ML Model Documentation

Team Kepler

June 11, 2022

1 Data

1.1 Data Collection

Dataset was provided in the problem statement and we used more training data from this link.

1.2 Pre-Processing

- Annuled data from the following columns : *kepid, tce_plnt_num, tce_rogue_flag, tce_insol, tce_impact, tce_insol_err, tce_period_err, tce_time0bk_err, tce_impact_err, tce_duration_err, tce_depth_err, tce_prad_err, tce_eqt_err, tce_eqt_err, tce_steff_err, tce_slogg_err, tce_sradius_err* as they were either unique attributes, NaN values or not useful for classification.
- Deleted rows having *UNK* (unknown) label.
- Dropped rows having any NaN values.

Parameters used for training : *tce_period, tce_time0bk, tce_duration, tce_depth, tce_model_snr, tce_prad, tce_eqt, tce_steff, tce_slogg, tce_sradius.*

2 Model Architecture

2.1 Random Forest Classifier

Random forest classifier uses many decision trees in order to train the model and predict on test data. We have used 13000 decision trees for estimation. The general architecture of a random forest classifier is as shown in Figure 1.

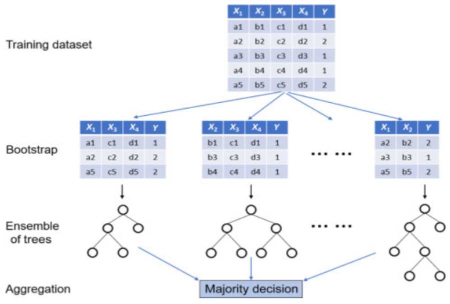


Figure 1: Architecture of Random Forest Classifier

2.2 Neural Networks

Here, we initially normalized the data using Standard Scalar function. We used Multi Layer Perceptron Classifier(MLP) with 3 hidden layers each consisting of 64 neurons whose basic architecture is as shown in Figure 2.

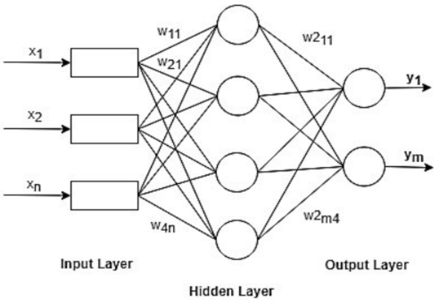


Figure 2: Architecture of MLP Classifier

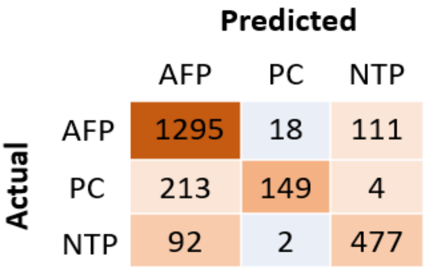
3 Evaluation Metrics :

3.1 Random Forest Classifier

3.1.1 Precision, Recall, F1-Score

Precision	0.82
Recall	0.38
F1 - Score	0.52

3.1.2 Confusion Matrix



		Predicted		
		AFP	PC	NTP
Actual	AFP	1295	18	111
	PC	213	149	4
	NTP	92	2	477

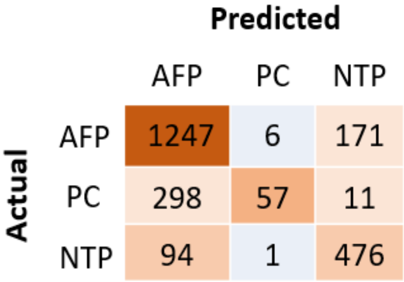
Figure 3: Confusion Matrix of RFC

3.2 Neural Networks

3.2.1 Precision, Recall, F1-Score

Precision	0.34
Recall	0.57
F1 - Score	0.43

3.2.2 Confusion Matrix



		Predicted		
		AFP	PC	NTP
Actual	AFP	1247	6	171
	PC	298	57	11
	NTP	94	1	476

Figure 4: Confusion Matrix of MLP Classifier

4 Limitations and Possible Enhancements

- The number of features given to us were less. If given more number of parameters, we could modify the app accordingly.
- The Random Forest Classifier Model is large (3.2GB) due to the 13000 decision trees. Given more time, we could find a way to incorporate it in our app.
- Currently, we have dropped all columns containing error related information and all rows which have NaN values in any of the columns. In future, we could derive some inferences from these columns, rows and use it for model training.

5 Team Members

- `Anand Hegde`
- `Arvind Kumar M`
- `Harriththa B`
- `Kavali Sri Vyshnavi Devi`
- `Shashank P`