**Capstone Project Definition – Airline Delays**

**Problem Definition**

Every year approximately 20% of airline flights are delayed or cancelled, resulting in significant costs to both travelers and airlines. As a Capstone project, I’d like to build a supervised learning model that predicts airline delay from historical flight.

**Impact to Client**

Typically, the problem defined above will impact an airline industry in terms of significant costs to both travelers and airlines. If the airlines are able to predict what factors cause the airline delays, they will be able to implement control mechanisms to check those factors and improve efficiencies.

**Dataset**

I will be exploring the airline delay dataset that is available on Bureau of Transportation Statistics. This dataset includes details about flights in the US from the years 1987-2008. For project purposes, I will be using a sample of 2-3 years of data. Every row in the dataset includes 29 variables listed below:

**Variable descriptions**

|  |  |  |
| --- | --- | --- |
|  | **Name** | **Description** |
| 1 | Year | 1987-2008 |
| 2 | Month | 1-12 |
| 3 | DayofMonth | 1-31 |
| 4 | DayOfWeek | 1 (Monday) - 7 (Sunday) |
| 5 | DepTime | actual departure time (local, hhmm) |
| 6 | CRSDepTime | scheduled departure time (local, hhmm) |
| 7 | ArrTime | actual arrival time (local, hhmm) |
| 8 | CRSArrTime | scheduled arrival time (local, hhmm) |
| 9 | UniqueCarrier | [unique carrier code](http://stat-computing.org/dataexpo/2009/supplemental-data.html) |
| 10 | FlightNum | flight number |
| 11 | TailNum | plane tail number |
| 12 | ActualElapsedTime | in minutes |
| 13 | CRSElapsedTime | in minutes |
| 14 | AirTime | in minutes |
| 15 | ArrDelay | arrival delay, in minutes |
| 16 | DepDelay | departure delay, in minutes |
| 17 | Origin | origin [IATA airport code](http://stat-computing.org/dataexpo/2009/supplemental-data.html) |
| 18 | Dest | destination [IATA airport code](http://stat-computing.org/dataexpo/2009/supplemental-data.html) |
| 19 | Distance | in miles |
| 20 | TaxiIn | taxi in time, in minutes |
| 21 | TaxiOut | taxi out time in minutes |
| 22 | Cancelled | was the flight cancelled? |
| 23 | CancellationCode | reason for cancellation (A = carrier, B = weather, C = NAS, D = security) |
| 24 | Diverted | 1 = yes, 0 = no |
| 25 | CarrierDelay | in minutes |
| 26 | WeatherDelay | in minutes |
| 27 | NASDelay | in minutes |
| 28 | SecurityDelay | in minutes |
| 29 | LateAircraftDelay | in minutes |

**Solution Approach**

1. Explore raw data to determine various properties of features and how predictive these features might be for the task on hand.
2. Using Python, prepare the feature matrix from the raw data. Then, with each iteration, improve the feature set, resulting in better overall predictive performance.
3. Using Python’s Scikit-learn, build various models, such as Logistic Regression or Random Forest.
4. Using Scikit-learn, evaluate performance of the models and compare between iterations.

**Deliverables**

1. Python Code that will implement solution.
2. Project Documentation
3. Summary Powerpoint Presentation