Flight Delay Prediction problem.

//Introduction

For this assignment, we were tasked with predicting flight delays based on a dataset of flights from 2017. Our training set has 4,911 examples, which is not a huge amount of data, but certainly is sufficient for most methods. We also don’t have a huge number of features; our training dataset has 21 columns. Of those 21, several are redundant (e.g., Airline ID, Unique Carrier), and one is the label. Still, before selecting algorithms to test, we should do some data pre-processing and dimensionality reduction. We will then evaluate three algorithms to find the best performing one for our dataset. We used Tableau for initial dataset evaluation, and python with pandas, sci-kit learn and matplotlib for all subsequent steps, from data pre-processing to result visualization.

As we have discussed, some features of our dataset are clearly redundant, and some, such as flight number, we wish to discard to prevent overfitting. After removing such fields, we are left with the following features:

|  |  |
| --- | --- |
| Feature Name | Type |
| Day of Week | Continuous |
| Departure Time | Continuous |
| Taxi Out | Continuous |
| Taxi In | Continuous |
| Actual Elapsed Time | Continuous |
| Distance Group | Continuous |
| Day of Year | Continuous |
| Carrier | Categorical |
| Origin City Market | Categorical |
| Destination City Market | Categorical |
| Origin Airport | Categorical |
| Destination Airport | Categorical |
| First Departure Time | Continuous / sparse |

To get to this list, we removed:

UID – removed to prevent overfit

Airline Id – redundant to Carrier field

Flight Number – removed to prevent overfit

Origin City Name – redundant to Origin Airport

Origin State – redundant to Origin Airport

Destination City Name – redundant to Destination Airport

Destination State – redundant to Destination Airport

Distance – redundant to Distance Group

Before we begin evaluating our features, we decided to add one additional feature: Flight Month. We believe that some pattern of flight delays may be seasonal, due to weather or high load related delays, but this may be more clear when looking at the month of travel rather than the specific day. We would normally at this point develop some meta features, such as flights per day of week, or flights per month, to examine the relationship between load and flight delay, but it appears that when this dataset was sampled from the database of flight delay data, the sample evened out the number of examples per day and month, so we do not expect such features would be meaningful.

To select which features we should use with our algorithm, we first evaluated our categorical variables for Information Gain

|  |  |  |
| --- | --- | --- |
| Field | Entropy | Information Gain |
| Carrier |  |  |
| Origin City Market |  |  |
| Destination City Market |  |  |
| Origin Airport |  |  |
| Destination Airport |  |  |
| First Departure Time |  |  |

//Algorithm selection

//Algorithm evaluation

Of the algorithms we studied in class, Linear Regression, K-Nearest Neighbor, Decision Trees, and Neural Networks can be used for regression.

//Conclusions