0.695, respectively. On the contrary, Logistic Regression and SVM models struggled to discriminate between classes, exhibiting lower scores of 0.628 and 0.646, respectively.

In the context of unbalanced data, the Neural Network model emerged as the top performer again with an impressive ROC-AUC of 0.749. Despite the challenges posed by imbalanced datasets, the Naive Bayes model also demonstrated resilience with a score of 0.585. However, other models faced difficulties in discriminating between classes on unbalanced data.

In summary, a ROC-AUC over 0.7 achieved by the Neural Network model is a positive sign, indicating that the model has a reasonably good ability to discriminate between classes.

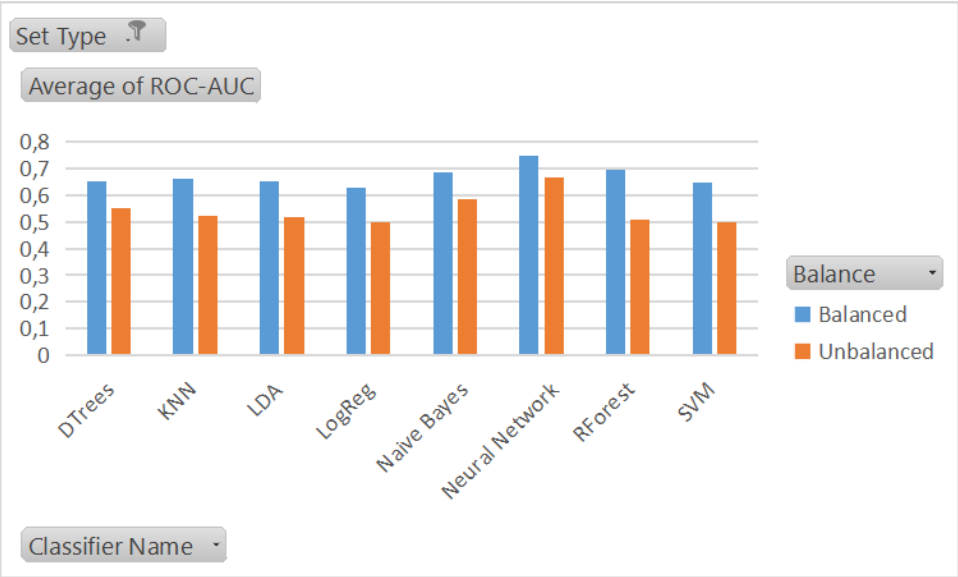

Figure 9: Average ROC-AUC

F1 Score

Presenting the F1 score is essential when evaluating the performance of a binary classification model, as it offers a comprehensive assessment that balances precision and recall. Unlike accuracy, which might be misleading in the presence of unbalanced datasets, the F1 score considers both false positives and false negatives, providing a more nuanced understanding of a model's effectiveness. This metric is particularly valuable when the consequences of missing positive instances or incorrectly classifying negative instances are significant. By encapsulating the trade-off between precision and recall in a single metric, the F1 score offers a more holistic measure of a model's overall accuracy in distinguishing between positive and negative classes.

All models exhibit an F1 score below 0.5 which generally suggests that the models are not performing well, as it means that they are not achieving a good balance between precision and recall.

This is because the primary focus was placed on achieving high recall, aiming to minimize the likelihood of false negatives and the potential oversight of companies at risk of bankruptcy. This emphasis on recall came with a trade-off, however, as the model exhibited lower precision. The decision to prioritize recall was driven by the critical nature of correctly identifying companies facing financial distress to prevent false negatives, which could have severe consequences. Consequently, the compromise led to a lower overall F1 score, reflecting the delicate balance between precision and recall in the context of bankruptcy pre-

diction, where avoiding false negatives is paramount.

However, the Neural Network model stands out again with the highest average F1 score, showcasing its effectiveness in handling imbalanced datasets and capturing positive instances.

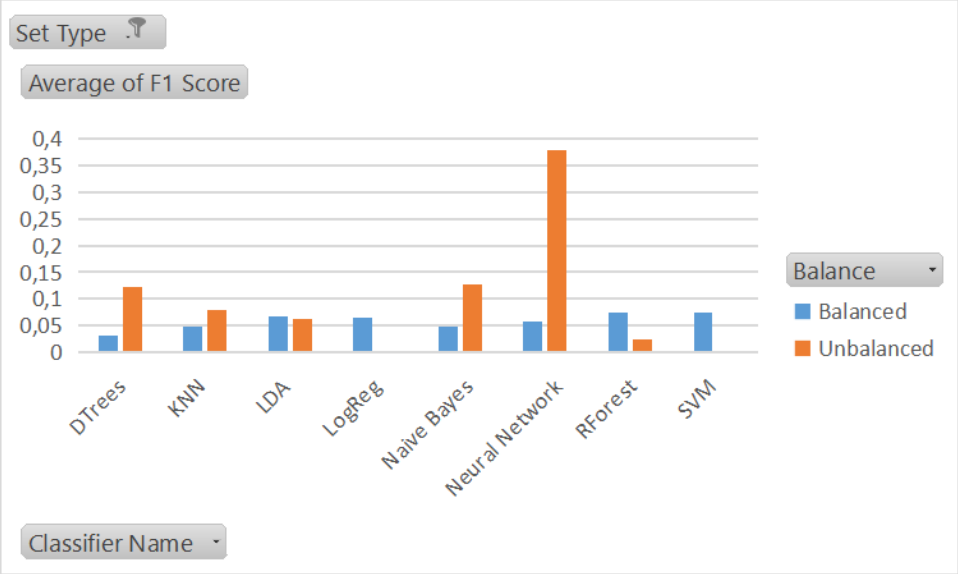

Figure 10: Average F1 Score

Informedness

This metric takes into account both sensitivity and specificity, offering a nuanced understanding of the model's performance by considering both the true positive and true negative rates. In essence, the informedness metric provides a more robust and insightful assessment of the bankruptcy classification model, enhancing the transparency and reliability of the findings presented.

The neural network model achievied the highest values across both balanced and unbalanced datasets. The model's superior performance on both balanced and unbalanced datasets indicates its robustness in handling varying class distributions. This versatility is crucial in real-world scenarios where imbalances in data are common. The neural network's capacity to discern between classes, striking a balance between sensitivity and specificity, showcases its adaptability and generalization capabilities. Such consistent satisfactory performance on diverse datasets underscores the reliability and effectiveness of the neural network in classification tasks, making it a compelling choice.

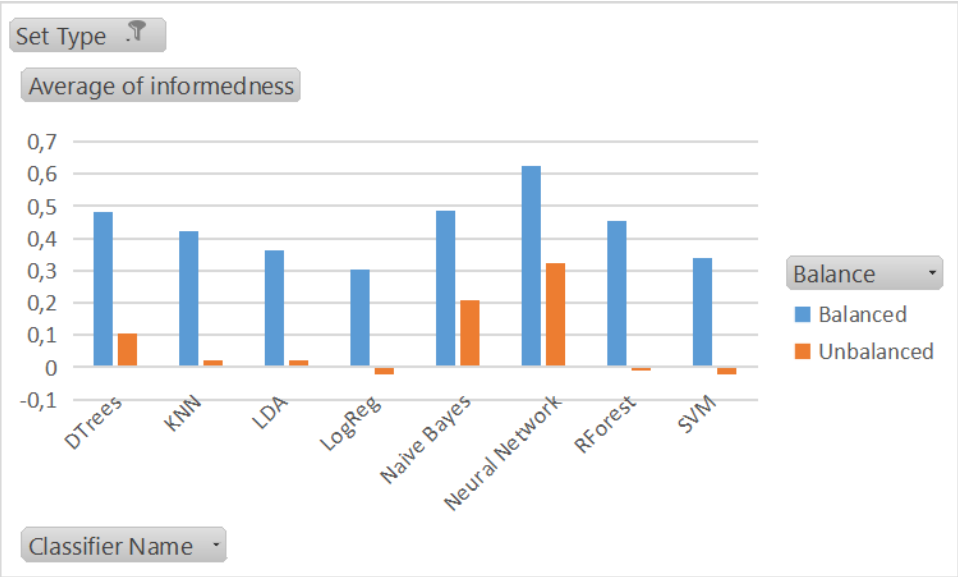

Figure 11: Average Informedness

Is there a model that satisfies the two performance constraints? Are there metrics related to them?

1. Constraint 1 is referring to recall, and as illustrated in Figure 12, no model meets this criterion on the unbalanced data. However, upon data balancing, the Neural Network stands out as the sole model that barely achieves an average recall of over 60% . This significant improvement highlights the importance of addressing class imbalances in the dataset, as it allows the Neural Network to better capture patterns associated with the minority class. While other models may have struggled to meet the recall criterion on both data, the Neural Network's adaptability to the balanced dataset showcases its potential for addressing class imbalances and improving overall performance.

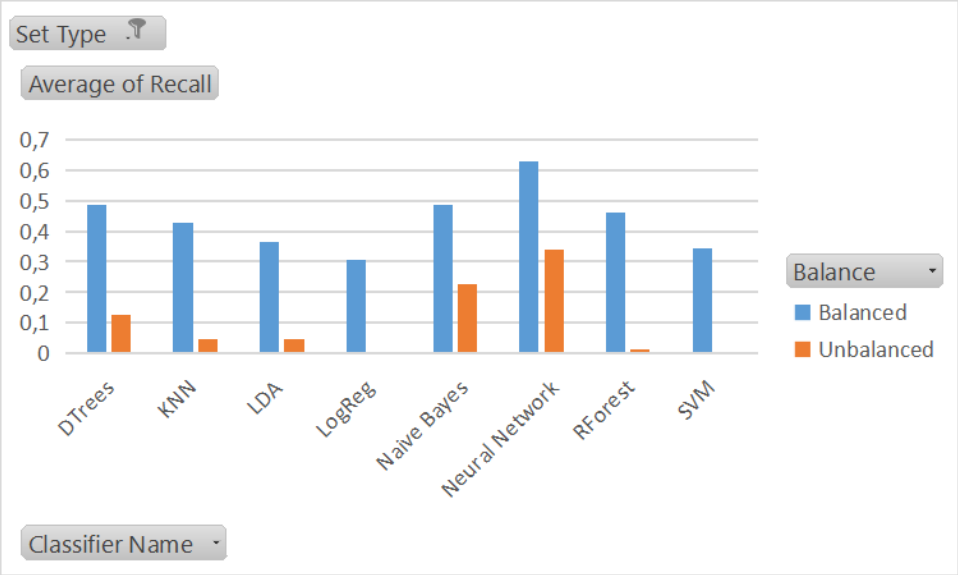

Figure 12: Average Recall

2. Constraint 2 is referring to specificity, and as depicted in Figure 13, all models, on both unbalanced and balanced data, surpass this criterion. Notably, a higher value is observed on the balanced dataset. Despite only a slight improvement, the Neural Network model outperforms others by achieving the highest values on both datasets.
The consistent surpassing of the specificity criterion by all models suggests that they effectively distinguish the majority class from the minority class.
The Neural Network's consistent superiority in achieving the highest specificity values indicates its robustness in correctly classifying instances of the majority class. This strength, coupled with its notable improvement in recall on the balanced dataset, positions the Neural Network as a promising candidate for applications where both recall and specificity are critical performance indicators.

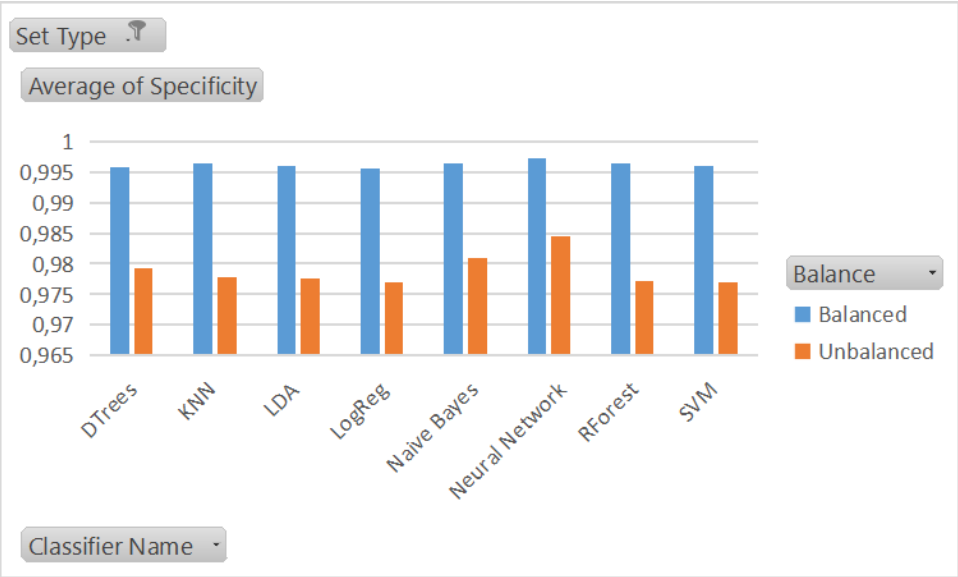

Figure 13: Average Specificity

Upon experimenting with a 0.3 threshold in the neural network and comparing the results with those obtained using a 0.5 threshold (above case):

The Neural Network model demonstrates enhanced average recall with a 0.3 threshold when applied to balanced data, while maintaining consistent specificity. Notably, there is no discernible variance in average recall when applied to unbalanced data.

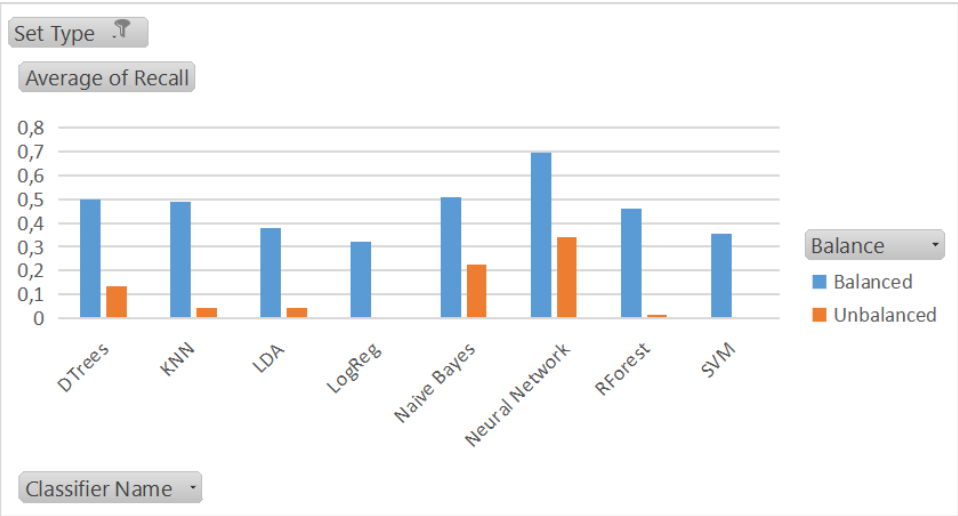

Figure 14: Average Recall

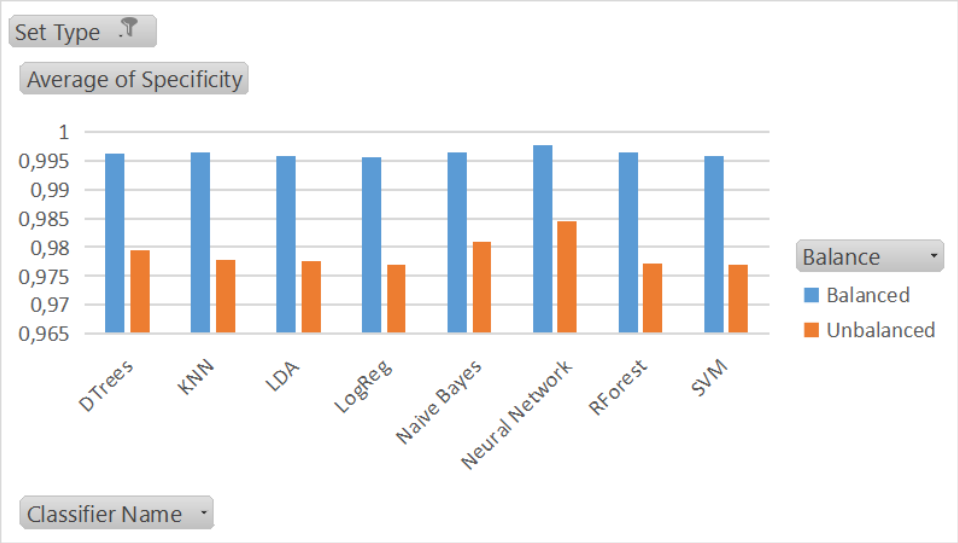

Figure 15: Average Specificity

Conclusion

[Selecting the Best Model]

The neural network model showcases outstanding performance within the dataset, specifically excelling in recall and roc-auc compared to alternative models. Given the objective of accurately identifying companies at risk of bankruptcy, emphasizing recall is paramount. The model's ability to effectively capture and retrieve instances of impending financial distress sets it apart, making it the optimal choice for the critical task of discerning potential bankruptcies within the dataset. Its consistent performance across various metrics further solidifies its standing, leaving no compelling reason to overlook or disregard its selection.