Company Bankruptcy Prediction (Kaggle)

Ritesh Kumar

2024WI_MS_DSP_422-DL_SEC61: Practical Machine Learning

Module 5 Assignment

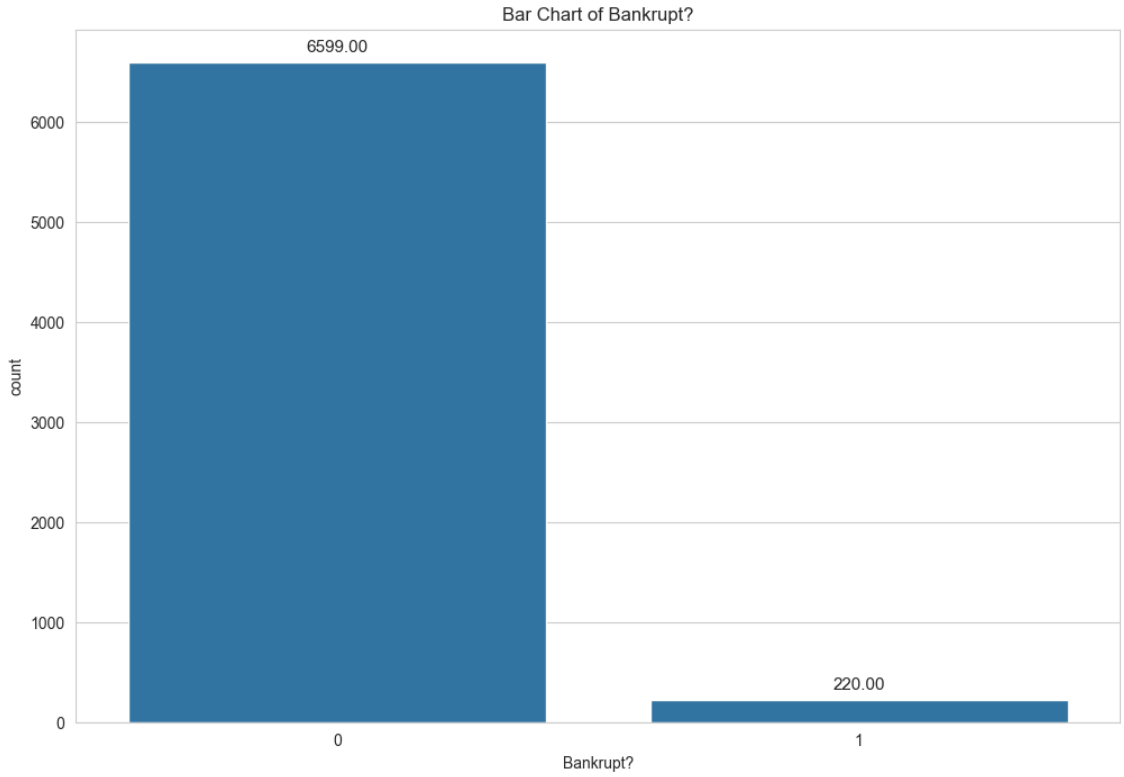
Company Bankruptcy Prediction

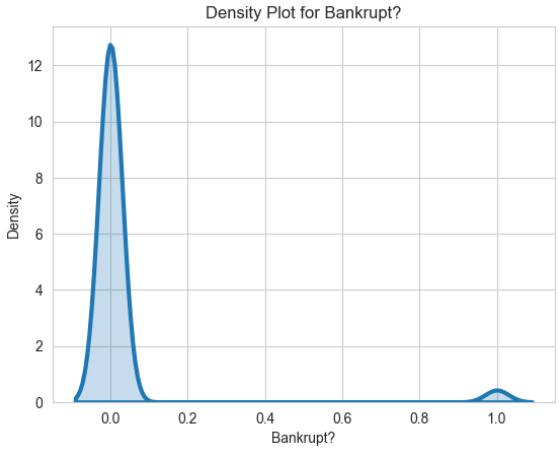
Donald Wedding and Narayana Darapaneni

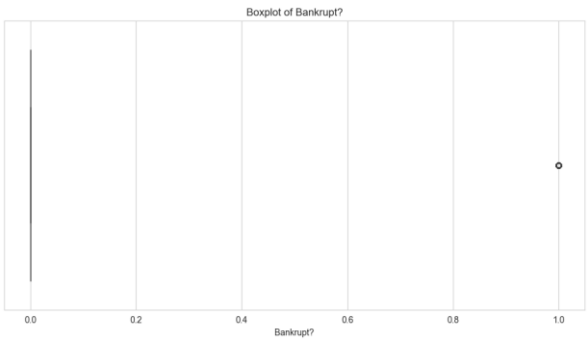
February 1, 2024

Approach

We start with a comprehensive Exploratory Data Analysis (EDA), commencing with the extraction of descriptive statistics for the 'Bankrupt?' column, serving as the target variable for prediction. We also summarize the dataset by computing the count, mean, standard deviation, minimum, quartiles, and maximum values, for all the 96 columns.







There were no missing values or duplicate rows in the dataset.

Outlier removal in the independent variables for a dataset with a highly imbalanced dependent variable like this one (where bankrupt cases are the minority) could potentially eliminate valuable information. Since bankruptcies are rare events, the characteristics that lead to bankruptcy may be present as outliers in the independent variables. These "outliers" might be critical in predicting the rare event of bankruptcy. If they were removed, the model's ability to generalize and identify the risk of bankruptcy could be significantly impaired. Therefore, we decided to not identify or remove the outliers

Next, we identified the top-20 features that have the highest correlation with Bankrupt, followed it up by plotting a heatmap depicting the strength of these 20 features amongst themselves and with ‘Bankrupt?’. To avoid multi-collinearity, we removed 20 variables that had a significantly high corelation (>0.95) amongst themselves.

We plotted the correlation heatmap of ‘Bankrupt?’ and the new top-20 highly correlated features, and boxplots and distribution-plots of these 20 features.

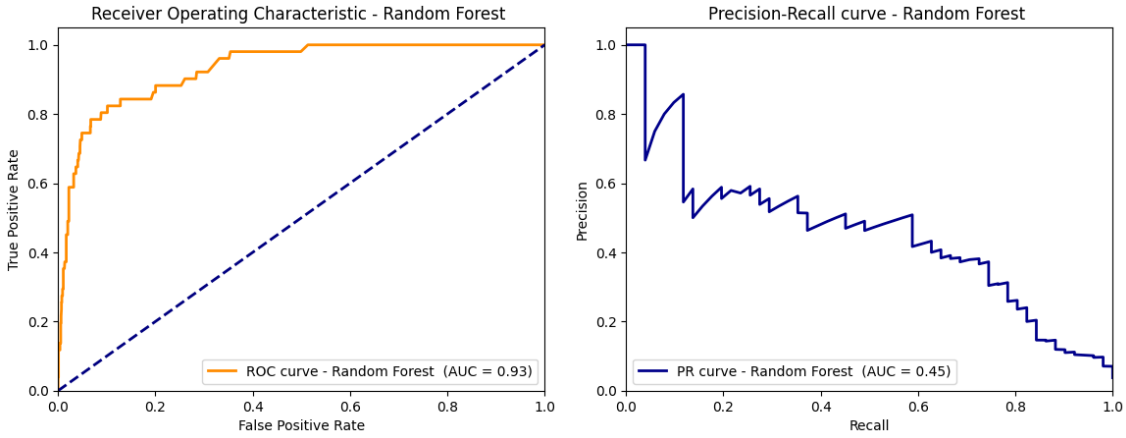
The dataset was split into train and test, in 80:20 ratio, and the data was scaled using StandardScaler.

Our dataset showcases a significant class imbalance with a vast majority of cases being non-bankrupt (6599) and a small minority being bankrupt (220). In such scenarios, logistic regression models tend to be biased towards the majority class, leading to poor classification performance on the minority class. SMOTE (Synthetic Minority Over-sampling Technique) generates synthetic samples for the minority class, helping to balance the dataset. This balance allows the logistic regression model to learn a more generalized decision boundary, improving its ability to correctly identify cases of bankruptcy, which is critical for the model's predictive performance.

By enhancing the representation of the minority class, SMOTE helps in improving the sensitivity (recall) and precision of the model, ensuring that both classes are predicted more accurately, rather than the model overwhelmingly predicting the majority class. We use SMOTE.

We start modelling by deploying the **Random Forest Classifier** and trying different hyperparameters in the model. The best model returns:

```
Best parameters: {'bootstrap': False, 'class_weight': 'balanced', 'criterion': 'entropy',
'max_depth': None, 'max_features': 'log2', 'min_samples_split': 5, 'n_estimators': 100,
'random_state': 42}
Best accuracy score (on the training dataset): 0.9839197881195613
Accuracy:  0.9618768328445748
Precision:  0.49019607843137253
Recall:  0.49019607843137253
F1 Score:  0.49019607843137253
```



The Receiver Operating Characteristic (ROC) curve for this model displays an area under the curve (AUC) of 0.93, signifying a strong discriminative ability of the model to correctly classify the positive cases. This performance is substantially better than random guessing, which would result in an AUC of 0.50, indicating that the model has a high true positive rate while maintaining a low false positive rate.

In contrast, the Precision-Recall (PR) curve has an AUC of 0.45, which is relatively low and signals that the model is not as effective when it comes to precision and recall. The precision of

the model is 0.4902, suggesting that when the model predicts a positive class, it is accurate less than half the time. This level of precision can result in a high number of false positives, which may be costly or undesirable depending on the application.

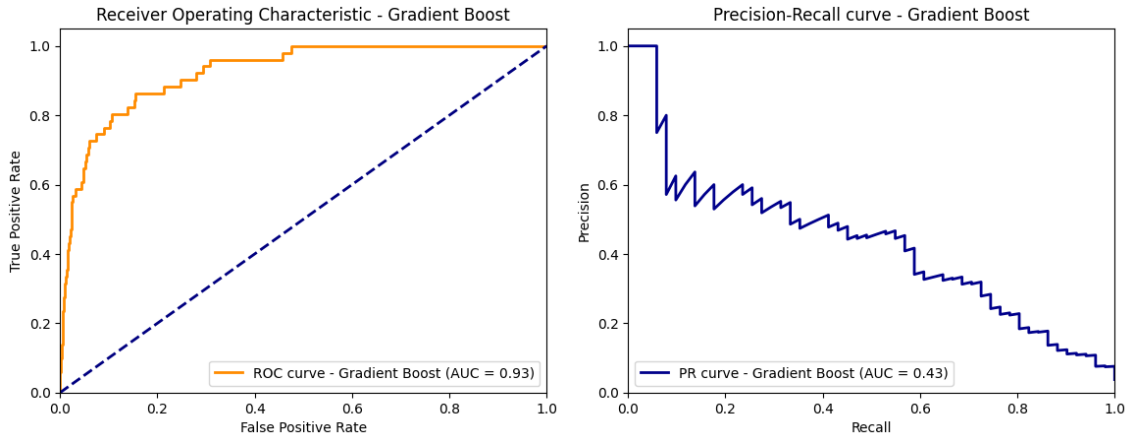
The recall value, also at 0.4902, means the model identifies 49.02% of all actual positive cases. This indicates that the model is capable of detecting nearly half of the positive instances but also misses a substantial portion, which could be critical if the positive class is of particular importance.

The model's accuracy is high at approximately 0.9619, yet this figure may be somewhat deceptive. High accuracy can occur in imbalanced datasets where one class dominates, and it does not necessarily mean the model is effective at classifying the positive class correctly.

Finally, the F1 Score, a measure that balances precision and recall, is 0.4902. This metric confirms the challenges seen in the precision and recall values and underscores the model's moderate effectiveness in classifying the positive class accurately. The identical values for precision, recall, and F1 score suggest a balance between the ability to identify positive cases and the accuracy of these identifications, but it also highlights the need for improvement to enhance the model's performance.

The **Gradient Boost Classifier** returns:

```
Best parameters: {'loss': 'exponential', 'max_depth': 10, 'max_features': 'sqrt',
'min_samples_split': 5, 'n_estimators': 250, 'random_state': 42}
Best accuracy score: 0.9863791146424518
Accuracy:  0.9596774193548387
Precision:  0.46
Recall:  0.45098039215686275
F1 Score:  0.45544554455445546
```



The Receiver Operating Characteristic (ROC) curve for this model displays an area under the curve (AUC) of 0.93, which indicates a strong ability to discriminate between the positive and negative classes, significantly better than random guessing, which would have an AUC of 0.50. This suggests that the model's discriminative ability to correctly classify the positive cases is quite good.

The Precision-Recall (PR) chart, however, shows a lower AUC of 0.43, suggesting that the model has room for improvement in terms of precision and recall. This is corroborated by the model's precision of 0.5217, indicating that when the model predicts a positive outcome, it is correct about 52.17% of the time. This level of precision may lead to a considerable number of false positives.

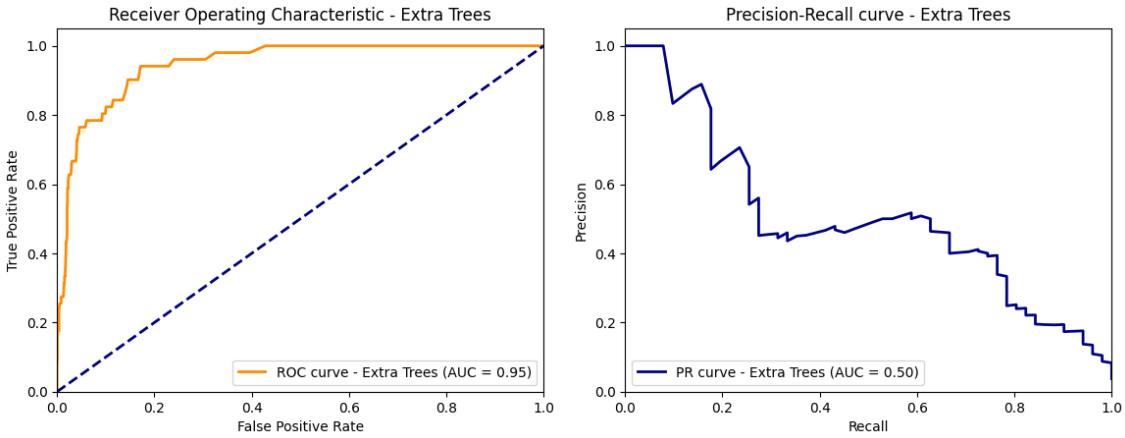
Furthermore, the recall of the model is 0.4706, which means it identifies about 47.06% of all actual positive cases. This moderate recall suggests that the model is missing a significant number of positive instances.

Despite these challenges, the model achieves an accuracy of approximately 0.964, which might be misleading as it does not capture the model's struggles with precision and recall — a common issue in datasets with class imbalance where accuracy is not the most informative metric.

Finally, the F1 Score, which is the harmonic mean of precision and recall, is at 0.4948. This score, being below 0.50, is indicative of the model's inadequate performance in precisely and reliably classifying the positive class.

The **Extra Trees Classifier** returns:

```
Best parameters: {'bootstrap': False, 'class_weight': 'balanced', 'criterion': 'gini',
'max_depth': None, 'max_features': 'sqrt', 'min_samples_split': 2, 'n_estimators': 200,
'random_state': 42}
Best accuracy score: 0.9859061672342037
Accuracy:  0.9596774193548387
Precision:  0.46
Recall:  0.45098039215686275
F1 Score:  0.45544554455445546
```



The Receiver Operating Characteristic (ROC) curve for this model shows an area under the curve (AUC) of 0.95, which is indicative of an excellent ability to distinguish between the positive and negative classes, far surpassing random guessing, which would have an AUC of 0.50. This high AUC value suggests that the model is very effective at correctly classifying the positive cases as compared to a random classifier.

The Precision-Recall (PR) chart, on the other hand, tells a different story with an AUC of 0.50, which is no better than random guessing. This low AUC on the PR curve is indicative of the

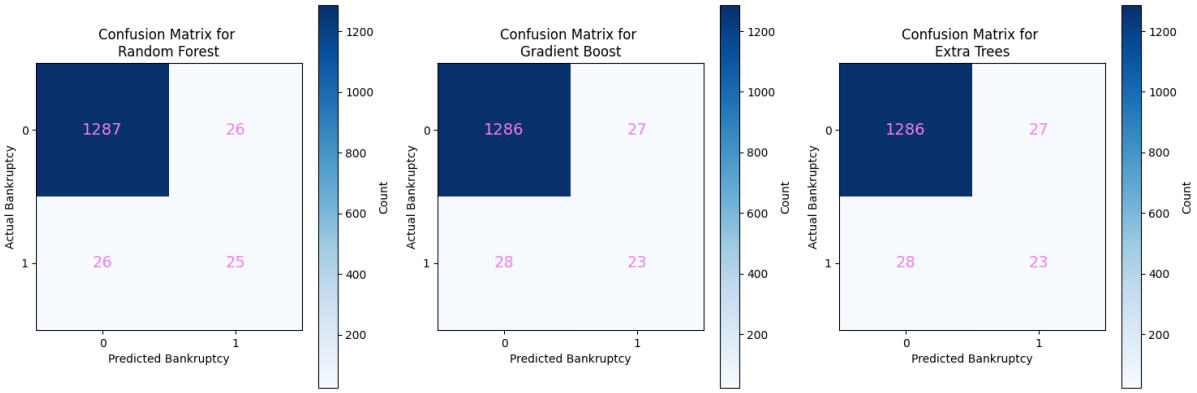
model's poor performance in terms of both precision and recall. The precision of the model is 0.46, meaning that when the model predicts a positive outcome, it is correct less than half of the time, leading to a significant number of false positives.

Additionally, the model's recall is 0.4509, indicating that it correctly identifies only 45.09% of all actual positive cases. This suggests that the model is missing a substantial number of positive instances, which is concerning for a classifier, especially in contexts where detecting true positives is crucial.

Despite these shortcomings, the model has an accuracy of approximately 0.9597, which can be misleading because it does not account for the model's low precision and recall. This is a typical scenario with imbalanced datasets where a high accuracy doesn't necessarily mean good predictive performance, particularly for the minority class.

Lastly, the F1 Score of 0.4554, which balances precision and recall, is not impressive and reflects the model's suboptimal performance in accurately and consistently classifying the positive class. This score, combined with the low precision and recall, points towards the need for further model tuning or consideration of alternative modeling approaches to improve its predictive power for the positive class.

Management Recommendations: To compare the three models, we plotted the confusion matrices:



Random Forest model exhibits a slightly better balance between false positives and false negatives, with both being equal at 26. Meanwhile, both Gradient Boost and Extra Trees models present a similar performance to each other, with 27 false positives and 28 false negatives, indicating a marginal increase in the false negatives compared to the Random Forest model.

In terms of true positives, all three models show relatively close numbers, with Random Forest at 25, and both Gradient Boost and Extra Trees at 23. This suggests that all models have a comparable ability to correctly identify bankruptcies. However, the true negatives, which represent the correct identification of non-bankruptcy cases, are highest for Random Forest at 1287, followed by both Gradient Boost and Extra Trees at 1286, which is an indication of a very slight edge for Random Forest in correctly predicting non-bankrupt cases.

These confusion matrices suggest that while all three models perform similarly, the Random Forest model has a minor advantage in terms of maintaining a balance between type I and type II errors (false positives and false negatives, respectively). This could potentially make it a more reliable choice for scenarios where it's important to maintain a balance between detecting bankruptcies and avoiding false bankruptcy alarms. However, the differences are marginal, and the choice between these models might also depend on other factors such as model

interpretability, computational efficiency, and performance on other metrics not visible in the confusion matrices.

The submission code and results can be viewed at https://www.kaggle.com/code/riteshrk/trees-rf-gb-et.

Code