A close-up of a note

Description automatically generated

Went to office hours, helped/got help from TAs and various students.

A math equations on a piece of paper

Description automatically generated

A close-up of a note

Description automatically generated

A close-up of a paper

Description automatically generatedA paper with mathematical equations

Description automatically generated

Q4.1:

­­­A screen shot of a computer program

Description automatically generated

4.2



4.3

1. How did you deal with categorical features and missing values?
   1. Substituted for categorical features the most frequent occurrence, based on the idea that most likely, the missing features would most likely be the most common feature. For the missing values, replaced them with median value.
   2. I completed them by using scikit pipeline coupled with simpleimputer, onehotencoder, and ColumnTransformer.
2. What was your stopping criterion?
   1. When depth was hit, it is a tunable hyperparameter.
3. How did you implement random forests?
   1. Leveraged the decision tree. The random forest implementation is basically just a decision tree in disguise.
4. Did you do anything special to speed up training?
   1. Nothing special other than changing the hyperparameters (slowly and painfully)
5. Anything else cool you implemented?
   1. I’m proud of my pre-processing, took me a very long time to figure it out and a lot of reading python libraries commands.

4.4

A black background with white text

Description automatically generated

A black background with white text

Description automatically generated

Spam Kaggle: 0.817

Titanic Kaggle: 0.822

4.5

2.

A computer screen shot of code

Description automatically generated

A screenshot of a computer

Description automatically generated

3.

A screen shot of a computer program

Description automatically generated

A graph with numbers and lines

Description automatically generated

Depth 10, 20, 27, 34, 37 gives the highest accuracy. This is broken. It should not be like this. I do not know why, I swear it was working earlier.

4.6

A computer screen shot of code

Description automatically generated

A screenshot of a computer

Description automatically generated

References:

* <https://scikit-learn.org/stable/modules/generated/sklearn.impute.SimpleImputer.html>
* <https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html>
* <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html>
* <https://en.wikipedia.org/wiki/Decision_tree>
* <https://www.geeksforgeeks.org/decision-tree/>
* <https://www.ibm.com/topics/decision-trees>
* <https://scikit-learn.org/stable/modules/tree.html>
* <https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/>
* <https://blog.paperspace.com/decision-trees/>
* <https://en.wikipedia.org/wiki/Decision_tree_learning>
* <https://www.mastersindatascience.org/learning/machine-learning-algorithms/decision-tree/>
* <https://en.wikipedia.org/wiki/Random_forest>
* <https://www.ibm.com/topics/random-forest>
* <https://towardsdatascience.com/random-forest-3a55c3aca46d>
* <https://builtin.com/data-science/random-forest-algorithm>
* <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>
* <https://careerfoundry.com/en/blog/data-analytics/what-is-random-forest/>
* <https://towardsdatascience.com/mastering-random-forests-a-comprehensive-guide-51307c129cb1>