Data Mining on IBM Airline Carrier on Time Dataset

Mukund Kalantri
Reid Glaze
Rohit Kharat

Description

We will analyze data from the “Airline Reporting Carrier On-Time Performance Dataset” and focus exclusively on flights that have occurred in the past 5 years. We will be performing data mining on attributes like flight delays, cancellations, etc with respect to flight details to answer some interesting questions. Some of the questions that we have in mind are:

1. Which airports should be focused more toward infrastructure development in order to mitigate the frequency and severity of delays?

2. What are the relationships between the airline carriers and flight delays/cancellation, so that the airlines could improve their services and customers can make an informed decisions about which airlines should be preferred?

3. What are the most common reasons for delays in flights based on airports, so that the airport authorities could improve their services?

4. What are the relationships between times in an year or times in a day when there are a lot of flight cancellations/delays, so that customers can make an informed decision while booking flights with respect to their flight timings?

Additionally, we will also consider building a prediction model that will be used to calculate the delay for a flight based on the other attributes.

Prior Work

- The dataset had been used in prior work for the following data mining tasks in the capstone project of the Coursera course ‘IBM Data Analyst’:

1. Yearly no. of flights under cancellation categories
2. Average delay time by reporting airline
3. Monthly average delay for each type based on airline
4. Yearly number of flights delayed based on departure/arrival state

Datasets

- Airline Reporting Carrier On-Time Performance Dataset
- URL: Airline Dataset
- This dataset was compiled from data available on the Bureau of Transportation Statistics.
- We are using ‘requests’ and ‘tarfile’ library to load dataset on Google Colab instead of downloading it to a local machine

Proposed Work

DATA CLEANING AND DATA INTEGRATION

The full dataset has about 194 million data points which have data from the year 1984 till 2020. Since the dataset is too resource extensive, we are going to randomly sample 2 million data points from this dataset. We plan to do this step multiple times, and each time we will just pick the data points for the last 5 years. In the end, we will compile all these small datasets and take out the unique rows, to come up with a big dataset that has values for the last 5 years.

DATA PREPROCESSING

We would be doing data cleaning, data transformation, feature engineering, feature selection, and data reduction for preprocessing our data.

Proposed Work

DATA ANALYSIS

We will do statistical analysis to describe our data and data mining to find out hidden patterns based on the flight details. We would do predictive analysis to identify the likelihood of future outcomes based on historical data.

DATA VISUALIZATION

We will use data visualization techniques to present our findings for our interesting questions as well as to show any trends, correlations, or patterns in the data.

PREDICTION AND CLASSIFICATION MODELS

At the end, we would be focusing more towards building a delay prediction model, which could be used in providing reliable information by the airline carrier to the passengers.

List of Tools

Evaluation

For statistical analyses we will be doing Correlation tests, Chi-Square tests, Log Transformation, Normality tests, Student’s t-tests, ANOVA, Time series tests, etc.

For evaluating our delay prediction model, we would be using evaluation metrics, such as, accuracy, precision, recall, F1 score, Log Loss, ROC, confusion matrix, etc.

Thank You