Jorge Acevedo A15834274

Shendo Yafuso A15528931

Albert Henderson A16384282

SVM Modeling: Bankruptcy Prediction

Introduction

Our project focuses on utilizing machine learning techniques to predict company bankruptcy using data gathered from the Taiwan Economic Journal. By using historical financial data and relevant features, our project aims to develop a predictive model that can identify companies at risk of bankruptcy. This is important for investors, creditors, and financial institutions because it makes proactive risk management easier, enhances investment decision-making, and helps prevent potential financial losses. As mentioned, our data was collected from the Taiwan Economic Journal, a reputable source known for its comprehensive coverage of financial information. The dataset spans a period of eleven years, from 1999 to 2009, providing a significant time frame to capture various economic and financial factors that influence bankruptcy. The dataset consists of a wide range of financial attributes, including liquidity ratios, profitability ratios, leverage ratios, and other relevant indicators. In total, there are 95 different features and 6819 observations. By utilizing these attributes, we can identify significant patterns that contribute to the likelihood of bankruptcy.

Based on previous studies and our understanding of the data, we hypothesize that an SVM model trained on our dataset will achieve higher prediction accuracy compared to other machine learning algorithms. Furthermore, we believe that by employing feature selection techniques to identify the most informative features for the SVM model, prediction accuracy can be enhanced while reducing computational complexity.

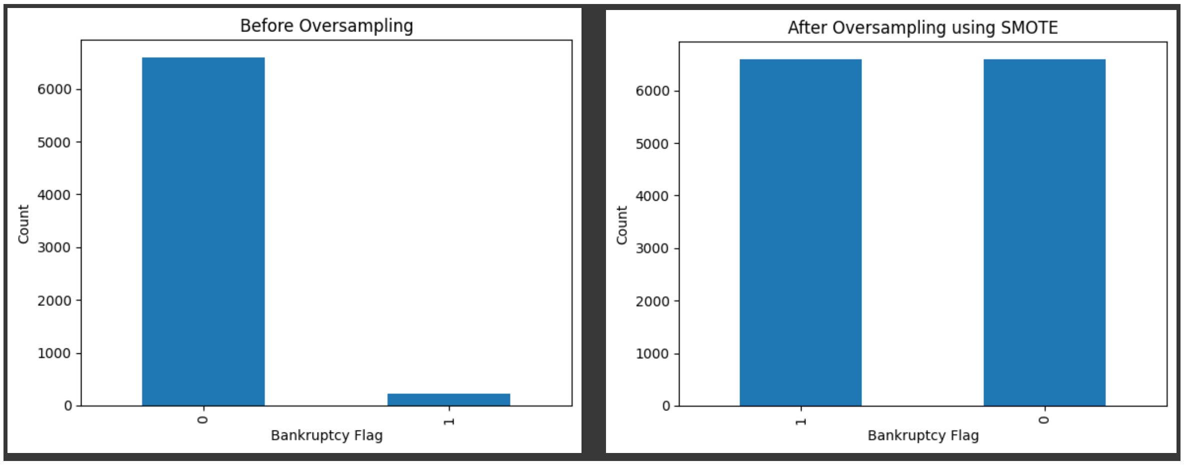
Methods

We decided on the SVM algorithm based on its ability to handle high-dimensional data and capture complex relationships between variables. The large number of features in the dataset provides a rich set of information about each company, including financial attributes, profitability ratios, leverage ratios, and other relevant indicators. These features may collectively contribute to the prediction of bankruptcy. The SVM algorithm's ability to handle high-dimensional data allows it to consider the interactions and nonlinear relationships among these features, which could be crucial for accurately predicting bankruptcy.

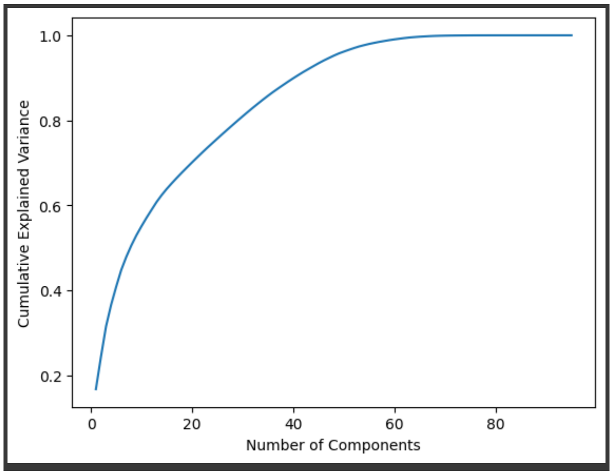
Furthermore, SVMs are known for their ability to handle imbalanced datasets, which is often the case with bankruptcy prediction where the number of bankrupt companies is significantly smaller compared to non-bankrupt companies. SVMs use a kernel function to map the data into a higher-dimensional feature space, allowing for better separation between the classes, even in imbalanced datasets.

Additionally, SVMs are less prone to overfitting, as they aim to maximize the margin of separation between classes, leading to better generalization on unseen data. This property is particularly important when dealing with prediction tasks like bankruptcy, where the model's performance on new, unseen data is crucial.

Although SVMs have the ability to perform well on datasets with high dimensionality and imbalanced observations, using Principal Component Analysis (PCA) and upsampling techniques could help improve our SVM model’s predictions. First, we applied Synthetic Minority Oversampling Technique (SMOTE) to our data to account for the large imbalance of non-bankrupt companies over bankrupt companies (6599 to 220). This increased the size of the minority class and ensured a more balanced dataset for training the machine learning model which helped to alleviate the bias introduced and can lead to better generalization of the model.



Next, we applied PCA to help overcome the curse of dimensionality. By reducing the dimensionality using PCA, we can eliminate irrelevant or redundant features, focusing only on the most informative ones. This simplification improves the SVM model's performance by reducing noise, improving generalization, and enhancing its ability to capture the underlying patterns in the data. The below graph showed us that 48 features of the original 96 were the best after PCA (n = 0.95).



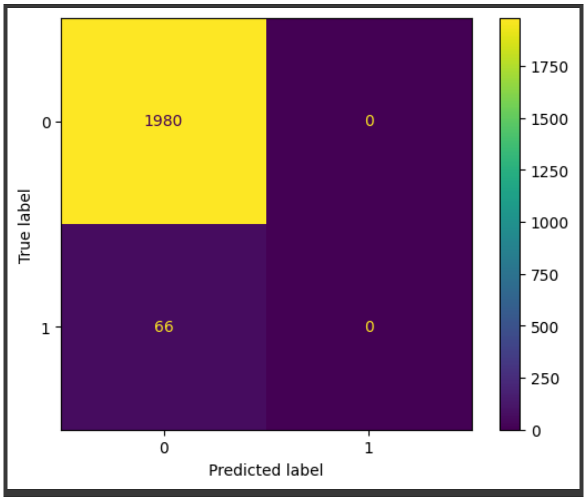
For the cross-validation aspect of our project, we made use of GridSearchCV within the sklearn library. This allowed us to exhaustively search for the best combination of hyperparameters for our SVM model. The parameters we tuned were the regularization parameter, kernel type, class weight, max iterations, and imputer strategy. To test the parameters, 5-fold cross validation was performed inside the grid search where the data was split into training and validation sets. This is repeated 5 times and each partition is used as the validation set once. Grid search cross validation methods enabled us to extract the parameters of the model that gave the best performance which we used for our final model.

Results:

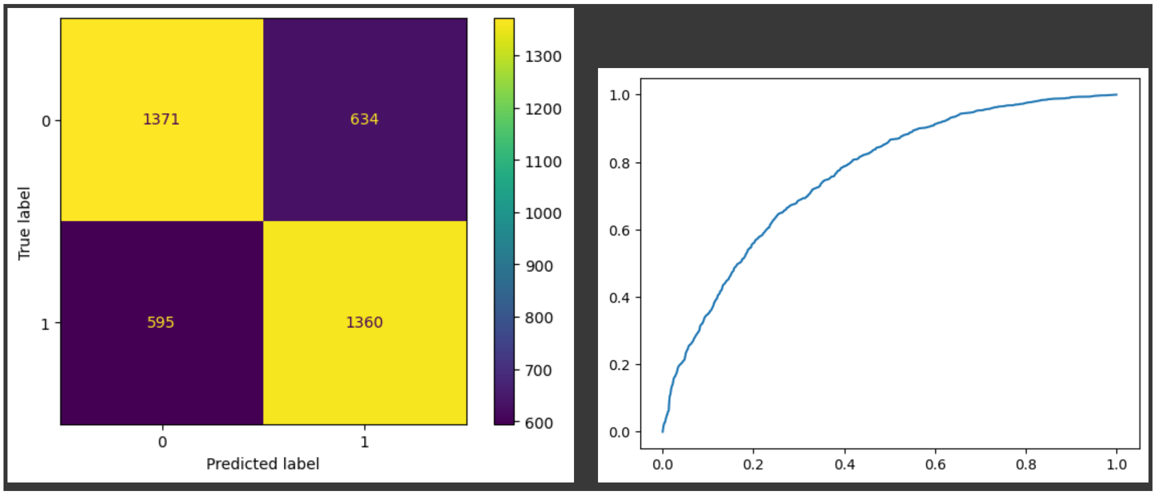
As stated previously we used SVM as our base model for this report. To compare the models' performances, we analyzed them through assessing: accuracy, precision, recall, f1 score, ROC AUC score and confusion matrices. We initially trained a SVM model in order to get a baseline of results which as we expected were not very good at creating a good model for prediction based on the scores it gave us. Next we trained a SVM model with upsampled data and performed PCA on the data as well. The main reason we decided to do this was because of the high dimensionality of the dataset we used. We believed that performing PCA and oversampling the data would yield much better results than the base model. With model two we saw a significant improvement in the model scores which also showed through the confusion matrix. Our third model was similar to the second except for this model we utilized cross validation grid search on the SVM parameters which improved our model even further in terms of prediction.

Model Selection

Our base model was trained using the default parameters of the SVC class. The results showed an accuracy of 0.9677, but the precision and recall were both 0.0, indicating that the model failed to correctly classify positive instances. While this model performed very well in terms of accuracy, this did mean that there was a high chance the model would more often than not overfit the data due to its high flexibility. You can also see in the confusion matrix that the test data did not do such a good job of analyzing true positive and false positive components.



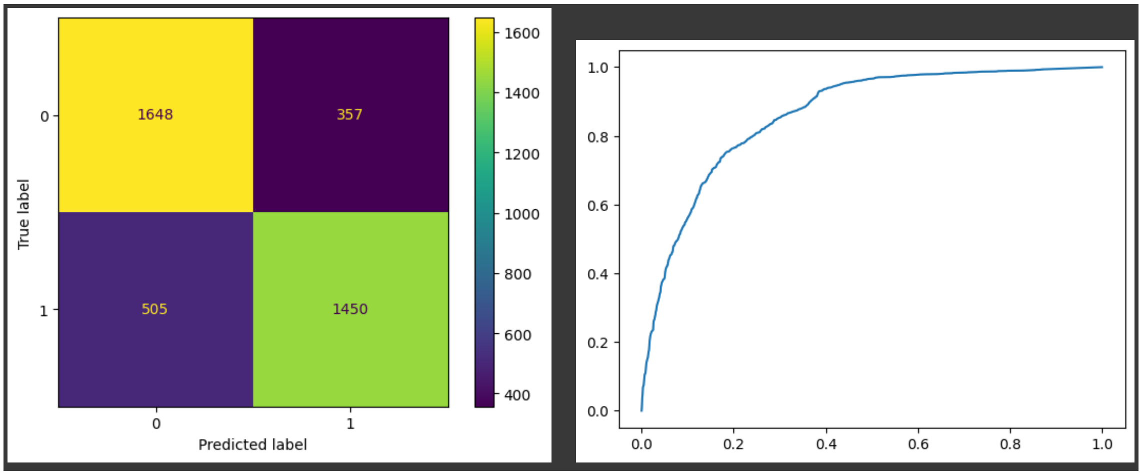
Our second model, with upsampling and PCA applied, had the following results: precision = 0.682, recall = 0.696, accuracy = 0.690, F1-score = 0.689, and ROC AUC = 0.690. Although this model showed improvements compared to the base model, there was still room for further enhancement. For example, its accuracy is much lower than that of the first model to the point where it would give less than acceptable results. This model is noticeably more complex than the previous one because of its use of PCA and upsampling. You can see in the data that the SMOTE oversampling definitely yielded more usable results that were now predicting true positive and false positive elements in our confusion matrix. Looking at the ROC graph however, you can see that there is a lot of potential for improvement on discerning these predictions.



For our third and final model we applied GridSearchCV to explore different hyperparameter combinations and found the optimal configuration for the SVM model which yielded the best results out of the three trained models with an accuracy of 0.782%. While its accuracy is still lower than that of the first model, this does help dismiss much worry of the model overfitting date. However, due to the use of grid search cross validation along with upsampling and PCA, this model has the highest complexity of the three by a noticeable margin. Through looking at the AUC ROC score however in comparison to the rest of the models this model outperformed the previous ones in all aspects and it shows through the confusion matrix as well.

Model Estimation

Our final model was the model in which we used cross validation grid search. For this final model, we found that the best parameters to be, 'imputer__strategy': 'mean', 'C_regularization': 10, 'class_weight': None, 'kernel': 'rbf', 'max_iter': -1. The evaluation metrics for this model were: precision = 0.802, recall = 0.742, accuracy = 0.782, F1-score = 0.771, and ROC AUC = 0.782.



All scores of models:

	Model 1	Model 2 (SMOTE PCA)	Model 3 (GSCV)
ROC AUC	0.5	0.690	0.782
Precision	0.0	0.682	0.802
Recall	0.0	0.700	0.741
Accuracy	0.968	0.690	782

Conclusion and Discussion

Based on the results we obtained from this experiment we concluded that our SVM does perform well in predicting if a company is bankrupt. But, we do think there is more to be desired in terms of accuracy. This desire led us to conclude that there are better models to be used instead of SVM or at least there are better methods in training SVM models than what we did in this report.

Further investigation could focus on additional preprocessing techniques such as different feature selection methods or alternative data imputation strategies. Researchers could also experiment with other algorithms beyond SVM on bankruptcy prediction tasks. Additionally, it would be valuable to analyze the misclassified instances to identify the underlying reasons for the models' limitations and potentially develop targeted solutions. Understanding the specific characteristics of companies that lead to bankruptcy and incorporating domain knowledge into the modeling process could further improve prediction accuracy.