Nick Palacio

Raj Dasgupta

Intro to AI

Project Proposal

# Predicting Flight Arrival Delays

## Problem Statement

The logistics of predicting airline flight delays can be incredibly complex. It can depend on many different variables: airline company, time of year, departure time, etc. This is a worthwhile problem to solve for airlines because if a flight will be arriving late it could have a chain reaction on other flights that are using the same airplane or crew which can be costly. This is a worthwhile problem to solve for passengers so that if a flight is going to be late they can plan accordingly. The sooner both of these stakeholders know about a delayed flight the better.

## Existing Techniques

Several machine learning methods have been employed to solve this very problem. A paper published in 2016 (found here <https://ieeexplore.ieee.org/abstract/document/7777956>) talks about using decision trees, random forests, and the AdaBoost algorithm to try and predict flight delays. However, they turned it into a binary classification problem to simply say if a flight would be late or not. Another paper also published in 2016 (found here <https://ieeexplore.ieee.org/abstract/document/7778092>) used Deep Learning to try and predict flight delays.

## Your Approach

My approach to this problem will involve comparing several different machine learning algorithms on a dataset of flight and weather information from Microsoft’s Azure Machine Learning Studio. The algorithm’s I will compare are **Decision Trees**, **Regression** and **Support Vector Machines**. I will also compare the model performance with and without an additional weather dataset that ties an airport to weather conditions at a point in time.

I will be using several tools/libraries for this project:

* **IPython** notebooks for sharing
* **Matplotlib** Python library for rendering charts and graphs of the data
* **SciKit-Learn** Python library for the implementation of the machine learning algorithms

I will evaluate the algorithms based on their accuracy, recall and precision.

## Bi-weekly Milestones

* Tuesday, November 6th
  + Python code (IPython)
    - Load dataset
    - Clean data
    - Analyze a few interesting features with charts
* Tuesday, November 20th
  + Python code (IPython)
    - Run 3 algorithms on a clean dataset without weather data
    - Produce some visual element to compare the 3 algorithms
* Tuesday, December 4th
  + Python code (IPython)
    - Run 3 algorithms on clean dataset with weather data
    - Produce some visual element to compare the 3 algorithms both with and without the weather data