Instalocate Machine Learning Internship

Flight Delay Analysis

Feature Selection: I removed all the Null values from the dataset first. Also the flight status and flight number are not useful in determining if a flight gets delayed or not. Also it brings sparsity to the dataset and makes it imbalanced. Hence it is removed. Python's Datetime format is convenient to work with dates and times and we use it to convert the dates in this format to extract all the information such as day, month, year from scheduled arrival and departure times. We do not use expected or actual times since they will not be available while prediction a new entry. Since the dataset did not have a separate validation or test set, I split the dataset as 80% train, 15% test, 15% validation.

Approach: Since the project required to build a simple model, I decided to implement a simple baseline along with a 3 layer deep neural I decided to use Tensorflow's newly available estimators. Estimators are pre-build computation graphs. The simple baseline classifier returns the probability distribution of the classes as seen in the labels. For multi-label problems, this will predict the fraction of examples that are positive for each class. I decided to use a 1D target i.e a simple True or False for our baseline model (since we only want to observe how the baseline performs) whereas the target variable for DNN is 2D. This is because we want confidence score for each prediction and not just a binary classification (‘delayed’ or ‘not delayed’). Using the softmax(categorical crossentropy) as a probability distribution.

Improvements: The model would definitely improve with more variables since flight delays depend upon a number of factors. It is also possible to use deep learning models such as Recurrent Neural Networks, LSTM, Deep Belief Networks(DBF) to study air traffic data since temporal information about time is available.