1. **Overview**

The problem that my code is trying to solve is classifying accounting fraud cases from 25 years of data. Accounting fraud is a major issue not only to the company’s direct stakeholders, but also to the indirect stakeholders of other companies since it gives the fraudulent company unjust competitive advantage. Hence, identifying accounting fraud is crucial. Additionally, identifying accounting fraud accurately is important - incorrect suspicions may lead to unnecessarily high amounts of human effort in verifying the predictions.

1. **Methodology**

* Data Import: Data is loaded from CSV files. This dataset contains measurements and features that are possibly indicative of fraud.
* Data Cleaning: No missing values were found, data did not require much cleaning.
* Data Preprocessing: Data is standardized, correlation matrix is calculated and VIF values are computed. Undersampling in the ratio 0.2 to 1 for positive - negative class was used.
* Model Training:
  + Logistic model with all features and full data was trained, 10-fold cross validation was used
  + Logistic model with L1 penalty was trained on significant features from a simple logistic model (on the same data), using undersampled data, 70/30 train-test sample (with stratified sampling) validation was used
  + Random Forests model with all features and undersampled data, validated using the 70/30 train-test sample with stratified sampling
  + AdaBoost model with reduced VIF features and undersampled data, validated using the 70/30 train-test sample with stratified sampling
* Model Evaluation: Each model's performance is assessed using recall and precision metrics, with additional focus on AUC.
* Model Selection: The comparison revealed that the Random Forests model with all features and undersampled data had the highest precision on the test set, and comparable recall and AUC to other models. The selection is based on the balance between precision and recall, which is critical in this context.
* Conclusion: The analysis concludes with the selection of the Random Forests model trained with all features and undersampled data as the best-performing model due to the best precision score among three models while maintaining a comparable recall score. While overall the recall was comparable but not very high, the model significantly outperformed others in terms of precision.
* Trade-off considerations: For this problem, both precision and recall are important. While recall may be relatively more important given that it signifies how many of the fraud cases the model is able to capture, most models failed to achieve high recalls. This was primarily due to high data imbalance where not enough information about the fraud cases is available for the model to train on, leading to poor recall. Precision still remains important because it denotes how correct/accurate the model is in identifying the fraud cases. Hence, when it came down to a comparison between models all of which cannot capture the positive class too well, emphasis is placed on which model more accurately captures whatever portion of fraud cases it can. The final model I chose outperforms the others in this regard.
* Trade-off implications: This precision-recall tradeoff has implications in the context of the problem. On one hand, because the cost to stakeholders is high for accounting fraud cases, it is important to capture as many fraud cases as we can - which is the focus of recall. On the other hand, models predictions in problems like these most certainly need to be followed up by professionals to verify if the case is actually fraud. Inaccurately capturing positive cases hence results in increased human effort to verify the predictions, as well as increased unjust cost to the stakeholders in cases where the prediction is actually inaccurate. This calls for a simultaneous focus on the accuracy of the captured fraud cases - which is the focus of precision. Maximizing both precision and recall is usually not doable, since an increase in one generally results in a decrease in the other given how they are computed. Considering this, the final model remains the best choice among other models considered in this analysis.