Data Analytics: Assignment 1 Report

2018111032
Vivek Pamnani

1. Objective

Objective is to correctly predict an earthquake given the four features YEAR, LATITUDE, LONGITUDE, and DEPTH for a given *threshold* **T** in the range **[4, 5]** (inclusive).

To enable the use of classifiers, we must first clean the data and either remove or fill invalid data points. In this assignment, I chose to remove any invalid data points.

Resulting data points were 40,107 out of the original 52,898 data points.

2. Cleaning the data

To clean the data, I first dropped all the features that were either redundant, unnecessary, or incomprehensible.

- Origin Time was not in a parse-able format and was dropped.
- Magnitude values in other than Mw were simply different scales and were dropped.
- Intensity values were largely invalid (NaN) and were dropped.
- Location values were only valid for roughly 20% of the dataset and was dropped.
- Reference would not be a useful feature in predicting earthquake and was dropped.

Values for the features LATITUDE and LONGITUDE were a mix of float values and strings (represented as <float>^{o}N). They were parsed to get the float values (by removing ^{o}N, etc.).

Values for MAGNITUDE, DEPTH, and YEAR were parsed as float, float, and int respectively.

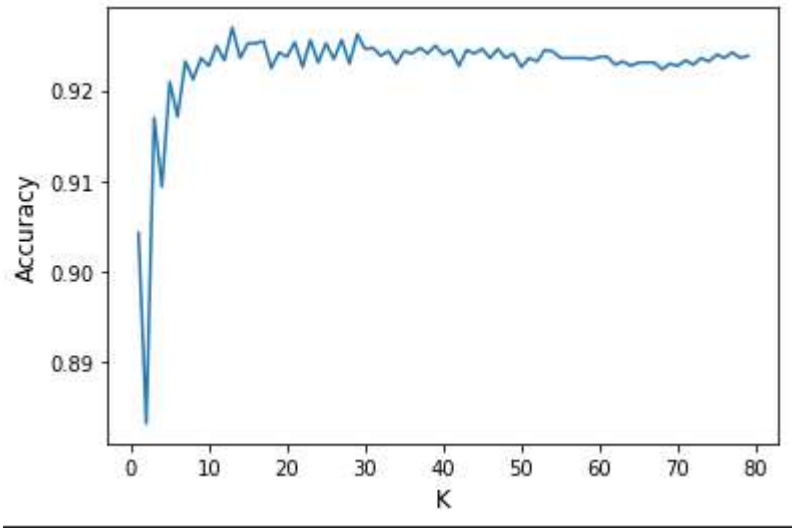
Any entries with invalid values for any single chosen feature were dropped.

3. Analysis of implemented classifiers

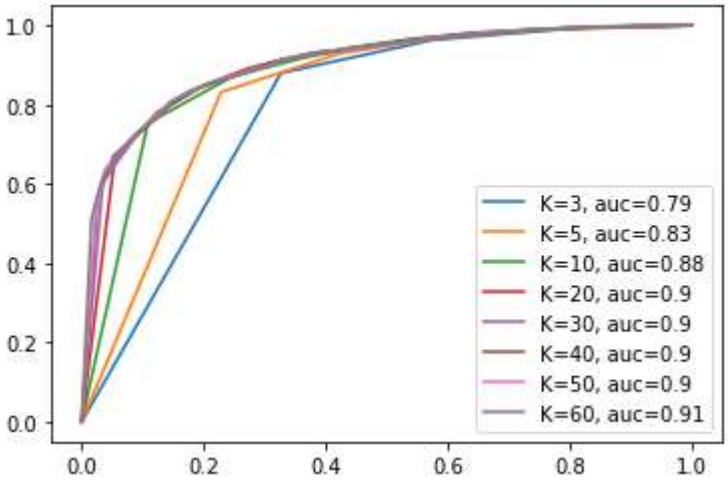
NOTE: Task specific questions have been answered in the python notebook under the “Analysis” section.

KNN Classifier:

The accuracy of the classifier for differed for K values as follows:



The ROC plots for the KNN classifier for various values of K were as follows:

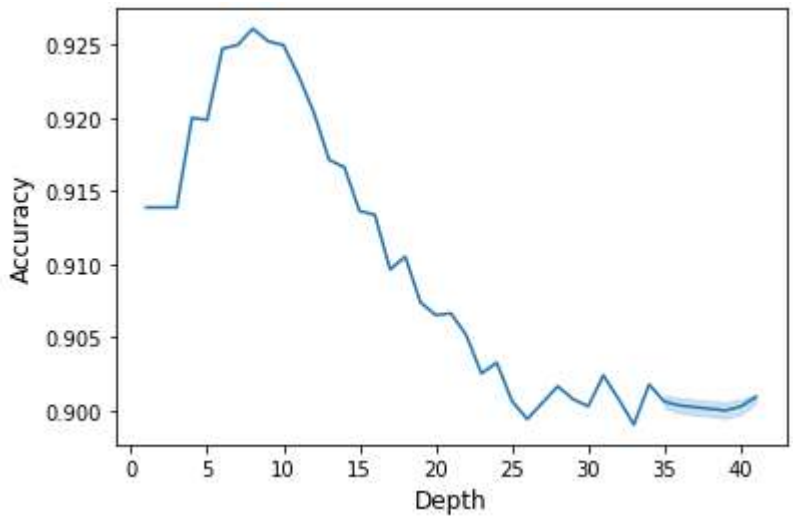


From here we can infer that K = 30 is the optimal parameter. It covers the maximum area of the ROC plot with AUC score as 0.9.

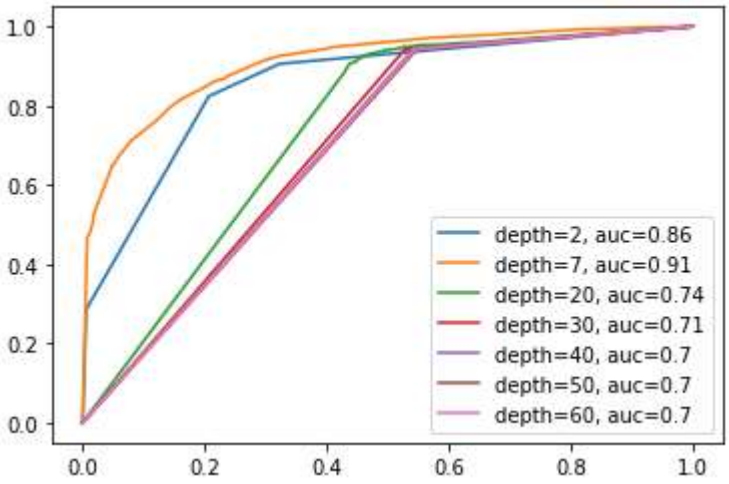
Accuracy and F1-scores for K=30 were optimal at 92.2% and 0.96.

Decision Tree Classifier

The accuracy of the classifier for differed for K values as follows:



The ROC plots for the Decision Tree classifier for various pre-pruned depths were as follows:

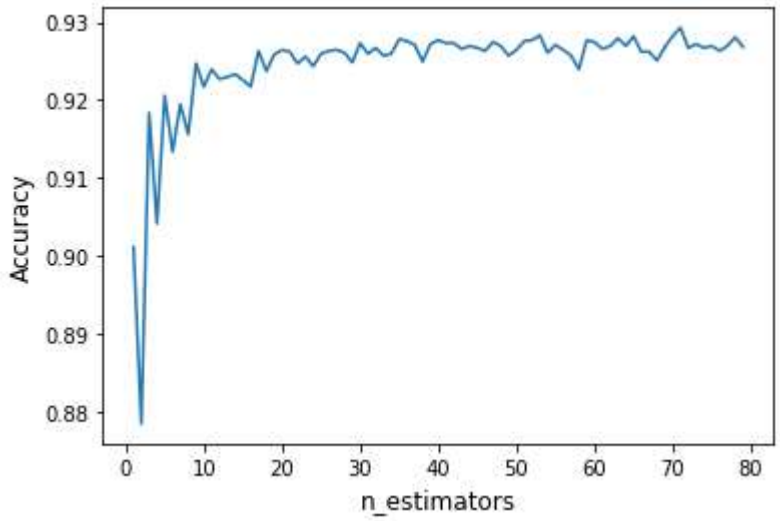


From here we can infer that max_depth = 7 is the optimal parameter. It covers the maximum area of the ROC plot with AUC score as 0.91.

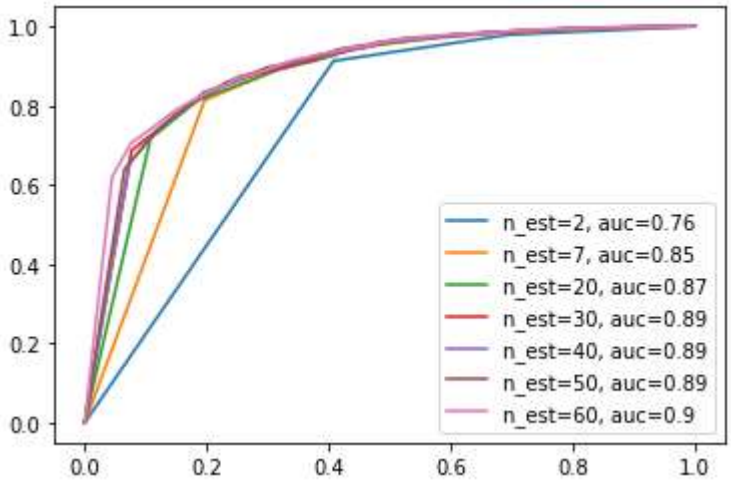
Accuracy and F1-score for max_depth=7 were optimal at 92.3% and 0.96.

Ensemble Learning: Random Forest Classifier

The accuracy of the classifier differed for number of estimators as follows:



The ROC plots for the Random Forest classifier for various number of estimators were as follows:



From here we can infer that n_est=50 is the optimal parameter. It covers near maximum area of ROC plot with AUC = 0.89.

Accuracy and F1-score for n_est=50 were optimal at 92.7% and 0.96.

4. Inferences

- **A.** Threshold **T = 4** gave the best predictions from all three models.
- **B.** Out of the three classifiers (at T = 4), **Random Forest Classifier** was the most accurate with 92.7% accuracy. F1-scores and AUC were similar for all three classifiers.
- **C.** For T = 5, optimal values were:
    - i. KNN : The optimal K seems to be 10 with Accuracy = 87.3%, F-score = 0.711 and AUC =0.80.
    - ii. Decision Tree : The optimal depth seems to be 7 with Accuracy = 87.5%, F-score = 0.693 and AUC = 0.0.78.

iii. Random Forest : The optimal n_est seems to be 50 with Accuracy = 86.9%, F-score =0.732 and AUC = 0.82.

D. For T = 4, optimal values were:

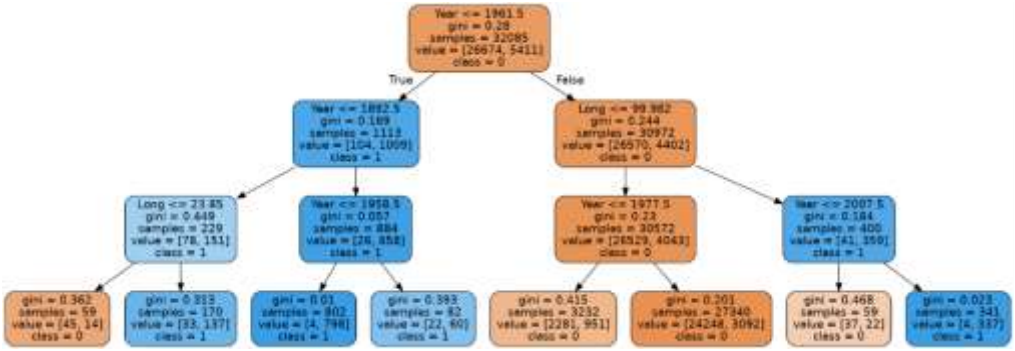
i. KNN : The optimal K seems to be 30 with Accuracy = 92.2%, F-score = 0.96 and AUC =0.91.

ii. Decision Tree : The optimal depth seems to be 7 with Accuracy = 92.3%, F-score = 0.96 and AUC = 0.91.

iii. Random Forest : The optimal n_est seems to be 50 with Accuracy = 92.7%, F-score = 0.96 and AUC = 0.90.

E. For T values between 4 and 5, the results were not optimal. This can be verified by changing the threshold (T) under the “Adjusting Threshold T” section of the notebook and running the respective code blocks for each of the classifiers.

F. Decision Tree revealed that the primary classifying factors were YEAR and LONGITUDE out of the four selected features.



G. Normalizing the features and performing Principal Components analysis to reduce the dataset to two dimensions slightly improved accuracy at T = 5.

END OF REPORT