Ian Brandenburg & Zsófia Rebeka Katona
https://github.com/Iandrewburg/DA3_Brandenburg/tree/main/Assignment_3

Default Prediction of Firms for 2015

This analysis develops predictive models to identify which small or medium-sized firms in the "Manufacture of computer, electronic, and optical products" industry might fail in 2015, based on their activity in 2014.

1. Data cleaning, feature and label engineering

Data preprocessing involved creating a binary target variable ("default" and "staying alive"), flag creation, winsorization of variable tails, addition of financial variables (e.g., gross profit margin, ROE, ROA, current and quick ratio, net profit margin, debt-equity ratio), removal of NaN and infinite values, and imputation of missing predictor values. The sales variable was cleaned, and a holdout sample was defined with industry code '26', sales between 1,000 and 10,000,000, and the year set to 2014.

2. EDA

We conducted an unconditional regression of log sales and squared log sales on the probability of company default. The negative coefficient for log sales (-0.0443), significant at the 1% threshold, indicates that a decrease in sales correlates with an increased likelihood of default. Given the significant skewness observed in the sales variable (Figure 1), we found that log transformation is more appropriate. Figure 2 illustrates the log-transformed sales distribution, which closely resembles a normal distribution and exhibits multimodal behavior, with peaks around log values of 11 and 12. We created box plots (Figure 3) to display the distribution of log sales values in both the training and holdout sets, grouped by default and staying alive companies. While sales are generally higher for operating companies compared to defaulted ones, the differences are not substantial. The average sales for defaulted (9.91 and 9.92) and still operating companies (11.34 and 11.43) in both datasets are similar, differing only by a few decimals. This suggests that sales figures alone may not decisively predict default status, highlighting that we need to add further variables to improve prediction accuracy. Figure 4, the lowess plot for log sales regressions and default binary, reveals consistency between the linear model and the lowess method. An increase in log sales from negative to 0 potentially indicates the company has stayed alive, suggesting a clear relationship between log sales and default likelihood. However, a small upward slope appears as log sales increase from 0, indicating that beyond a certain threshold, there are diminishing returns on reducing default risk.

3. Model building

For LASSO, we interacted CEO characteristics with financial ratios, log sales values with our defined financial ratios, and explored the impact of CEO origin, age, and gender on firm defaulting in 2015. Our initial six models (M1-M6) shared similar variable sets, but we assigned different variable sets for each predictive model to optimize predictive accuracy and generalization ability.

Linear probability models

We used logit and LASSO with cross-validation to predict probabilities. From the training set, LASSO (0.225), M5 (0.226), and M6 (0.227) had the lowest RMSE scores, while LASSO (0.777), M5 (0.77), and M2 (0.75) scored highest in AUC. Considering both RMSE and AUC, LASSO and M2 performed best. Due to the variable count, we chose M2 (12 variables) over LASSO (19), performing the calibration curve on M2.

Figure 5 shows that the highest predicted event probability was around 0.55, cutting the calibration curve relatively short. Actual event probability decreased between predicted probabilities of 0.37 to 0.45, implying that the model is overconfident in these cases. At a predicted probability of 0.45, actual event probability was zero, indicating high certainty but no actual occurrence. From 0.45, as predicted probability increased, actual event probability rose until reaching 1 at predicted probability 0.55. The plot suggests the model performs well but tends to be overly confident, especially with lower predicted event probabilities. According to Figures 6 and 7, our Model 2 ROC curve represents good predictive power. Namely, our highest FP rate is 0.268828, cutting the curve short. It consistently surpasses the 45-degree line, which indicates a better performance compared to random guessing. The curve shows rapid TPR increase at low FPR (0-0.2), indicating our model is strong at correctly identifying positive cases early on, with a relatively few incorrect identifications of negative cases. However, beyond TPR of 0.65, FPR increases rapidly, suggesting heightened sensitivity but more false negatives. This trade-off means that our model might be excellent at identifying defaulting firms, it should be used cautiously, especially when minimizing false alarms (misclassifying non-defaulting firms) is crucial.

We defined our cost function, by setting FP to 3 and the FN to 15. In Figure 8, for Model 2 Fold5, a threshold of 0.26 minimizes loss, indicating its effectiveness. Loss values peak around 0.5 for higher thresholds but are lowest between 0.2 and 0.3, indicating the best performance. The expected loss of 0.76 suggests that, on average, the cost associated with the model's predictions is 76% of the maximum possible cost. The sudden drop in expected loss reflects increased sensitivity to positive cases when the threshold is set from 0 to 0.26. Figure 10 shows the optimal threshold coordinates for M2, with Specificity at 0.98 and Sensitivity at 0.17. The AUC results for the linear models range between 67 and 80, meaning that these models have a 67 to 80% chance that the model will be able to distinguish between positive class and negative class.

RandomForest and Gradient Boosting Method

In Figure 11, our decision tree consists of one terminal node and two other nodes, despite setting a maximum depth of 3. With only two branches and a single split, the tree shows minimal complexity. The terminal node has a higher impurity, while the other nodes represent more homogeneous subsets with lower Gini indexes. In the left sample, approximately 82.9% belong to class 0 (stayed alive), while around 17.1% belong to class 1 (defaulted). In the larger sample, approximately 96% belong to class 0, indicating a high proportion of companies staying alive. We conducted a feature importance analysis and added the variables with the highest importance to the decision tree, but it remained unchanged,at our best model's (M2) feature importance values which were gross profit margin, annual profit/loss (pl) and ROA. Despite adding variables with the highest feature importance, our decision tree remained unchanged. This indicates that while certain features may have high importance in M2, their predictive power may not translate well to other types of models.

Our GridSearch determined the best model with a maximum of 6 variables and a minimum of 11 samples per node. We used these hyperparameters to train the final model on the entire training set. The feature importance analysis for the RandomForest model highlighted material expenses, dummy log sales and net profit margin as the most impactful variables. RF's AUC scores, ranging from 77% to 82%, indicate moderate to high ability in distinguishing between classes.

Figures 13 and 14 reveal that for RandomForest in Fold5, setting the threshold to 0.14 minimizes loss, making it the optimal threshold. Similarly, loss values peak around 0.5 for higher thresholds across models M2, LASSO, and RandomForest, suggesting random predictions beyond this point. The optimal ROC coordinates for RF are 0.89 (Specificity) and 0.56 (Sensitivity). In Figure 15, RandomForest demonstrates better predicted event probabilities compared to linear models, as indicated by the splines closer to the 45-degree line. The calibration plot for Gradient Boosting extends completely, possibly because the model utilizes all predictors rather than a specific number of variables. We chose to include all the variables in the GB model to capture complex relationships and help to identify potentially important variables that might have been overlooked. (Figure 16).

4. Evaluation on the holdout set

Evaluation criteria based on the training set are shown in Figure 17: CV RMSE values range from 0.229 to 0.244, with RF achieving the best value (0.223). AUC scores range between 0.71 and 0.85, with GB leading with a score of 0.854. M5 has a threshold of 0.12, while RandomForest has the highest threshold at 0.229. GB achieves the lowest expected loss at 0.621. Expected loss directly integrates the costs associated with different types of prediction errors, we can consider this as the decisive criteria. Therefore, based on the training set, GB emerges as our best-performing model.

M2: Testing M2 on our holdout set caused RMSE score to decrease to 0.208 and the AUC to slightly decrease from 0.754 to 0.752 (Figure 18). The optimal threshold also decreased from 0.169 to 0.133. Most importantly, the expected loss on the hold out set decreased from 0.76 to 0.707, suggesting that the M2 model performs better on the holdout set than on the training set. This leads to more accurate predictions and better decision making in real-world applications.
GB: The scores of our best-performing model from our training set have also changed (Figure 19). Despite GB initially having the highest RMSE value on the training set, it decreased to 0.197, lower than M2 and RF models on the holdout set. This significant drop suggests improved predictive performance. The AUC score increased by only 0.05, and the threshold decreased from 0.13 to 0.033. However, the expected loss increased from 0.621 to 0.763, making RF the most accurate predictive model on the holdout set.
RF: Observing Figure 20, we note a decrease in RMSE by 0.017, an increase in AUC by 0.03, and the optimal threshold remaining at 0.229. However, the expected loss shifted from 0.712 to 0.584. These changes suggest improved performance of RF on the holdout sample. The unchanged threshold value suggests that the model's decision boundary remains consistent between the training and the holdout sets. We can also notice that the change in the expected loss is more significant than in other scores, further justifying that the RF model is a better predictor on the holdout sample.

5. Conclusion

RandomForest's cautious approach, favoring predictions of companies staying alive, despite the higher cost of false negatives, still results in the lowest expected loss, making RandomForest the preferred choice. In practical scenarios, such a strict model is preferred when avoiding false positives and a conservative risk assessment approach is crucial.

Appendices

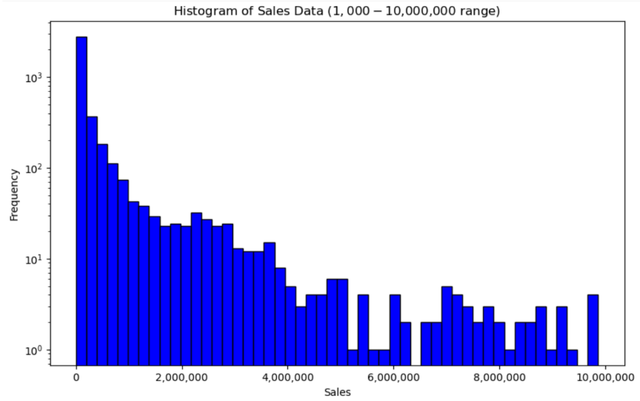


Figure 1: Histogram of sales data (from 1,000 - 10,000,000)

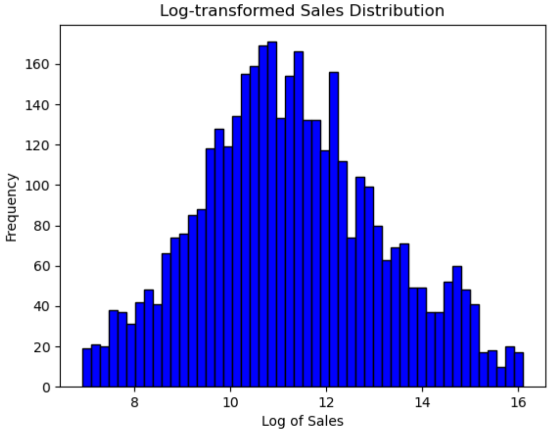


Figure 2: Histogram of logarithmic sales

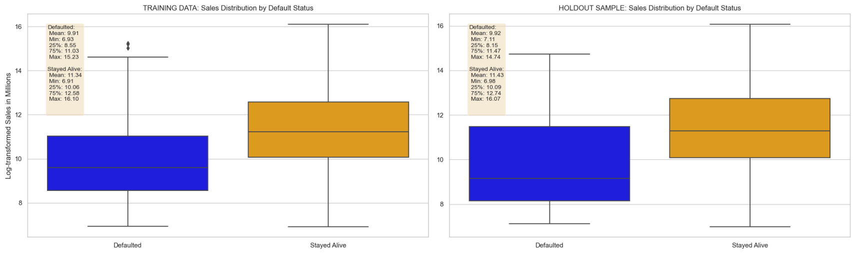


Figure 3: box plots displaying the distribution of sales for companies grouped by their default status (defaulted and still alive) in both the training and holdout datasets

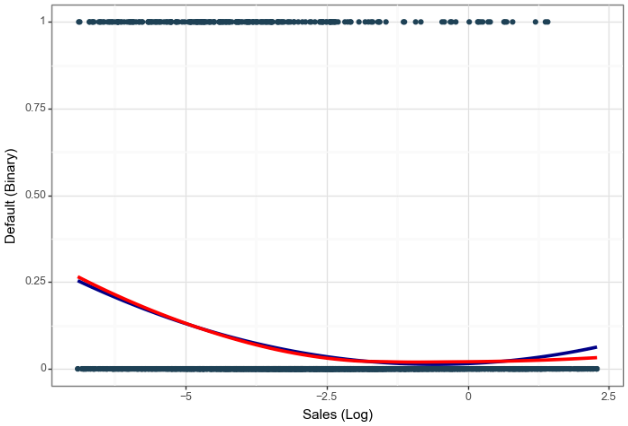


Figure 4: Lowess plot of the relationship between log sales and default status (binary: defaulted or still alive)

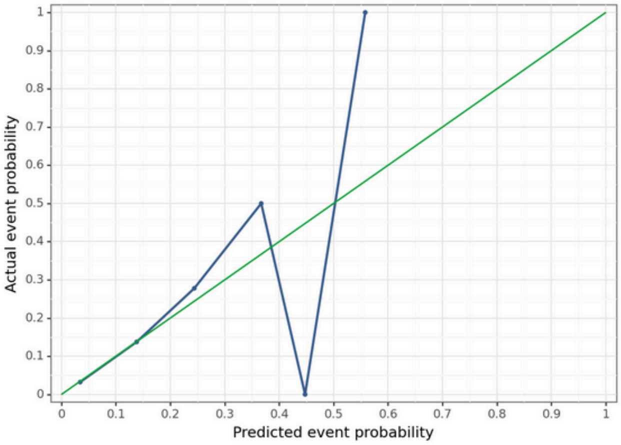


Figure 5: Calibration curve for actual and predicted event probability for Model 2

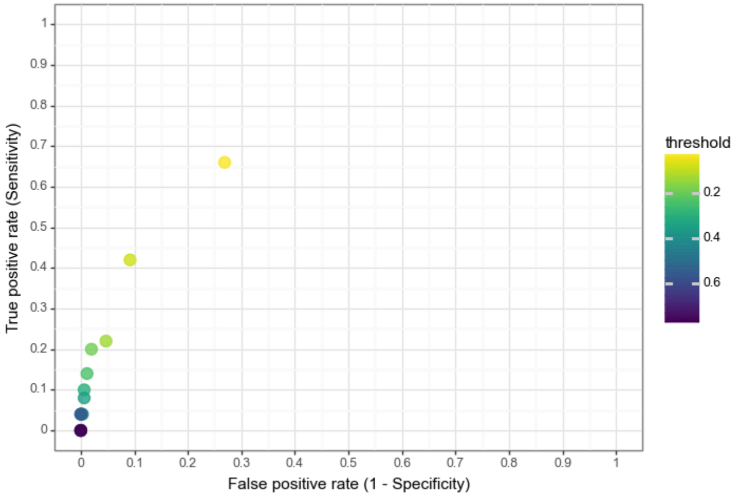


Figure 6: Receiver Operating Characteristics (ROC) curve for Model 2

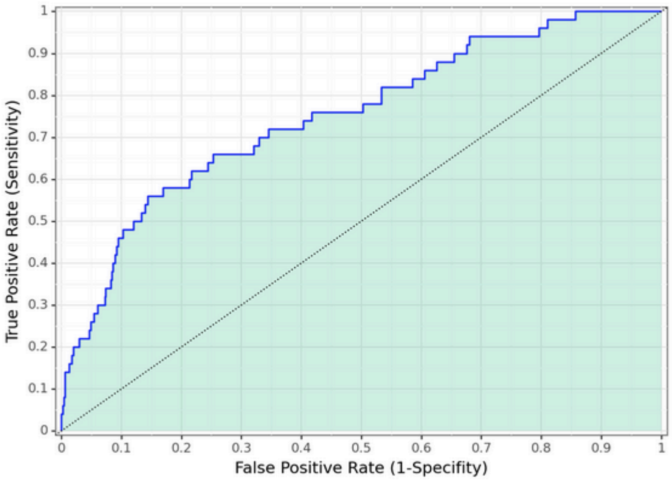


Figure 7: Continuous Receiver Operating Characteristics (ROC) curve for Model 2

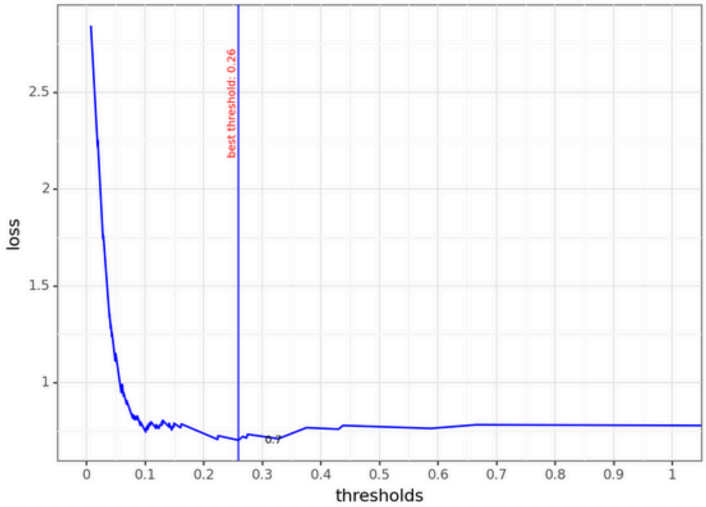


Figure 8: Expected loss vs threshold plot for Model 2

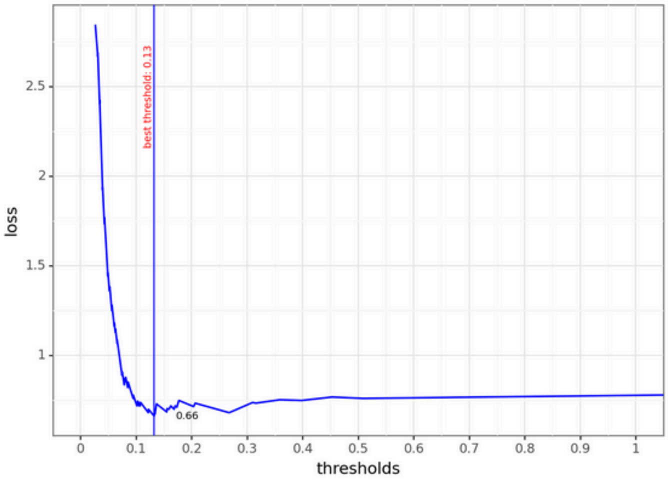


Figure 9: Expected loss vs threshold plot for LASSO model

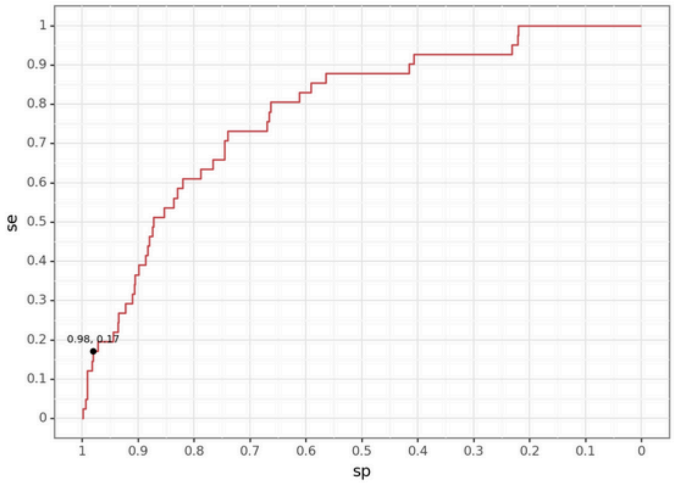


Figure 10: Optimal ROC coordinates of the best threshold for Model 2

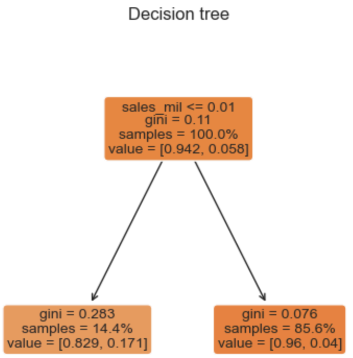


Figure 11: Decision Tree for the RandomForest model

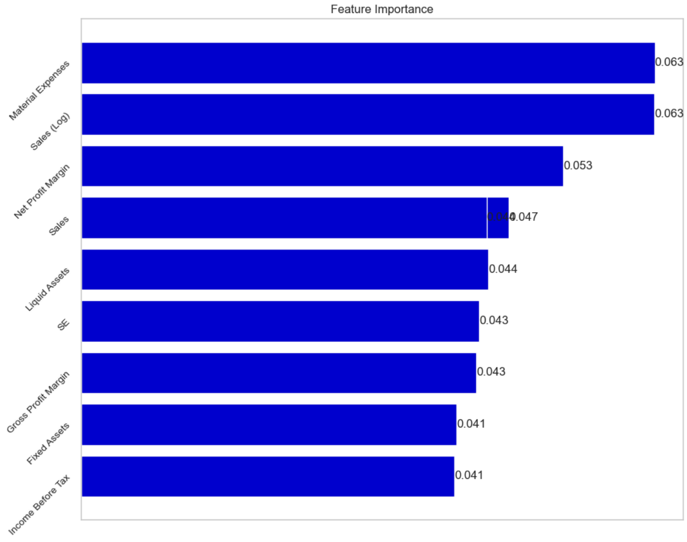


Figure 12: Feature Importance values for the RandomForest model

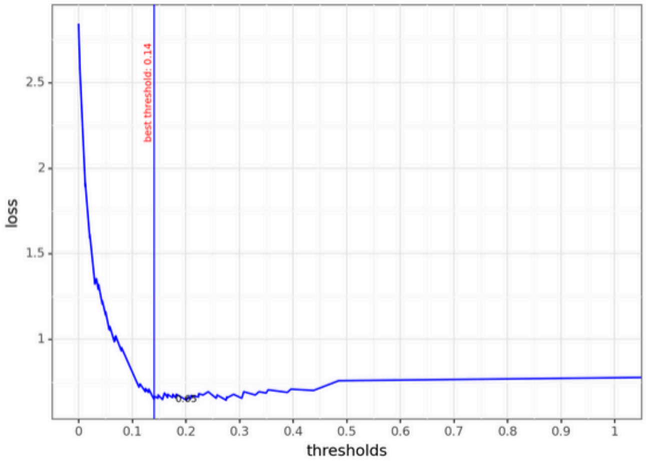


Figure 13: Expected loss vs threshold for RandomForest model

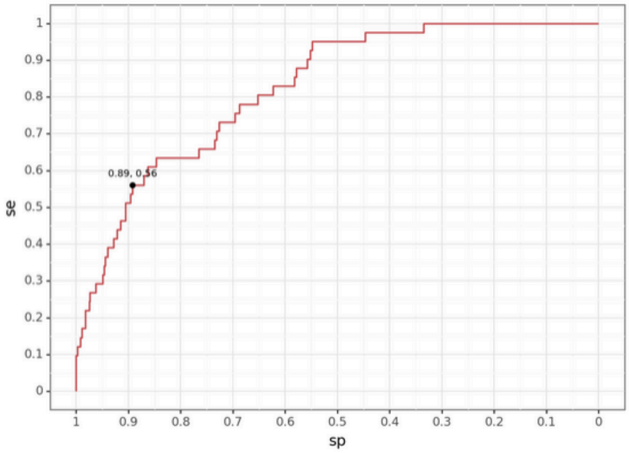


Figure 14: Optimal ROC coordinates of the best threshold for the RandomForest model

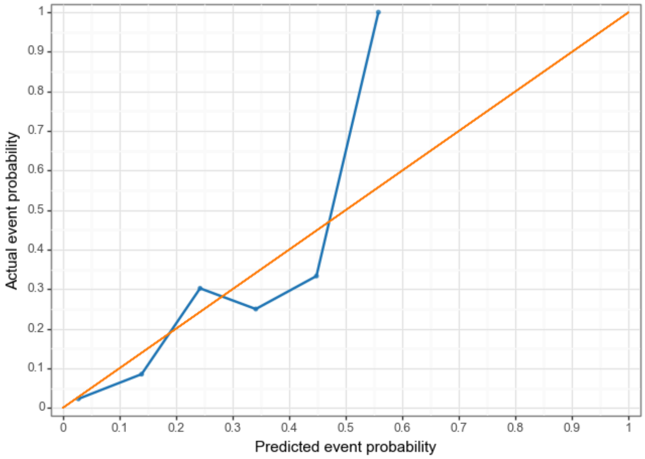


Figure 15: Calibration curve for Random forest

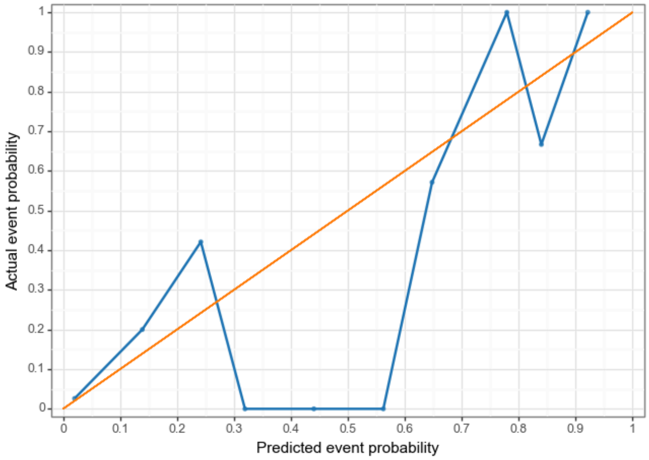


Figure 16: Calibration curve for Gradient Boosting model

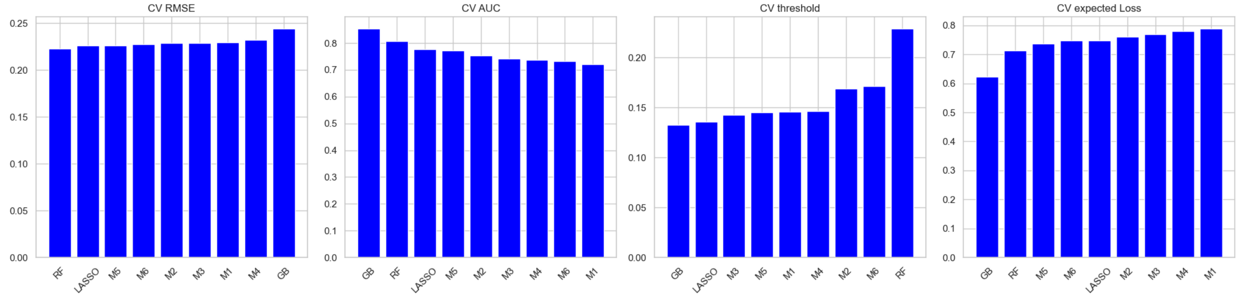


Figure 17: Summary of evaluation scores for all models for training set, including RMSE, AUC, threshold, and expected loss

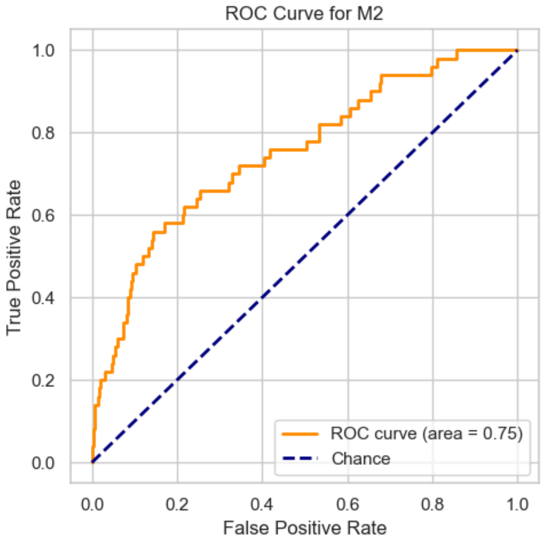


Figure 18: ROC curve of the holdout set for Model 2

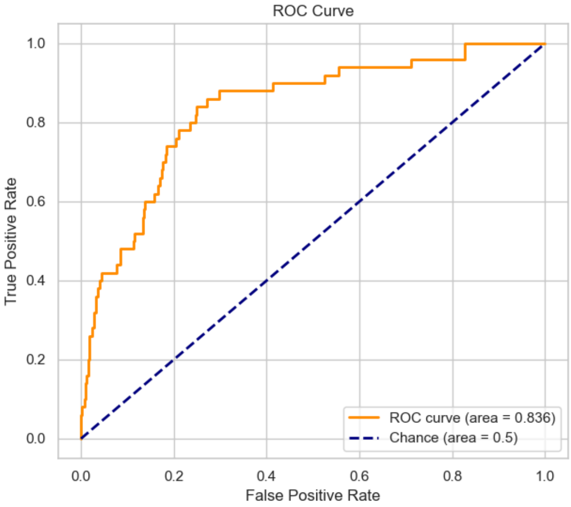


Figure 19: ROC curve of the holdout set for RandomForest

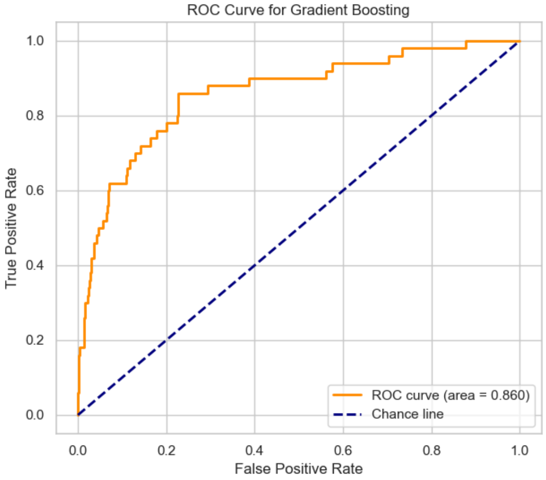
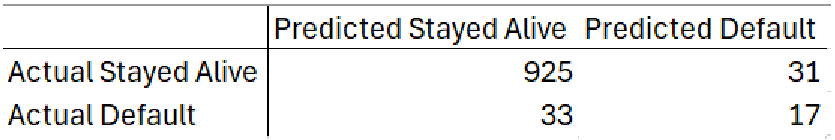


Figure 20: ROC curve of the holdout set for Gradient Boosting model



	Predicted Stayed Alive	Predicted Default
Actual Stayed Alive	925	31
Actual Default	33	17

Figure 21: Confusion Matrix for the holdout set for RandomForest