Executive Summary of Default Prediction Project

Student: Mai Pham
Github link: https://github.com/thimaipham/DA3_Assignment3

This report summarizes the development and evaluation of predictive models to identify potential defaults among SMEs in the 'Manufacture of computer, electronic and optical products' industry.

The data includes 2 parts:

- A portion is kept as a hold-out sample filtered according to the following conditions: belongs to the industry ‘Manufacture of computer, electronic and optical products’ industry, existed in 2014, but did not exist in 2015, sales in 2014 was between 1000 EUR and 10 million EUR
- The remaining Design data sample for training and testing. Filter Data for Years Before 2014: Select data from 2013 and earlier. This ensures that we do not use any information from the hold-out sample for training the model. The industry focuses on ind2 == 26, similar to what we did with the hold-out sample. Just like with the hold-out sample, we will identify SMEs based on revenue in 2013.

For data featuring and exploration parts, I will explain in more detail in the technical report.

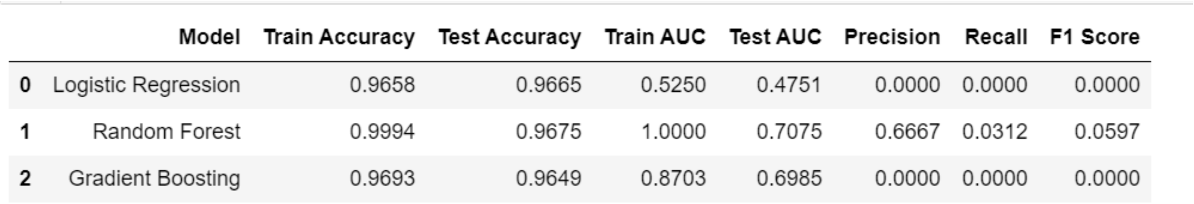
I. Model Development and Training

I have employed three different predictive models:

- **Logistic Regression:** Established as the baseline model, it demonstrated high training (96.58%) and testing accuracy (96.65%). However, the Area Under the Curve (AUC) for both training (0.525) and testing (0.4751) was close to 0.5, indicating that the model was only marginally better than random guessing at distinguishing between the classes.
- **Random Forest:** Showed nearly perfect training accuracy (99.94%) and high testing accuracy (96.75%), with an AUC of 1.0 for training, indicating excellent model fit to the training data. However, the test AUC of 0.7075, though reasonably good, suggests potential overfitting as the model performed exceptionally well on training data but less so on unseen data.
- **Gradient Boosting:** This model achieved a training accuracy of 96.93% and a testing accuracy of 96.49%, with respective AUCs of 0.8703 and 0.6985, which are indicative of a robust model with good generalization capabilities.

II. Model Evaluation and Selection

- Random Forest leads in terms of AUC on the test set and is the only model with positive Precision and F1 Score, although these values are still quite low. This indicates that the model has the best ability to distinguish between classes among the three models.
- Gradient Boosting also shows a relatively high Test AUC; however, like Logistic Regression, it fails to successfully detect default cases (with Precision, Recall, and F1 Score all being 0).
- All three models exhibit high Test Accuracy, reflecting the ease of correctly predicting the dominant class in imbalanced data. However, this metric does not fully reflect the model's performance in detecting default cases.



	Model	Train Accuracy	Test Accuracy	Train AUC	Test AUC	Precision	Recall	F1 Score
0	Logistic Regression	0.9658	0.9665	0.5250	0.4751	0.0000	0.0000	0.0000
1	Random Forest	0.9994	0.9675	1.0000	0.7075	0.6667	0.0312	0.0597
2	Gradient Boosting	0.9693	0.9649	0.8703	0.6985	0.0000	0.0000	0.0000

In this specific context, even though Random Forest is overfitting, it still appears to be the best choice based on its ability to differentiate between classes and some success in detecting default cases.

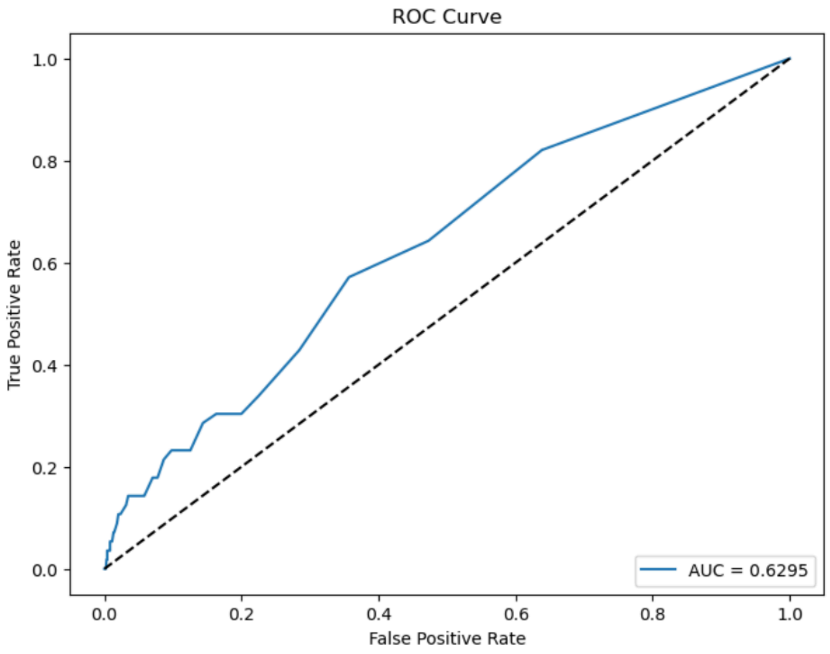
III. Hold-Out Sample Results

When applied to the hold-out sample, the Random Forest model yielded an accuracy of 94.41%, which was misleadingly high, as Precision, Recall, and F1 Score were all zero with the default threshold. This revealed the model's inability to correctly predict any true default cases. Adjusting the threshold to 0.03 resulted in improved Recall (57.14%) but with compromised Precision and increased expected loss, highlighting a trade-off between detecting defaults and maintaining accuracy for non-default predictions.

VI. Detailed Analysis and Recommendations

- **AUC:** The AUC provided by the Random Forest model on the test set (0.7075) suggests that the model has a moderate ability to distinguish between the default and non-default classes.
- **Brier Score:** The Brier score of 0.0524 on the hold-out sample is relatively low, suggesting that the model's predicted probabilities are, on average, not far from the true outcomes. However, given the imbalanced nature of the dataset, Brier score may not fully capture the model's performance in distinguishing between the two classes.
- **Accuracy (63.93%):** At the optimal threshold, the accuracy is significantly lower than the default threshold. This drop suggests that while attempting to capture more true positives (defaults), the model also increases the number of false positives, thus reducing overall accuracy.

- **Precision (8.38%):** The precision is considerably low, indicating that of all the cases the model predicts as default, only a small fraction are actually default. This low precision can be problematic, as it may lead to unnecessary actions being taken against firms incorrectly identified as likely to default.
- **Recall/Sensitivity (57.14%):** A recall rate of over 57% is a marked improvement and indicates that the model can now correctly identify more than half of the actual default cases. In the context of default prediction, a higher recall is usually more desirable than high precision, as failing to detect a default can have significant consequences.
- **Specificity (64.32%):** The specificity at the optimal threshold suggests that about 64% of non-default cases are correctly identified. However, this also means that around 36% of non-defaulting firms are being incorrectly flagged as defaulting, which could potentially lead to unwarranted credit decisions.
- **Expected Loss:** The increase in expected loss from 0.8158 to 1.3597 upon adjusting the threshold indicates a higher cost associated with misclassification, underscoring the importance of choosing an appropriate threshold that balances the cost of false positives and false negatives.
- **Area Under the Curve (AUC):** The AUC value is 0.6295, which is average, indicating that the model is able to distinguish between default and non-default classes slightly better than random prediction.
- **Classification Performance:** The ROC curve is above the dashed diagonal line, which shows that the model has better classification performance than random prediction. However, this performance is not great, because the curve is not close to the top left corner of the graph, where a perfect model would lie.



Given the current model's limitations in accurately predicting defaults, we recommend further exploration of feature selection, data rebalancing techniques, and advanced machine learning algorithms that might offer improved predictive power for imbalanced datasets.

V. Conclusion

Although the Random Forest model has shown potential, the results on the hold-out sample indicate that significant improvements are needed, particularly in correctly classifying default cases. The imbalance in the dataset and the dependence on threshold selection suggest a need for continued optimization of the model and evaluation methods.