**Machine Learning Engineer Nanodegree**

**Capstone Proposal**

Puneet Mathur  
March 12th, 2017

**Proposal**

*(approx. 2-3 pages)*

**Domain Background**

*(approx. 1-2 paragraphs)*

The proposal is to create a FlightDelay Python application which predicts flight delays for airlines and passengers. Flights delays globally contribute to huge losses not just to the passengers but also to the national economies. United States of America which has the highest number of commercial flights has conducted a detailed study on flight delays and impact on its GDP in a study by NEXTOR[1]. Flight delays happen due to various reasons and this project will try to unearth those reasons which cause flight delays. The purpose being to build a model for predicting the next flight delay.

**Problem Statement**

*(approx. 1 paragraph)*

In its published work by NEXTOR[1] it has become evident that in the year 2007 alone the cost of all flight delays to the commercial airlines was $8.3 billion. Cost to passengers was even higher at $16.7 billion dollars. Estimates on cost to GDP was $4 billion overall. This study is relevant to our project as it is the one sponsored by the Government of United States and was carried out by consortium of Universities at its behest.

The problem we are trying to address is the Flight Delays and the problem is with respect to United States of America based on data from DoT[1]. The problem statement is the following:

“Why do Flights get delayed? Can we predict which flights are prone to delays and their reasons?”

Flight delays are fairly common nowadays and any passenger on a commercial airline would like to know if their flight is predicted to be on-time or not. To address this problem statement machine learning would be used along with GUI to create an application which would predict whether a flight is going to be on time or not.

**Datasets and Inputs**

*(approx. 2-3 paragraphs)*

The dataset that I am going to use in this project is from the US Department of Transportation website[2] on flight delays. <https://www.transtats.bts.gov/Tables.asp?DB_ID=110&DB_Name=Air%20Carrier%20Statistics%20%28Form%2041%20Traffic%29-%20%20U.S.%20Carriers&DB_Short_Name=Air%20Carriers>

All inputs in form of flight time performance datasets have been taken from US DoT[1] website. The years that have been chose for this project include from the year 2000 to 2016. The data however is available only till Nov 2016

**Solution Statement**

*(approx. 1 paragraph)*

The solution would involve creating a python GUI based application which would take in appropriate inputs based on the variables identified through Unsupervised learning and then fed into a suitable Supervised learning model. The unsupervised learning model would help us unearth those key parameters which lead to flight delays. The solution would classify and classify a flight as “Delayed” or “Not Delayed” meaning it is on time and there are no expected delays.

**Benchmark Model**

*(approximately 1-2 paragraphs)*

There is no model as far as I am aware which predicts flight delays currently. However when we are building our model I can use Naïve Bayes classifier as an initial score to benchmark the prediction that our final model gives. Naïve Bayes classifier assumes independence between variables and initially we could check if there is any correlation between the variables to see if they are independent or not.

**Evaluation Metrics**

*(approx. 1-2 paragraphs)*

The model being built should be able to predict with an accuracy score of more than 70% flight delays on the production dataset or live dataset fed from online data on flight schedules.

**Project Design**

*(approx. 1 page)*

In this final section, summarize a theoretical workflow for approaching a solution given the problem. Provide thorough discussion for what strategies you may consider employing, what analysis of the data might be required before being used, or which algorithms will be considered for your implementation. The workflow and discussion that you provide should align with the qualities of the previous sections. Additionally, you are encouraged to include small visualizations, pseudocode, or diagrams to aid in describing the project design, but it is not required. The discussion should clearly outline your intended workflow of the capstone project.

Preliminary investigation:

Workflow for prediction of a flight delay:

Model building:

Visualizations on Wireframes(GUI):

**Before submitting your proposal, ask yourself. . .**

* Does the proposal you have written follow a well-organized structure similar to that of the project template?
* Is each section (particularly **Solution Statement** and **Project Design**) written in a clear, concise and specific fashion? Are there any ambiguous terms or phrases that need clarification?
* Would the intended audience of your project be able to understand your proposal?
* Have you properly proofread your proposal to assure there are minimal grammatical and spelling mistakes?
* Are all the resources used for this project correctly cited and referenced?