Homework -4

CSCE 633: Machine Learning

Manav Gurumoorthy

830000011

Introduction:

In this HW, we dealt with Decision Tree Regression and Random Forest Regression. The task was to use a Decision Tree Regression model and Random Forest Regression to predict the number of shares of an article from Mashable.com.
The dataset consists of approximately 40k entries, 60 features to learn from and a final parameter shares that was the result to be computed.

Decision Tree Regression:

In this part, a decision tree regression model was constructed to estimate the number of shares of an article. In order to find a robust model we tuned the depth of the decision tree. This was done using a 5-fold cross validation.
The accuracy of the model was computed using the mean square error metric. The results are as shown below. Depth was varied from 1 to 10.

```
Average Scores for different Depths of tree:  [1016.3140978011204, 894.0392691925359, 1050.7181523765987, 943.7
Best Score: 605.2357596807856 at a tree depth of: [5]
```

Using this depth of the best performer the model was tested on the test data.

```
The final accuracy on the test set is: 2239.607809448783
```

Random Forest Regression:

In this section we look to create a random forest for different input features and for a small percentage of the train dataset. We performed 5-fold cross validation on this reduced dataset. The next step was to average the prediction across all tree to find the optimal size of the forest. Once the hyperparameters were determined i.e. the depth of the trees and number of trees in the forest, this forest was used on the test dataset. The final prediction is the average of each tree in the forest.

```
Best performer error is 1477.6864650270134 using a forest of 8 trees
```

Feature exploration:

In this part we are interested to determine what the most effective features were in the creation of the model. This was done using feature_importaces_)_ API of the decision tree regression library. This gives us a list of features and their importance. I have plotted the 10 more important features.
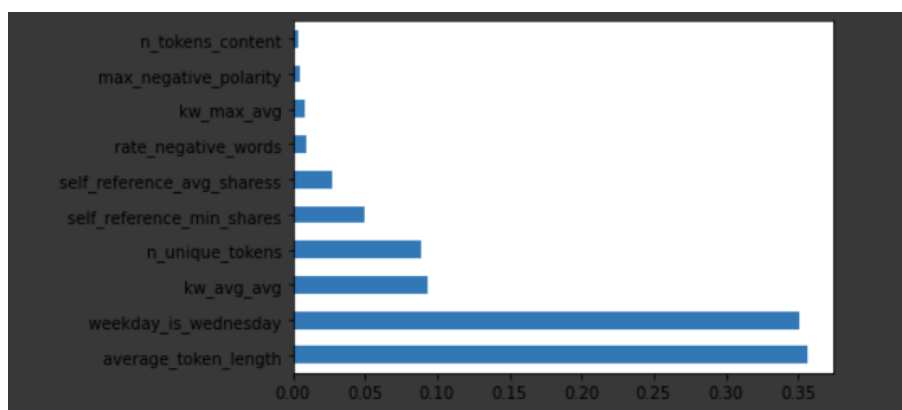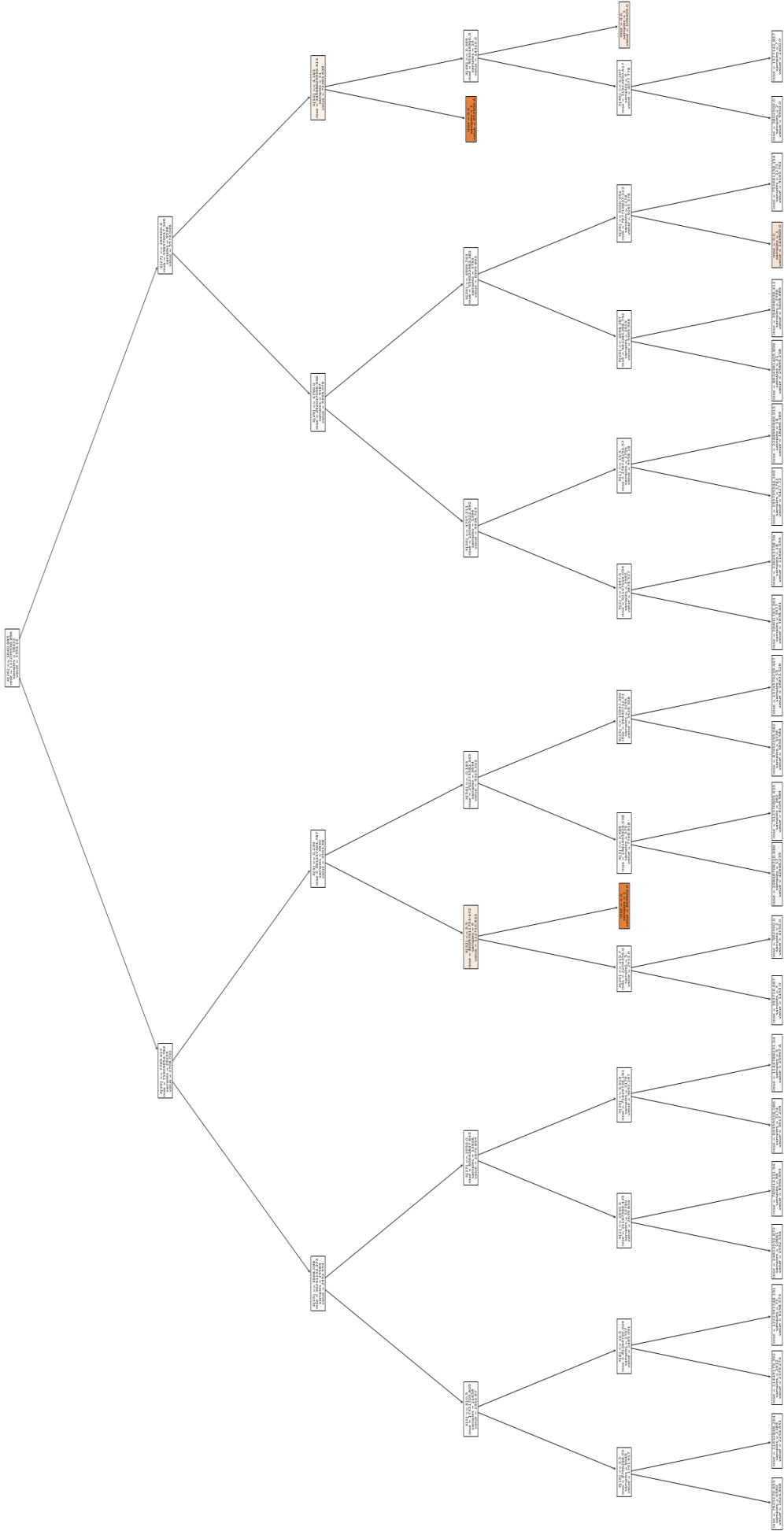
The figure above shows the final decision tree. Each parent node represents a feature of higher importance than the child. This tree is 5 layers deep.

APPENDIX A: COLAB WORKBOOK