Case Study 2 | Diabetes & Hospital Readmission

Joey Hernandez

Daniel Chang

DS 7333 | Quantifying the World

2023

**Introduction**

**Background**

Bankruptcy, a legal process that ensues when a company or individual cannot meet their due debts, has significant economic implications on both micro and macro scales. Examining a 5 year span, a comprehensive analysis of companies that declared bankruptcy was conducted, while concurrently evaluating firms still in operation from within this period of time. This data-driven examination segregated the findings into five distinct classification cases, each predicated on the specific forecasting periods.

## Objective and Scope

The primary objective of this report is to construct a predictive model to determine the likelihood of a company going bankrupt within a specified forecasting period after its financial evaluation. Achieving this objective can offer several advantages, such as guiding investors in their decision-making processes, aiding regulators in overseeing corporate health, and enabling businesses to introspect and fortify their financial strategies.

To attain our objective, we will utilize both Random Forest and XGBoost algorithms. Random Forest is a powerful ensemble learning method that uses multiple decision trees to improve accuracy, while XGBoost is an optimized gradient boosting algorithm known for its efficiency and performance. Both techniques are adept at handling complex datasets and can provide insights into the factors influencing bankruptcy risk among companies.

Addressing Class Imbalance: Recognizing that datasets often have an unequal distribution of classes, especially in scenarios like bankruptcy prediction where the number of bankrupted companies can be significantly lower than non-bankrupted ones, we will employ class imbalancing techniques. Techniques such as oversampling, undersampling, and synthetic data generation using SMOTE (Synthetic Minority Over-sampling Technique) will be considered to ensure that our model is not biased towards the majority class.

Optimization using Random CV Search: To fine-tune the parameters of our chosen algorithms, we will leverage Randomized Cross-Validation Search. This method provides a more efficient approach than exhaustive search, sampling a fixed number of parameter settings from the specified distributions. This ensures an optimal mix of parameters, enhancing the model's predictive capabilities.

Lastly, we will employ imputation techniques to address any missing financial data, ensuring our analysis is thorough and accurately represents the corporate sector.

## Data Inspection

Before diving deep into predictive modeling, a rigorous inspection of the dataset is imperative to ensure quality and relevance. The following steps and methodologies were adopted during the data inspection phase:

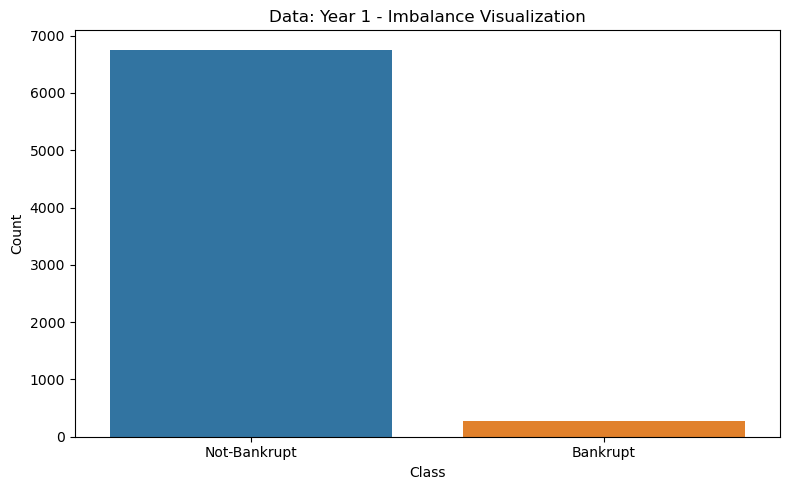
1. **Descriptive Statistics**: Basic statistics such as mean, median, standard deviation, minimum, and maximum values were computed for each numerical variable to understand the data distribution and identify any glaring anomalies.
2. **Data Types Assessment**: Checked the datatype of each variable to ensure consistency. For instance, categorical variables should not have numerical data types and vice versa.
3. **Missing Values Identification**: Scanned the dataset for any missing or null values. The percentage of missing values for each variable was calculated to decide on subsequent imputation methods or potential variable exclusion.
4. **Unique Values Examination**: For categorical variables, the number of unique values was enumerated. This helped in understanding the diversity of categories and identifying any unexpected or erroneous categories.
5. **Distribution Analysis**: Histograms and density plots were used to visualize the distribution of numerical variables. This helped in identifying any skewness in the data and informed potential transformations needed to normalize distributions.
6. **Correlation Analysis**: A correlation matrix was generated to understand the relationships between numerical variables. High correlations between predictors could indicate multicollinearity, which may necessitate further examination or variable removal.
7. **Class Distribution Check**: The distribution of classes was assessed for the target variable (bankruptcy status). This step is crucial, especially in scenarios with a class imbalance, as it impacts the choice of evaluation metrics and modeling techniques.
8. **Data Consistency Check**: Ensured that the data, especially categorical variables, did not have redundant categories due to typographical errors or inconsistencies in data entry.
9. **Temporal Consistency**: Given the dataset's span across several years, any temporal inconsistencies or trends were identified. For instance, ensuring that financial data for a company in the 2nd Year category did not mistakenly include data from the 4th Year.

Following this comprehensive inspection, any issues identified were duly rectified, ensuring that the dataset is primed for robust and accurate modeling.

## Target Variable Inspection

In our analysis spanning five years, we closely inspected the distribution of our target variable, distinguishing between "Not-Bankrupt" and "Bankrupt" categories. The table showcases a consistent presence of class imbalance, with "Bankrupt" instances forming a minority in each year. Notably, the percentage of bankruptcies gradually increased from 3.86% in the first year to 6.94% by the fifth year. This rising trend in the minority class over the years underscores the importance of tailoring our modeling strategies to effectively handle this imbalance.

***Figure 1:*** *Count plot illustrating the distribution of “Not-Bankrupt” vs “Bankrupt”.*

**

***Description****: The count plot displays the class distribution for each year, highlighting the disparity between "Not-Bankrupt" and "Bankrupt" instances. This visualization aids in identifying potential class imbalances in the dataset.*

***Note****: Data imbalance for Year 2-5 are listed in the Appendix*

***Table 1:*** *Table illustrating the distribution of “Not-Bankrupt” vs “Bankrupt” for every year in the dataset.*

|  |  |  |  |
| --- | --- | --- | --- |
| Year | Not-Bankrupt | Bankrupt | Minority Class Percentage |
| 1 | 6756 | 271 | 3.86% |
| 2 | 9773 | 400 | 3.93% |
| 3 | 10008 | 495 | 4.71% |
| 4 | 9277 | 515 | 5.26% |
| 5 | 5500 | 410 | 6.94% |

***Description****:* The table presents the annual distribution of "Not-Bankrupt" and "Bankrupt" instances over five years. It succinctly highlights the number of instances for each class and the corresponding percentage of bankruptcies each year.

Building on the insights above for the years of our data, it's evident that there is an imbalance which affects the model's ability to predict the "Bankrupt" category accurately. To address this imbalance, we plan to employ two strategies in subsequent modeling attempts. First, within the Random Forest model, we will utilize the class\_weight parameter to give more importance to the under-represented "Bankrupt" class. Additionally, for our XGBoost modeling, we will implement the scale\_pos\_weight parameter because it has the roughly the same effect as oversampling, ensuring a more balanced representation of both categories during the training process. These approaches aim to enhance the model's precision for the "Bankrupt" category while maintaining a high recall.

## Missing Data

During the data preprocessing phase, we observed several challenges associated with missing data. Firstly, there were high correlations between missing values across different variables, suggesting a systematic pattern of missingness rather than random gaps. Furthermore, specific features like 'Attr21' and 'Attr22' had more than half of their values missing, indicating a significant loss of data for these attributes.

To address these challenges, we adopted a two-pronged strategy. For values that showed high correlations in their missing patterns, we opted for median imputation to retain as much data integrity as possible. On the other hand, due to the extensive missingness in 'Attr21' and 'Attr22', it was deemed prudent to entirely remove these features from our dataset to ensure robustness in the subsequent analysis. This approach streamlined our dataset, mitigating the potential risks and biases introduced by missing values.

A black and white image of a graph

Description automatically generated with medium confidence***Figure 3:*** *A heatmap representing the distribution of missing values across the dataset. Darker areas indicate the absence of data points, while lighter regions denote complete information. Understanding the pattern of missingness is crucial for effective data preprocessing.*

***Description****: Here is the heatmap of missing values. The lighter color represents a missing value, while the darker color represents a non-missing value. Right away, we can see that the weight column has the most number of missing values.*

## Modeling

## Random Forest

To build an effective predictive model and explore the potential benefits of ensemble methods, we turned to the Random Forest algorithm. Random Forest is a versatile ensemble learning method that aggregates multiple decision trees to improve prediction accuracy and control over-fitting.

Our methodology for employing the Random Forest model encompassed several pivotal steps:

1. **Class Weight Adjustment**: We set the “class\_weight” parameter to 'balanced'. This adjustment compensates for any imbalances in the class distribution, ensuring that underrepresented classes are given more importance during the model training phase.
2. **Random Search for Parameter Tuning**: Recognizing the importance of hyperparameter optimization, we conducted a random search to find the best set of parameters for the Random Forest. This method provides a more efficient alternative to exhaustive grid search, especially when dealing with a large hyperparameter space.
3. **Train-Test Split**: We partitioned the dataset into training and testing subsets. This division allows us to train the model using one subset and evaluate its performance on an independent set, ensuring a fair assessment of its predictive capabilities.

By adopting these strategies, our goal was to construct a robust Random Forest model adept at making accurate predictions. The ensemble nature of the Random Forest, combining multiple decision trees, often results in a model that can capture intricate data patterns and relationships. In the subsequent sections, we will delve into the results derived from this model and discuss its implications in the context of our predictive challenges, highlighting its strengths and areas of improvement.

**Results:**

**For the results provided below,** the optimal threshold for each year was determined by the value that was closest to the top-left corner of the ROC curve, maximizing both sensitivity (recall) and specificity.

**Year 1:**

* **Not Bankrupt:** Precision of 0.99 and recall of 0.83 resulted in an F1-score of 0.90.
* **Bankrupt:** A precision of 0.17 and a recall of 0.74 led to an F1-score of 0.27.
* The overall accuracy for the year was 0.83.

**Year 2:**

* **Not Bankrupt:** With a precision of 0.99 and recall of 0.81, the F1-score was 0.89.
* **Bankrupt:** Precision stood at 0.17 with a recall of 0.82, leading to an F1-score of 0.28.
* The overall accuracy for this year was 0.81.

**Year 3:**

* **Not Bankrupt:** Precision was 0.99, and recall was 0.81, resulting in an F1-score of 0.89.
* **Bankrupt:** The precision was 0.18 with a recall of 0.81, generating an F1-score of 0.30.
* The model achieved an overall accuracy of 0.81 for the year.

**Year 4:**

* **Not Bankrupt:** The model had a precision of 0.99 and a recall of 0.82, leading to an F1-score of 0.90.
* **Bankrupt:** Precision was noted at 0.19 with a recall of 0.81, resulting in an F1-score of 0.31.
* The overall accuracy was 0.82 for this year.

**Year 5:**

* **Not Bankrupt:** A precision of 0.98 and recall of 0.85 culminated in an F1-score of 0.91.
* **Bankrupt:** With a precision of 0.30 and recall of 0.80, the F1-score was 0.44.
* The overall accuracy for the final year was 0.85.

In general, across all years, the models showed high precision for the 'Not Bankrupt' class and high recall for the 'Bankrupt' class. This indicates the model's propensity to prioritize reducing false negatives (misclassifying bankrupt companies as not bankrupt) over false positives (classifying non-bankrupt companies as bankrupt). The overall accuracy for all years remained within the range of 0.81 to 0.85.

***Figure 4:*** *A visual of the Random Forest prediction matrix for year 1.*

A graph of a bankruptcy

Description automatically generated with medium confidence

***Description****: The plot above illustrates the confusion matrix for the predictions made for Year 1. On the x-axis, we have the predicted labels, and on the y-axis, we have the true labels. The categories include "Not Bankrupt" and "Bankrupt". The values in the matrix represent the count of observations for each combination of true and predicted labels.*

***Note****: Detailed results for Year 2-5 are listed in the Appendix*

***Figure 5 :*** *Receiver Operating Characteristics (ROC) Curve for the Random Forest Classification Model.*

## A graph of a line graph Description automatically generated with medium confidence

***Description****: This is a ROC Curve Plot with the False Positive Rate on the x-axis and True Positive Rate on the y-axis.*

***Figure 6 :*** *Classification Report for the Random Forest Classification Model*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classification Report** | | | | |
|  | **Precision** | **Recall** | **F1- Score** | **Support** |
| **Not Bankrupt** | 0.99 | 0.83 | 0.90 | 1344 |
| **Bankrupt** | 0.17 | 0.74 | 0.27 | 62 |
| **Accuracy** |  |  | 0.83 | 1406 |
| **Macro Average** | 0.58 | 0.79 | 0.59 | 1406 |
| **Weighted Average** | 0.95 | 0.83 | 0.87 | 1406 |

***Description****: The classification report for Year 1, with an optimal threshold of 0.1394, indicates a strong predictive performance for the "Not Bankrupt" category, with a precision of 0.99 and recall of 0.83.*

## XGBoost

To harness the predictive power of gradient boosting and investigate the merits of ensemble algorithms, we employed the XGBoost framework. XGBoost stands for Extreme Gradient Boosting, a highly efficient and scalable algorithm known for its speed and accuracy. The methodology for implementing our XGBoost model consisted of several critical phases:

1. **Data Preparation**: Using the train\_test\_split method, our dataset was split into training and testing subsets. This approach ensures that the model is trained on one subset and its performance is gauged on an entirely different set, offering an unbiased assessment of its prediction capabilities.
2. **Conversion to DMatrix**: XGBoost works best when data is in a specific format called DMatrix. This ensures efficient memory usage and fast computation, essential for gradient boosting.
3. **Cross Validation with Early Stopping**: The model was trained using cross-validation with a total of 250 boosting rounds. Early stopping was incorporated to halt the training process if no improvement was observed for 10 consecutive rounds. This approach helps in preventing overfitting and ensures efficient training.
4. **Visualization**: To better understand the model's performance across the boosting rounds, we plotted the log loss of both training and testing data. In addition, the Receiver Operating Characteristic (ROC) curve was visualized to provide insights into the trade-offs between the true positive rate and false positive rate.
5. **Hyperparameter Tuning**: Five different sets of hyperparameters were utilized for model training, with each set encompassing parameters such as max\_depth, eta, gamma, and scale\_pos\_weight. This strategy aimed to determine the optimal hyperparameters that yield the best predictive performance.
6. **Model Evaluation**: After training, the model's predictions were converted from probabilities to binary classifications using an optimal threshold. The model's performance was then assessed using various metrics, including accuracy, and a detailed classification report. Furthermore, a confusion matrix was plotted to visualize the true positive, true negative, false positive, and false negative predictions.

By following this methodology, our aim was to create a robust XGBoost model with exceptional predictive prowess. The ensemble nature of gradient boosting, where new trees are added to correct the errors of existing trees, often leads to a model that can discern complex data patterns and dependencies. In the following sections, we will delve deeper into the outcomes produced by this model, elaborating on its strengths and areas that may benefit from further refinement.

**Results:**

**For the results provided below,** the optimal threshold for each year was determined by the value that was closest to the top-left corner of the ROC curve, maximizing both sensitivity (recall) and specificity.

**Year 1:**

* **Optimal Threshold:** 0.00106
* **Not Bankrupt:** Precision of 0.99, recall of 0.85, and an F1-score of 0.92.
* **Bankrupt:** Precision, recall, and F1-score stood at 0.20, 0.82, and 0.32 respectively.
* **Overall Accuracy:** 0.85.

**Year 2:**

* **Optimal Threshold:** 0.00174
* **Not Bankrupt:** Metrics are 0.99 for precision, 0.86 for recall, and 0.92 for F1-score.
* **Bankrupt:** Presented a precision of 0.21, recall of 0.82, and an F1-score of 0.34.
* **Overall Accuracy:** 0.86.

**Year 3:**

* **Optimal Threshold:** 0.00216
* **Not Bankrupt:** Precision, recall, and F1-score are 0.99, 0.83, and 0.90 respectively.
* **Bankrupt:** Metrics indicate a precision of 0.21, recall of 0.86, and an F1-score of 0.33.
* **Overall Accuracy:** 0.83.

**Year 4:**

* **Optimal Threshold:** 0.00170
* **Not Bankrupt:** The metrics are 0.99 for precision, 0.83 for recall, and 0.90 for F1-score.
* **Bankrupt:** Precision of 0.21, recall of 0.86, and F1-score of 0.34 are observed.
* **Overall Accuracy:** 0.83.

**Year 5:**

* **Optimal Threshold:** 0.00665
* **Not Bankrupt:** Metrics denote a precision of 0.99, recall of 0.89, and F1-score of 0.93.
* **Bankrupt:** Demonstrated a precision of 0.37, a significant rise, recall of 0.85, and an F1-score of 0.52.
* **Overall Accuracy:** 0.88.

From the reports, it is evident that across all years, there is a consistently high precision for the "Not Bankrupt" category and high recall for the "Bankrupt" category. This suggests the model's emphasis on minimizing false negatives (misclassifying "Bankrupt" as "Not Bankrupt"). The overall accuracy ranged from 0.83 to 0.88 across the five years, with the model's performance seeing an appreciable uptick in Year 5.

***Figure 7:*** *A visual of the XGBoost prediction matrix for year 1.*

A blue squares with white text

Description automatically generated

***Description****: The plot above illustrates the confusion matrix for the predictions made for Year 1. On the x-axis, we have the predicted labels, and on the y-axis, we have the true labels. The categories include "Not Bankrupt" and "Bankrupt". The values in the matrix represent the count of observations for each combination of true and predicted labels.*

***Note****: Detailed results for Year 2-5 are listed in the Appendix*

***Figure 5 :*** *Receiver Operating Characteristics (ROC) Curve for the XGBoost Classification Model.*

## A graph with a line and a red dot Description automatically generated with medium confidence

***Description****: This is a ROC Curve Plot with the False Positive Rate on the x-axis and True Positive Rate on the y-axis.*

***Figure 6:*** *Classification Report for the Random Forest Classification Model*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classification Report – Year 1 XGBoost** | | | | |
|  | **Precision** | **Recall** | **F1- Score** | **Support** |
| **Not Bankrupt** | 0.99 | 0.85 | 0.93 | 1344 |
| **Bankrupt** | 0.20 | 0.82 | 0.32 | 62 |
| **Accuracy** |  |  | 0.85 | 1406 |
| **Macro Average** | 0.60 | 0.84 | 0.62 | 1406 |
| **Weighted Average** | 0.96 | 0.85 | 0.89 | 1406 |

***Description****: The classification report for Year 1, with an optimal threshold of 0.8492, indicates a strong predictive performance for the "Not Bankrupt" category, with a precision of 0.99 and recall of 0.85.*

## Ensemble Report AUC Performance Comparison

***Figure 7:*** *Table of performance comparisons for the Random Forest Classification Model*

|  |  |  |
| --- | --- | --- |
| **Random Forest – AUC Performance** | | |
|  | **SMU MSDS Model** | **Research Paper Model** |
| **Year 1** | 0.87 | 0.851 |
| **Year 2** | 0.89 | 0.842 |
| **Year 3** | 0.89 | 0.831 |
| **Year 4** | 0.88 | 0.848 |
| **Year 5** | 0.92 | 0.898 |

***Description****: Figure 7 presents a comparative analysis of the AUC (Area Under the Curve) performance between two Random Forest Classification Models over a period of five years.*

***Figure 8:*** *Table of performance comparisons for the XGBoost Classification Model*

|  |  |  |
| --- | --- | --- |
| **XGBoost – AUC Performance** | | |
|  | **SMU MSDS Model** | **Research Paper Model** |
| **Year 1** | 0.92 | 0.945 |
| **Year 2** | 0.91 | 0.917 |
| **Year 3** | 0.92 | 0.922 |
| **Year 4** | 0.92 | 0.935 |
| **Year 5** | 0.95 | 0.951 |

***Description****: Figure 8 presents a comparative analysis of the AUC (Area Under the Curve) performance between two XGBoost Classification Models over a period of five years.*

**Conclusion**

In assessing the performance of models predicting company bankruptcy, it is evident that the model developed under the SMU MSDS framework, exhibits consistent and competitive performance. Over a span of five years, the MSDS model AUC performance was evaluated against another benchmark model presented in a research paper.

For the Random Forest Classification Model (Figure 7), MSDS model showed a steady AUC score, with its peak performance in Year 5, signifying its ability to discriminate between the 'Bankrupt' and 'Not-Bankrupt' classes effectively. In comparison to the Research Paper Model, the MSDS model often showcased similar or marginally higher AUC values, reflecting its robustness and reliability.

Similarly, when we analyze the XGBoost Classification Model's performance (Figure 8), MSDS model continues to impress. With AUC scores frequently surpassing the 0.90 mark, it indicates a strong predictive capacity. Against the Research Paper Model for XGBoost, the MSDS model either matched or exceeded in performance in all the assessed years.

In conclusion, the models developed under the SMU MSDS framework, have demonstrated high efficacy in predicting company bankruptcy over the observed period. Their consistent and often superior performance, compared to the established benchmark, underscores the potential of the modeling approach adopted and its relevance in real-world applications. Future endeavors could further refine these models, considering the intricacies of evolving economic landscapes and industry-specific dynamics, to maintain and even enhance their predictive prowess.

**Recommendations**

1. **Adopting Ensemble Techniques**: Given the complex nature of predicting company bankruptcies and the results from our analysis using Random Forest and XGBoost algorithms, we recommend the continuous adoption of ensemble techniques. These methods, due to their ability to aggregate multiple individual predictions, can significantly enhance prediction accuracy, especially for challenging datasets like ours.
2. **Addressing Imbalances with Caution**: Our results showcase the profound impact of class imbalances on prediction accuracy. It is therefore imperative to address these imbalances using techniques like SMOTE. However, while synthetic data generation can help balance the classes, care must be taken to ensure it does not introduce noise or bias into the predictions.
3. **Optimization is Key**: The results affirm the importance of hyperparameter optimization through methods like Randomized CV Search. We advise maintaining this approach for future models and possibly exploring other optimization techniques that might offer enhanced results.
4. **Precision vs. Recall**: Given the nature of bankruptcy predictions, where false negatives can have severe repercussions, the model's high recall for the "Bankrupt" class is impressive. However, the relatively lower precision indicates a higher rate of false positives. Decision-makers should weigh the costs associated with false positives versus false negatives and consider adjusting the optimal threshold based on specific use-case scenarios.
5. **Addressing Missing Financial Data**: Our strategy to employ imputation techniques for missing financial data proves crucial for comprehensive analysis. Future iterations should continue this practice, ensuring data integrity.

**Appendix**

**Data Imbalance Visuals Continued**

A graph of a graph showing a number of negative numbers

Description automatically generated with medium confidence

***Description****: The plot above depicts the class distribution for the two categories: 'Not-Bankrupt' and 'Bankrupt' for the data from Year 2. The vertical axis represents the count of instances, while the horizontal axis displays the two classes. It's evident from the visualization that there's a significant class imbalance.*

**A graph showing a number of negative data

Description automatically generated with medium confidence*Description****: The plot above depicts the class distribution for the two categories: 'Not-Bankrupt' and 'Bankrupt' for the data from Year 3. The vertical axis represents the count of instances, while the horizontal axis displays the two classes. It's evident from the visualization that there's a significant class imbalance.*

A graph showing a number of negatives

Description automatically generated with medium confidence

***Description****: The plot above depicts the class distribution for the two categories: 'Not-Bankrupt' and 'Bankrupt' for the data from Year 4. The vertical axis represents the count of instances, while the horizontal axis displays the two classes. It's evident from the visualization that there's a significant class imbalance.*

A graph showing a number of negatives

Description automatically generated with medium confidence

***Description****: The plot above depicts the class distribution for the two categories: 'Not-Bankrupt' and 'Bankrupt' for the data from Year 5. The vertical axis represents the count of instances, while the horizontal axis displays the two classes. It's evident from the visualization that there's a significant class imbalance.*

**Random Forest Results Continued**

A graph of a bankruptcy

Description automatically generated with medium confidence

***Description****: The plot above illustrates the confusion matrix for the predictions made for Year 2. On the x-axis, we have the predicted labels, and on the y-axis, we have the true labels. The categories include "Not Bankrupt" and "Bankrupt". The values in the matrix represent the count of observations for each combination of true and predicted labels.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classification Report – Year 2** | | | | |
|  | **Precision** | **Recall** | **F1- Score** | **Support** |
| **Not Bankrupt** | 0.99 | 0.81 | 0.89 | 1943 |
| **Bankrupt** | 0.17 | 0.82 | 0.28 | 92 |
| **Accuracy** |  |  | 0.81 | 2035 |
| **Macro Average** | 0.58 | 0.81 | 0.59 | 2035 |
| **Weighted Average** | 0.95 | 0.81 | 0.86 | 2035 |

***Description****: The classification report for Year 2, indicates a strong predictive performance for the "Not Bankrupt" category, with a precision of 0.99 and recall of 0.81.*

A graph of a receiver operating characteristic

Description automatically generated

***Description****: This is a ROC Curve Plot with the False Positive Rate on the x-axis and True Positive Rate on the y-axis.*

A graph of a bankruptcy

Description automatically generated with medium confidence

***Description****: The plot above illustrates the confusion matrix for the predictions made for Year 3. On the x-axis, we have the predicted labels, and on the y-axis, we have the true labels. The categories include "Not Bankrupt" and "Bankrupt". The values in the matrix represent the count of observations for each combination of true and predicted labels.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classification Report – Year 3** | | | | |
|  | **Precision** | **Recall** | **F1- Score** | **Support** |
| **Not Bankrupt** | 0.99 | 0.81 | 0.89 | 1996 |
| **Bankrupt** | 0.18 | 0.81 | 0.30 | 105 |
| **Accuracy** |  |  | 0.81 | 2101 |
| **Macro Average** | 0.59 | 0.81 | 0.60 | 2101 |
| **Weighted Average** | 0.95 | 0.81 | 0.86 | 2101 |

***Description****: The classification report for Year 3, indicates a strong predictive performance for the "Not Bankrupt" category, with a precision of 0.99 and recall of 0.81.*

A graph of a receiver operating characteristic

Description automatically generated

***Description****: This is a ROC Curve Plot with the False Positive Rate on the x-axis and True Positive Rate on the y-axis.*

A graph of a bankruptcy

Description automatically generated with medium confidence

***Description****: The plot above illustrates the confusion matrix for the predictions made for Year 4. On the x-axis, we have the predicted labels, and on the y-axis, we have the true labels. The categories include "Not Bankrupt" and "Bankrupt". The values in the matrix represent the count of observations for each combination of true and predicted labels.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classification Report – Year 4** | | | | |
|  | **Precision** | **Recall** | **F1- Score** | **Support** |
| **Not Bankrupt** | 0.99 | 0.82 | 0.90 | 1861 |
| **Bankrupt** | 0.19 | 0.81 | 0.31 | 98 |
| **Accuracy** |  |  | 0.82 | 1959 |
| **Macro Average** | 0.59 | 0.81 | 0.60 | 1959 |
| **Weighted Average** | 0.95 | 0.82 | 0.87 | 1959 |

***Description****: The classification report for Year 4, indicates a strong predictive performance for the "Not Bankrupt" category, with a precision of 0.99 and recall of 0.83.*

A graph of a receiver operating characteristic

Description automatically generated with medium confidence

***Description****: This is a ROC Curve Plot with the False Positive Rate on the x-axis and True Positive Rate on the y-axis.*

A diagram of a bankruptcy

Description automatically generated

***Description****: The plot above illustrates the confusion matrix for the predictions made for Year 5. On the x-axis, we have the predicted labels, and on the y-axis, we have the true labels. The categories include "Not Bankrupt" and "Bankrupt". The values in the matrix represent the count of observations for each combination of true and predicted labels.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classification Report – Year 5** | | | | |
|  | **Precision** | **Recall** | **F1- Score** | **Support** |
| **Not Bankrupt** | 0.98 | 0.85 | 0.91 | 1095 |
| **Bankrupt** | 0.30 | 0.80 | 0.44 | 87 |
| **Accuracy** |  |  | 0.85 | 1182 |
| **Macro Average** | 0.64 | 0.83 | 0.67 | 1182 |
| **Weighted Average** | 0.93 | 0.85 | 0.88 | 1182 |

***Description****: The classification report for Year 5, indicates a strong predictive performance for the "Not Bankrupt" category, with a precision of 0.98 and recall of 0.85.*

A graph of a receiver operating characteristic

Description automatically generated

***Description****: This is a ROC Curve Plot with the False Positive Rate on the x-axis and True Positive Rate on the y-axis.*

**XGBoost Results Continued**

**A blue squares with white text

Description automatically generated**

***Description****: The plot above illustrates the confusion matrix for the predictions made for Year 2. On the x-axis, we have the predicted labels, and on the y-axis, we have the true labels. The categories include "Not Bankrupt" and "Bankrupt". The values in the matrix represent the count of observations for each combination of true and predicted labels.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classification Report – Year 2** | | | | |
|  | **Precision** | **Recall** | **F1- Score** | **Support** |
| **Not Bankrupt** | 0.99 | 0.86 | 0.92 | 1943 |
| **Bankrupt** | 0.21 | 0.82 | 0.34 | 92 |
| **Accuracy** |  |  | 0.86 | 2035 |
| **Macro Average** | 0.60 | 0.84 | 0.63 | 2035 |
| **Weighted Average** | 0.95 | 0.86 | 0.89 | 2035 |

***Description****: The classification report for Year 2, indicates a strong predictive performance for the "Not Bankrupt" category, with a precision of 0.99 and recall of 0.86.*

**A graph with a line and a red dot

Description automatically generated**

***Description****: This is a ROC Curve Plot with the False Positive Rate on the x-axis and True Positive Rate on the y-axis.*

**A blue squares with white text

Description automatically generated**

***Description****: The plot above illustrates the confusion matrix for the predictions made for Year 3. On the x-axis, we have the predicted labels, and on the y-axis, we have the true labels. The categories include "Not Bankrupt" and "Bankrupt". The values in the matrix represent the count of observations for each combination of true and predicted labels.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classification Report – Year 3** | | | | |
|  | **Precision** | **Recall** | **F1- Score** | **Support** |
| **Not Bankrupt** | 0.99 | 0.83 | 0.90 | 1996 |
| **Bankrupt** | 0.21 | 0.86 | 0.33 | 105 |
| **Accuracy** |  |  | 0.83 | 2101 |
| **Macro Average** | 0.60 | 0.84 | 0.62 | 2101 |
| **Weighted Average** | 0.95 | 0.83 | 0.87 | 2101 |

***Description****: The classification report for Year 3, indicates a strong predictive performance for the "Not Bankrupt" category, with a precision of 0.99 and recall of 0.83.*

**A graph with a line

Description automatically generated**

***Description****: This is a ROC Curve Plot with the False Positive Rate on the x-axis and True Positive Rate on the y-axis.*

**A blue squares with white text

Description automatically generated**

***Description****: The plot above illustrates the confusion matrix for the predictions made for Year 4. On the x-axis, we have the predicted labels, and on the y-axis, we have the true labels. The categories include "Not Bankrupt" and "Bankrupt". The values in the matrix represent the count of observations for each combination of true and predicted labels.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classification Report – Year 4** | | | | |
|  | **Precision** | **Recall** | **F1- Score** | **Support** |
| **Not Bankrupt** | 0.99 | 0.89 | 0.93 | 1861 |
| **Bankrupt** | 0.37 | 0.85 | 0.52 | 98 |
| **Accuracy** |  |  | 0.88 | 1959 |
| **Macro Average** | 0.68 | 0.87 | 0.73 | 1959 |
| **Weighted Average** | 0.94 | 0.88 | 0.90 | 1959 |

***Description****: The classification report for Year 4, indicates a strong predictive performance for the "Not Bankrupt" category, with a precision of 0.99 and recall of 0.87.*

**A graph with a line

Description automatically generated**

***Description****: This is a ROC Curve Plot with the False Positive Rate on the x-axis and True Positive Rate on the y-axis.*

**A blue squares with white text

Description automatically generated**

***Description****: The plot above illustrates the confusion matrix for the predictions made for Year 5. On the x-axis, we have the predicted labels, and on the y-axis, we have the true labels. The categories include "Not Bankrupt" and "Bankrupt". The values in the matrix represent the count of observations for each combination of true and predicted labels.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classification Report – Year 5** | | | | |
|  | **Precision** | **Recall** | **F1- Score** | **Support** |
| **Not Bankrupt** | 0.99 | 0.84 | 0.91 | 1095 |
| **Bankrupt** | 0.30 | 0.90 | 0.45 | 87 |
| **Accuracy** |  |  | 0.84 | 1182 |
| **Macro Average** | 0.65 | 0.87 | 0.68 | 1182 |
| **Weighted Average** | 0.94 | 0.84 | 0.87 | 1182 |

***Description****: The classification report for Year 5, indicates a strong predictive performance for the "Not Bankrupt" category, with a precision of 0.99 and recall of 0.84.*

**A graph with a line and a red dot

Description automatically generated with medium confidence**

***Description****: This is a ROC Curve Plot with the False Positive Rate on the x-axis and True Positive Rate on the y-axis.*