University of New Haven

TAGLIATELA COLLEGE OF ENGINEERING

Electrical & Computer Engineering and Computer Science

Distributed and Scalable Data Engineering
(DSCI-6007)

Electrical & Computer Engineering & Computer Science (ECECS)

SPRING 23

CONTENTS

TEAM #10

Team Members: **Questions? Contact**

Sushma Nagubandi – Team Leader ➔ **snagu12@unh.newhaven.edu**
Rishikeshwar Kankurthyshetty - Data Modeling ➔ **rkank3@unh.newhaven.edu**
Sai Kumar Reddy Tummeti – Data Scientist ➔ stumm14@unh.newhaven.edu
Sesha Sai Veerla – Data Visualization ➔ **sveer1@unh.newhaven.edu**

Title: AIRLINE DELAY PREDICTION

AIRLINE DELAY ANALYSIS

Executive Summary:

One of the most common problems in the world is flight delays. The most annoying thing that can happen to someone is a flight schedule delay, considering how important time is in everyday life. It is possible to miss an important meeting or an organized event. The earnings and losses of the airlines are significantly impacted by unavoidable flight delays. This problem has plagued all nations, not just the US, for many years. Many businesses today depend on different airlines to link them to other regions of the world, and the aviation industry is essential to the global transportation industry. According to statistics, 20% of airline flights are canceled or delayed annually, costing travelers $20 billion in lost revenue and time. To solve this issue, precise flight delay forecasting enables airlines to deal with the underlying causes of the delays in advance and enables passengers to be fully prepared for the disruption to their travel. Customer satisfaction and revenue for airline agencies can both rise with accurate time delay estimation or projection. This project's goal is to investigate the methods used to create models that predict flight delays due to different factors.

Highlights of Project:

We'll be employing a variety of supervised machine-learning techniques for this research to get the most precise forecast of whether there will be a flight delay. In this classification issue, the dependent variable (Delay) might have one of two values - 0 or 1, where 0 denotes the absence of a delay and 1 denotes the presence of a delay. This information only offers general, static flying information. Information that can change up until takeoff, like as the weather, is not included in this data because it is typically unknown until takeoff.

Submitted on: 03/03/2024

Abstract:

Flight delays can cause one to miss a crucial meeting or a scheduled function. Unavoidable flight delays have a big influence on the carriers' profits and losses. This issue has existed for many generations in various countries, not just the US. Nowadays, a lot of businesses rely on various airlines to link them to other regions of the world, and the aviation sector is crucial to the global transportation sector as well. With accurate flight delay predictions, airlines can counteract this and address the root causes of the delays, allowing passengers to be fully prepared for the disruption to their journey.

Video link is attached in project submission page.

GitHub link is attached here- UNH-DATA-ENGINEERING-TEAM10/TEAM-10-MIDTERM-PROJECT: MID TERM PROJECT (github.com)

Methodology:

The goal of this analysis is to foresee a delay. We will use the airline data from the airline dataset to predict a delay. We'll start by reviewing the data and doing some basic analysis. We will access, interact with, and visualize our data using Python.
We'll also use the data to train machine learning algorithms. The trained machine learning model will subsequently be used to create and distribute an application that will let any user provide input to determine whether or not the aircraft will be delayed.

Business Understanding:

The goal of this analysis is to foresee a delay. We will use the airline data from the airline dataset to predict a delay. We'll start by reviewing the data and doing some basic analysis.
We will access, interact with, and visualize our data using Python. We'll also use the data to train machine learning algorithms. The trained machine learning model will subsequently be used to create and distribute an application that will let any user provide input to determine whether or not the aircraft will be delayed.

Data Understanding:

The U.S. Department of Transportation's Bureau of Transportation Statistics (BTS) is the source of the data.
Large airlines' domestic flight schedule performance is monitored by the Bureau of Transportation Statistics (BTS) of the U.S. Department of Transportation (DOT). In addition to the summary tables provided on this website, the DOT also summarizes the number of on-time, delayed, canceled, and diverted flights in its monthly Air Travel Consumer Report, which is typically released thirty days following the end of the month. In June 2003, BTS started compiling information on the reasons behind aircraft delays. When the Air Travel Consumer Report is produced, raw data and summary statistics are made available to the general public.

Modeling:

We initially split the data into train and test data sets using the train test split () function of the sklearn toolkit. Pandas are used to organize and clean data. The machine learning models are trained using the data produced during data preparation.

Deployment:

Application deployment is done using the flask server.