Project Proposal
Info - 7390 Advanced Data Science and architecture

Predicting Airlines Delay

Presented by:

Dhanisha Damodar Phadate
001859234
phadate.d@husky.neu.edu

Palak Sharma
001828562
sharma.pala@husky.neu.edu

Dharani Thirumalaisamy
001887922
thirumalaisamy.d@husky.neu.edu

OVERVIEW :

Transportation is one important factor for a Country's growth. Air being one of the important modes of transport , the Airlines are responsible for any problem regarding that.

on an average , 200 flights are being delayed everyday around the world which affects the passengers in some or the other way. In order to improve the service provided by the airlines, they are investing a lot on delay prediction now-a-days.

GOALS :

The goal of this project is to :

1. Tell the airlines the parameters that influence the delay in flights.

2. Predict the airline delays

3. Classify the delays based on the external conditions such as time , weather.

DATA :

1. You can download the dataset from here :

https://raw.githubusercontent.com/hortonworks/data-tutorials/master/tutorials/hdp/predicting-airline-delays-using-sparkr/assets/airline_data.zip

We will be working on datasets which has details about flight delays in USA in 2015 in R-Studio.

PROCESS OUTLINE :

1. Data Ingestion from the above link.

2. Upload the data to HDFS

3. Create a SparkContext object that connects the program to cluster.

4. Exploratory Data Analysis in R

5. Feature Engineering on the dataset

6. Feature Selection on the dataset

7. Creating the testing and training datasets.

8. Use supervised methods to perform predictive analysis.

9. Use supervised methods to classify the delay based on weather.

TIMELINE :

TIME FRAME	PROCESS
Day 1-4	Data Preprocessing
Day 3-6	Model Building
Day 7-10	Deployment of models

FLOW DIAGRAM :