Project Proposal Paper

Analysis of Airline Data

Mukund Kalantri
Computer Science
CU Boulder
Boulder, CO USA
muka4041@colorado.edu

Rohit Kharat
Mechanical Engineering
CU Boulder
Boulder, CO USA
rokh4336@colorado.edu

Reid Glaze
Mechanical Engineering
CU Boulder
Boulder, CO USA
regl1257@colorado.edu

PROBLEM STATEMENT / MOTIVATION

In this project, we will analyze airline data taken from approximately 2 million US domestic flights and analyze the delays and cancellations of these flights. We will aim to answer the following questions:

1) Which airports should be focused more on infrastructure development in order to mitigate the frequency and severity of delays?
2) What are the relationships between airline carriers and flight delays/cancellations, so that airlines could improve their services and customers can make 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 a month or times in a day when there are a lot of cancellations and delays so that customers can make an informed decision while booking flights with respect to their timings?

Additionally, we will also consider building a prediction model that will be used to calculate the arrival delay for a flight based on the other attributes. This could directly be used by the customers and the airport authorities.

LITERATURE SURVEY

This dataset has been used for several different data mining tasks in a Coursera specialization course called “IBM Data Analyst”:

1) Yearly number of flights canceled
2) Average delay time by airline
3) Monthly average delay for each type based on airline
4) Yearly number of flights delayed based on the arrival and departure states

The link to this specialization is: https://www.coursera.org/professional-certificates/ibm-data-analyst

This dataset has also been used in another Coursera course called “Data Analysis with R” where it has been used for:

1) Performing Exploratory Data Analysis
2) Building a prediction model for predicting flight arrival delay
3) Using the R package tidymodels to do an evaluation on the model

The link to this course is: https://www.coursera.org/learn/data-analysis-with-r

PROPOSED WORK

Data Integration

Since this dataset, we are using contains information for over 194 million flights and this amount of data is hard to work with, we are focusing exclusively on flights that have taken place in early 2022. Since it is necessary to download each month separately, we have four different files containing flight data. We plan on combining these files in order to have one workable dataset. We will use the data for the months of January to April of 2022, which includes data for over 2 million US domestic flights.

Data Preprocessing

The initial data preprocessing steps will be importing required libraries and reading the collected data. We will be employing data cleaning techniques such as removing unnecessary features, handling missing data, and filtering outlier data. Data transformation methods will be employed for transforming variables to correct data types for our computations. We will standardize the data using Scikit-learn's libraries for reducing the effect of unscaled data. This data includes 109 attributes. We will not be using all of these attributes in our evaluations, so we will be using dimensionality reduction techniques (e.g. PCA) to keep the most important features for our interesting questions. We would also be using feature engineering techniques to compute new features such as features pertaining to different time periods for our analysis. We will be using data splitting techniques for splitting the feature array and label array into training and testing sets.

Data Analysis

We will perform statistical analysis to describe our data and pattern mining techniques to find hidden patterns in the data. Regression analysis will be done to estimate the relationship between the set of features. Correlation analysis will be done to understand the effect of dependent variables on the independent variable. This will help us to perform predictive analysis to identify future outcomes. Additionally, we will plan to perform a Monte Carlo simulation for analyzing the effect of unpredictable variables on the delay or cancellation of flights. Finally, time series analysis will be used to identify trends, seasonality, and cyclic patterns in our data.

Data Visualization

We will use data visualization techniques to display the findings of our project. This can be used to depict trends, correlations, or patterns in the data. For data reduction or showing important features, chart types such as bar graphs will be useful. We will be using univariate analysis techniques such as distribution plots, box and whisker plots for outliers, and violin plots for kernel density estimation. In the bivariate analysis we will be using line plots, bar plots, and scatter plots for showing important patterns or clusters. We plan on using Tableau and Python data visualization libraries for our visualization software.

Prediction and Classification

We will be focusing on building a delay prediction model. This can be used to predict how long a given flight will be delayed. This can be estimated by examining the significance of attributes in determining the probability and length of a delay, which can be applied to new testing data. Regression analysis will be used for prediction modeling and classification models will be employed for classifying delay or cancellation types based on other features. Monte Carlo simulation methods will be helpful for us to generate models for our analysis as well as to understand the probability distributions for possible outcomes.

DATA SET

We will be using a dataset called "Airline Reporting Carrier On-Time Performance Dataset" that contains information about US domestic flights that have occurred since 1987. This data was compiled from the Bureau of Transportation Statistics.

The link to the data set is:

https://developer.ibm.com/exchanges/data/all/airline/

EVALUATION METHODS

Model evaluation will be done based on our interesting questions and statistical methods will be utilized to infer confidence for our findings.

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 doing cross-validation on our model to predict problems like overfitting and underfitting, and using regularization methods to handle these problems.

Once our model is complete, we would be using evaluation metrics, such as accuracy, precision, recall, F1 score, Log Loss, ROC, etc to evaluate our final model.

Confusion matrix will be used to describe the performance of our classification model on a set of test data generated.

TOOLS

Software

- Google Colaboratory will be used so that we can all work on code at the same time.
- Jupyter Labs will be used for data analysis with Python to solve the questions provided.
- RStudio will be used for data cleaning and data pre-processing.
- Tableau will be used for Data Visualization

Libraries

- Numpy and Pandas for data preprocessing
- Matplotlib for data visualization
- Keras and Tensorflow for building a prediction model
- SciKit learn and SciPy for Exploratory Data Analysis and model evaluation

MILESTONES

Task	**Deadline**
Finish Data Cleaning	July 22
Finish Data Preprocessing	July 25
Brainstorm which techniques to use for Specific Questions	July 27
First Draft to answer questions (other than prediction model)	July 30
Determine which features should be used for the prediction model	Aug 3
Final answer to questions (other than prediction model)	Aug 6
Finalize prediction model	Aug 8
Present Work and submit all remaining assignments	Aug 10