Title: Predicting Airline Delay

First name	Last Name	Email	Monday or Tuesday class
Ayush	Goyal	agoyal18@hawk.iit.edu	Tuesday (Online)
Puneet	Ojha	pojha@hawk.iit.edu	Tuesday (Online)

Important Notes:

- Each group must submit ONLY one copy by a single team member!
- Your submission will get a score 0-10 which means your maximal grade in your final project. For example, if your submission (i.e., your team) got a score of 7, it implies that your team can get a maximal grade of 70 on your final project.
- You will have a chance to resubmit the proposal if you are not satisfied about your grade. But you must resubmit it by Nov 2.

Topics of Your Projects: Multiple linear regression

1. **Introduction**

   The aim of this project is to perform Multiple linear regression analysis on the data of airline delay. Multiple linear regression is considered as one of the perfect technique in data analytics, and for our future prediction we are driven towards multiple linear regression. Analysis on flight delay can be done based on the features such as Origin Airport, Destination Airport, weather, days (like – holidays, weekdays), distance, and elapsed time. With the given data and statistical technique, we will predict the future outcomes.

2. **Data Sets**

   Our data set is regarding the airline delay with the following attributes which are listed below. We have obtained are data from United States department of Transportation. This data contains various information about the flights in United States. The data in the associated file “AirlineDelay.csv”. The major cause of delay can be predicted by the following attributes – carrier_delay, weather_delay, national_air_system_delay, security_delay, and late_aircraft_delay. the data set like div_distance is set to ‘0’ if the flight is diverted from the actual rout.

   Data set variables are as follows.

   Year

   Month

   Day_Month

   Day_Week

Unique_Carrier
Origin_Airport_ID
Dest_Airport_ID
CRS_Dep_Time
CRS_Arr_Time
Arr_Delay
Distance
CRS_Elapsed_Time

3. Research Problems

We have chosen airline delay because many airlines make fake promises to its customers that their airline is always on time. However, it's not true because as our data set shows there are major delays in flight due to which the customer suffers a lot. So, our research problem is to predict the airline delay and the factors which are causing that delay. It is very important to learn about the factors which are causing the delay because based on learning, we can predict the solution to overcome the delay problem. For the example, if weather is one of the factor which cause the airline delay then from the previous data it can be predicted that when the weather is not good there are higher chances that the flight will be delayed.
So, by applying statistical programming on the dataset we will find the major factors which will cause the delay in airline.

4. Potential Solutions

To solve the delay problem, first we will identify the attributes which are affecting the time schedule of the airline using the R. Then based on the result of R2 and Adj-R2 and p-value the comparison will be done. Additionally, linear regression is applied on the dataset with backward elimination method such that we can obtain the best model by making time our dependent variable and rest factors as independent variables. Moreover, we will also find the co-relation and dependency between the variables such that we can measure our dataset precisely.

5. Evaluations

A solution can be reached by many numbers of methods. Choosing of that method is very important as it leads us to the right solution. We will be using Hold-out evaluation for the analysis, as our data set is very large. We will divide the data in to two different sets such as training data set and test set. The training data set will be 80% and test data set will be 20% of the entire data set. Building of the model will take place from the training data and evaluation of the model will be done by the test data. RMSE, Confidence interval, AIC, BIC will be taken in consideration for the evaluation.

6. Expected Outcomes

There are many expected outcomes but few which are more important are as follows:-

- Best fit model of the data set.
- Relationship between the different variables which are responsible for the delay in the flight.
- Some pattern which can be interpreted from the data set for the minimum or the maximum delay of the flight (such as during holidays or week days).
- Prediction of the delay time of the flight in minutes.
- Days or a time period when there will be minimum delay or no delay to the flights.