**CS 559-WS Group 1 Project Report**

**Team**: Shreya Nutakki, Gaoyi Wu, Ryan Savin, Aditya Kumaran  
**Course**: CS 559-WS   
**Instructor**: David Landaeta

1. **INTODUCTION**

This report presents the results of the CS 559 final project, where machine learning models were developed to predict company bankruptcies using a dataset of 5807 companies, 198 of which were bankrupt. The dataset was clustered into five subgroups using *K*-means to capture diverse financial profiles, ranging from high bankruptcy proportions (e.g. 21.92% for Subgroup 0) to low (e.g. 0.28% for Subgroup 1). Base models and stacking models were evaluated for each subgroup, with metrics including accuracy and confusion matrices. The report summarizes subgroup characteristics, model performance, and limitations, while also addressing challenges (such as partial data for Subgroup 0) and providing insights into bankruptcy prediction and machine learning applications.

1. **METHODOLOGY**

The dataset was preprocessed using Principal Component Analysis (PCA) to reduce dimensionality (e.g. 30 features for Subgroup 4, retaining 98.62% variance). Each subgroup trained base models (e.g. Random Forest, Gradient Boosting, XGBoost) and a stacking ensemble with a meta-learner. Models were evaluated with **random\_state** set to ensure reproducibility. Performance metrics included accuracy and confusion matrices.

1. **RESULTS**
   1. **Subgroup Summarization**

The dataset was divided into 5 subgroups using *K*-means clustering, as implemented in Group1\_TrainingData.ipynb. Table 1 summarizes the number of companies, bankruptcies (y = 1), non-bankruptcies (y = 0), bankruptcy proportions, and key insights.

**Table 1: Subgroup Characteristics**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Subgroup** | **Number of Companies** | **Bankrupt**  **( y = 1 )** | **Non-Bankrupt ( y = 0 )** | **Bankruptcy Proportion** | **Key Insight** |
| 0 | 479 | 105 | 374 | 21.92 % | High financial distress, likely in volatile sectors. |
| 1 | 1444 | 4 | 1440 | 0.277 % | Highly stable, likely in low-risk industries. |
| 2 | 371 | 0 | 371 | 0 % | Constant prediction |
| 3 | 2163 | 59 | 2104 | 2.728 % | Moderate risk, large dataset size. |
| 4 | 1350 | 30 | 1320 | 2.222 % | Balanced risk, robust model performance. |

**Summary of Subgroup Characteristics**:

The subgroups exhibit diverse bankruptcy rates, ranging from 0.28% (Subgroup 1) to 21.92% (Subgroup 0), demonstrating effective *K*-means clustering based on financial health. Subgroup 0, with the highest bankruptcy rate, likely comprises high-risk companies in volatile sectors. Subgroup 1, with a minimal bankruptcy rate, represents highly stable entities, possibly in low-risk industries. Subgroups 3 and 4, with comparable rates of 2.73% and 2.22% (respectively), indicate moderate risk profiles, with Subgroup 4 showing balanced risk and robust model performance.

* 1. **Model Performance Summary**

The following table summarizes the model performance for each subgroup, including the average base model accuracy, stacking model accuracy, and number of features. The base models for Subgroup 0 were KNN, Random Forest, and XG Boost. The base models for Subgroup 1 were SVM with an RBF kernel, Random Forest, and XG Boost. Subgroup 2 was constant. The base models for Subgroup 3 were KNN, Random Forest, and SVM with an RBF kernel. The base models for Subgroup 4 were Random Forest, Gradient Boost, and XG Boost.

**Table 2: Model Performance Across Subgroups**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Subgroup ID** | **Student Name** | **Avg. Acc. Score Base [ TT (TF) ]** | **Accuracy Score Meta [ TT (TF) ]** | ***N features*** |
| 0 | Gaoyi Wu | 0.97 [ 101.33 (3.67) ] | 1.00 [ 105 (0) ] | 14 |
| 1 | Ryan Savin | 1.00 [ 4 (0) ] | 1.00 [ 4 (0) ] | 30 |
| 2 | Constant | 1.00 [ 0 (0) ] | 1.00 [ 0 (0) ] | 0 |
| 3 | Aditya Kumaran | 0.94 [ 55.66 (3.33) ] | 1.00 [ 59 (0) ] | 18 |
| 4 | Shreya Nutakki | 0.44 [30(0)] | 0.33 [ 10 (20) ] | 30 |
| Team |  | 0.88 [ 174.67 (23.33) ] | 0.90 [ 178 (20) ] | 28.6 |

**Confusion Matrices for Subgroups**

**Subgroup 0 – Gaoyi Wu**

**Stacking:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | |  |  | | --- | --- | |  | **Predicted Non-Bankrupt** | | |  |  | | --- | --- | |  | **Predicted Bankrupt** | |
| **Actual Non-Bankrupt** | 374 | 0 |
| **Actual Bankrupt** | 0 | 105 |

**XG Boost:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | |  |  | | --- | --- | |  | **Predicted Non-Bankrupt** | | |  |  | | --- | --- | |  | **Predicted Bankrupt** | |
| **Actual Non-Bankrupt** | 374 | 0 |
| **Actual Bankrupt** | 0 | 105 |

**KNN:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | |  |  | | --- | --- | |  | **Predicted Non-Bankrupt** | | |  |  | | --- | --- | |  | **Predicted Bankrupt** | |
| **Actual Non-Bankrupt** | 374 | 0 |
| **Actual Bankrupt** | 11 | 94 |

**Random Forest:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | |  |  | | --- | --- | |  | **Predicted Non-Bankrupt** | | |  |  | | --- | --- | |  | **Predicted Bankrupt** | |
| **Actual Non-Bankrupt** | 374 | 0 |
| **Actual Bankrupt** | 0 | 105 |

**Subgroup 1 – Ryan Savin**

**Stacking:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | |  |  | | --- | --- | |  | **Predicted Non-Bankrupt** | | |  |  | | --- | --- | |  | **Predicted Bankrupt** | |
| **Actual Non-Bankrupt** | 1440 | 0 |
| **Actual Bankrupt** | 0 | 4 |

**Random Forest:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | |  |  | | --- | --- | |  | **Predicted Non-Bankrupt** | | |  |  | | --- | --- | |  | **Predicted Bankrupt** | |
| **Actual Non-Bankrupt** | 1440 | 0 |
| **Actual Bankrupt** | 0 | 4 |

**XG Boost:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | |  |  | | --- | --- | |  | **Predicted Non-Bankrupt** | | |  |  | | --- | --- | |  | **Predicted Bankrupt** | |
| **Actual Non-Bankrupt** | 1440 | 0 |
| **Actual Bankrupt** | 0 | 4 |

**SVM:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | |  |  | | --- | --- | |  | **Predicted Non-Bankrupt** | | |  |  | | --- | --- | |  | **Predicted Bankrupt** | |
| **Actual Non-Bankrupt** | 1440 | 0 |
| **Actual Bankrupt** | 0 | 4 |

**Subgroup 3 – Aditya Kumaran**

**Stacking:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | |  |  | | --- | --- | |  | **Predicted Non-Bankrupt** | | |  |  | | --- | --- | |  | **Predicted Bankrupt** | |
| **Actual Non-Bankrupt** | 2104 | 0 |
| **Actual Bankrupt** | 0 | 59 |

**KNN:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | |  |  | | --- | --- | |  | **Predicted Non-Bankrupt** | | |  |  | | --- | --- | |  | **Predicted Bankrupt** | |
| **Actual Non-Bankrupt** | 1999 | 105 |
| **Actual Bankrupt** | 0 | 59 |

**Random Forest:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | |  |  | | --- | --- | |  | **Predicted Non-Bankrupt** | | |  |  | | --- | --- | |  | **Predicted Bankrupt** | |
| **Actual Non-Bankrupt** | 2104 | 0 |
| **Actual Bankrupt** | 0 | 59 |

**SVM:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | |  |  | | --- | --- | |  | **Predicted Non-Bankrupt** | | |  |  | | --- | --- | |  | **Predicted Bankrupt** | |
| **Actual Non-Bankrupt** | 1852 | 252 |
| **Actual Bankrupt** | 10 | 49 |

**Subgroup 4 –** Shreya Nutakki

**Stacking:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | |  |  | | --- | --- | |  | **Predicted Non-Bankrupt** | | |  |  | | --- | --- | |  | **Predicted Bankrupt** | |
| **Actual Non-Bankrupt** | 1311 | 9 |
| **Actual Bankrupt** | 20 | 10 |

**Gradient Boost:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | |  |  | | --- | --- | |  | **Predicted Non-Bankrupt** | | |  |  | | --- | --- | |  | **Predicted Bankrupt** | |
| **Actual Non-Bankrupt** | 1320 | 0 |
| **Actual Bankrupt** | 30 | 0 |

**XG Boost:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | |  |  | | --- | --- | |  | **Predicted Non-Bankrupt** | | |  |  | | --- | --- | |  | **Predicted Bankrupt** | |
| **Actual Non-Bankrupt** | 1320 | 0 |
| **Actual Bankrupt** | 20 | 10 |

**Random Forest:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | |  |  | | --- | --- | |  | **Predicted Non-Bankrupt** | | |  |  | | --- | --- | |  | **Predicted Bankrupt** | |
| **Actual Non-Bankrupt** | 1320 | 0 |
| **Actual Bankrupt** | 0 | 30 |

1. **CONCLUSION**

This project successfully demonstrated the application of machine learning to predict corporate bankruptcies, leveraging *K*-means clustering to uncover distinct financial profiles and stacking ensembles to enhance predictive accuracy. The collective efforts of the team produced a robust analysis, despite challenges posed by incomplete data in certain analyses. The results underscore the potential of advanced ensemble techniques in financial risk assessment and highlight the importance of tailored preprocessing strategies (such as PCA) in handling complex datasets. Moving forward, integrating additional feature engineering and advanced class-balancing methods could further improve model performance, offering valuable insights for stakeholders in financial analytics. This collaborative endeavor has strengthened our understanding of machine learning applications and their practical implications in real-world scenarios.