

Gradient boosting trees model is chosen for this project.

Gradient Boosting Trees and Random Forest are both ensemble learning techniques that have proven effective for a wide range of machine learning tasks. However, Gradient Boosting often outperforms Random Forest in terms of predictive accuracy, especially in scenarios where the data is highly non-linear and complex. This report outlines the advantages of Gradient Boosting over Random Forest, including aspects like model performance, handling overfitting, and flexibility in loss function optimization.

GBT often provides higher accuracy than RF because it optimizes on a differentiable loss function, allowing for the minimization of errors in a more granular way. This is particularly effective for complex datasets with intricate patterns.

More testing of model config is done in this week. Learning rate is tested with the outcome of 0.1 is the best learning rate.

```
Learning rate: 0.05
Accuracy score (training): 0.994
Accuracy score (validation): 0.994

Learning rate: 0.1
Accuracy score (training): 0.995
Accuracy score (validation): 0.994

Learning rate: 0.25
Accuracy score (training): 0.995
Accuracy score (validation): 0.994

Learning rate: 0.5
Accuracy score (training): 0.901
Accuracy score (validation): 0.897

Learning rate: 0.75
Accuracy score (training): 0.568
Accuracy score (validation): 0.566

Learning rate: 1
Accuracy score (training): 0.730
Accuracy score (validation): 0.733
```

Number of trees did not influence the accuracy of validation set so it is set to be 101 trees.

Max depth is set to 4 after testing:

```
Max depth: 1
Accuracy score (training): 0.993
Accuracy score (validation): 0.993

Max depth: 2
Accuracy score (training): 0.994
Accuracy score (validation): 0.993

Max depth: 4
Accuracy score (training): 0.995
Accuracy score (validation): 0.994

Max depth: 8
Accuracy score (training): 0.998
Accuracy score (validation): 0.993

Max depth: 10
Accuracy score (training): 1.000
Accuracy score (validation): 0.993

Max depth: 20
Accuracy score (training): 1.000
Accuracy score (validation): 0.993
```

Any max depth more than 4 will cause overfitting problem.

The final config will be 101 trees, LR = 0.1 and max depth to be 4:

```
Click to add a breakpoint: boostingClassifier(n_estimators=101, learning_rate = 0.1, max_features=2, max_depth = 4, random_state = 0)
```