Supplementary Material

Glossary:

Overfitting: it is when the model learns the trained data too well, taking into consideration the noise as well, which in the long term results in bas generalization to newer data.

Cross-validation: a statistical technique for calculating machine learning models' skill levels. To train the model and test it, it separates the data into two sets: a validation set and a training set.

Hyperparameters: a parameter that needs to be set before the learning process can start and whose value is used to regulate it.

Regularization: methods for penalizing models with excessive coefficient values in order to prevent overfitting.

Feature Importance: a number that represents each feature's importance within the prediction model.

Intermediate Results:

- Regularization strategies were required since Logistic Regression had an excessive tendency to overfit with high variance scores in its initial trials.

- Class weights were adjusted to rectify the imbalance in class prediction that existed in early Random Forest models, which greatly favored the majority class.

- Several features were eliminated in later model iterations as it was discovered that they had little bearing on prediction accuracy.

Implementation Details:

Logistic Regression:

Regularization: To prevent overfitting, L1 and L2 penalties are applied; cross-validation is used to establish the ideal strength.

Solver: Because the 'liblinear' solver handles sparse datasets and binary outcomes well, it was used for binary classification.

Class Weight: In order to improve model fairness and minority class prediction, class weights were used to address the initial data imbalance difficulties.

Random Forest:

Number of Trees: The number of trees was determined by tweaking the model to balance its variance and bias, with 20 trees being the ideal number.

Criteria: The model was able to maximize the homogeneity of nodes by using 'Gini', known as 'gdi' in Matlab, to measure the quality of splits.

Bootstrapping: Made it possible to add randomization to the model, which decreased overfitting risk and variance.

Negative Results:

Perfect training scores but low validation scores point to the overfitting of the early iterations of both models.

At first, minority class predictions proved to be challenging for logistic regression, which resulted in a large number of false negatives.

With more trees, Random Forest experienced problems with computational efficiency but saw no appreciable improvement in performance.