Prediction of Passenger Count at an Air Terminal [Machine Learning based approach]

Submitted in partial fulfillment for the award of

Internship completion at

Diginique Tech Labs

Submitted by

P Harsha Satya Sai Sree

(B Tech III year,Department of Civil Engineering,IIT Roorkee)

under the guidance of

Mr. Bandenawz Bagwan

(Trainee at Diginique Techlabs)

ACKNOWLEDGEMENT

I would like to express my sincere gratitude towards Mr. Bandenawaz Bagwan for his intuitive lectures during the training period and extended support during the whole internship. I also want to thank Diginique Techlabs for providing me an opportunity to work at this organisation. I specially thank my parents for their support and encouragement.

Contents **Page no.**

Introduction:

As passenger flow at the airport increases, the passenger terminal becomes an important element of the airport. At large airport terminals ,design of terminals accounts for more than 70% of the overall infrastructure investment. The design that is ultimately adopted depends principally on the passenger volumes to be served and the type of passenger involved. Passengers are classified as business, chartered, originating or destined, transit. Each category of passenger has specific needs and in order to satisfy them the terminal should be designed accordingly. Generally airports consist of a dynamic variety of passengers and knowledge of estimated passenger count of every category of traveller at a terminal is required to make it serviceable throughout the design period.

Problem Statement:

The objective is to estimate the adjusted passenger count of a specific category of passengers at an air terminal. The passenger count depends upon many attributes like month and year of the schedule , type of price code etc . The given dataset contains all the information and with this we need to deploy a machine learning model which predicts the adjusted passenger count as accurately as possible.

Dataset description:

The given dataset consists of 15007 rows and 16 columns. Each row consists of a particular air terminal holding a unique set of attributes.

The attributes of the dataset:

- **Activity Period**: Integer formed using Year and month.
- **Operating Airline**: Name of the operating airline.
- **Operating Airline IATA Code**: Abbreviation of the corresponding Operating Airline.
- **Published Airline**: Name of the Published Airline.
- **Published Airline IATA Code**: Abbreviation of the corresponding Published Airline.
- **GEO Summary**: Type of travel (Domestic/International).
- **GEO Region**: Geographical location of the corresponding Airline.
- **Activity Type Code**: (Enplaned/Deplaned/Thru/Transit).
- **Price Category Code**: (Low fare/Other).

- **Terminal**: Specifies Terminal number.
- **Boarding Area**: Specifies Boarding location within the corresponding terminal.
- **Passenger Count**: Number of passengers.
- **Adjusted Activity Type Code**: (Enplaned/Deplaned/Thru/Transit*2).
- **Adjusted Passenger Count**: Adjusted number of passengers.
- **Year**: Year of the journey (2015 to 2016).
- **Month**: Month of the journey (January to December).

Exploratory Data Analysis:

To understand the data more clearly , visualising the data using bar plots and scatter plots is a good method.

- Monthly Passenger Count:

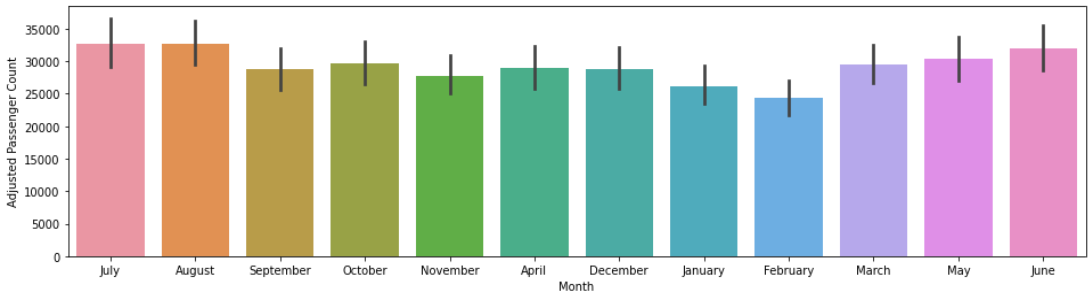

Figure:1

- Yearly Passenger Count: Figure:2 shows the increase of passenger count yearly.In 2016 , there is a drop ,but this is because of the limited data in the year.

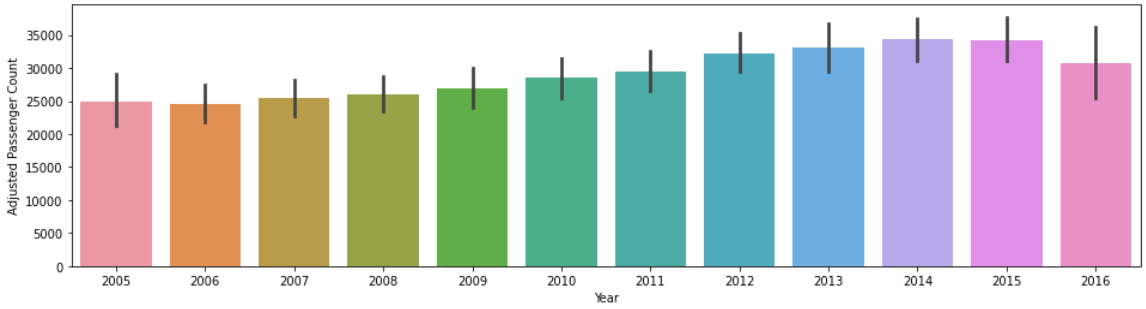

Figure:2

- Scatter plot of Adjusted Passenger Count: This scatter plot shows that there are very few examples of ‘Thru/Transit’ type in the dataset.So, the data is skewed.

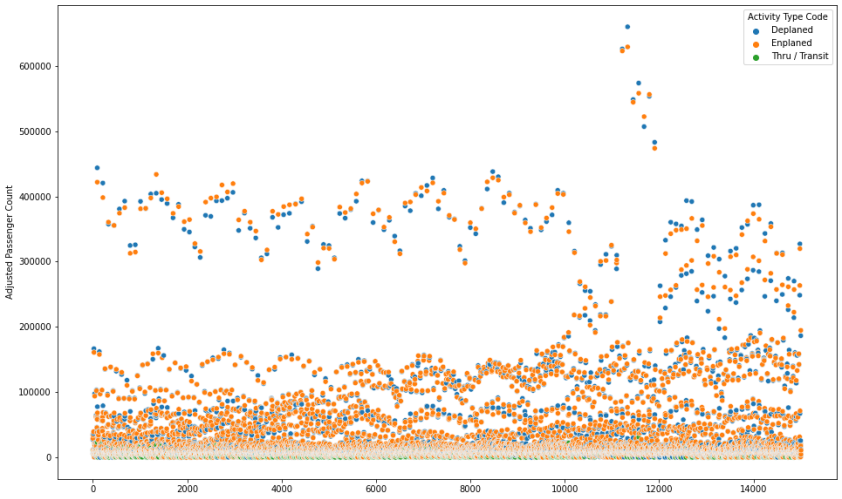

Figure:3

- The below count plot indicates that the data is slightly skewed towards International.

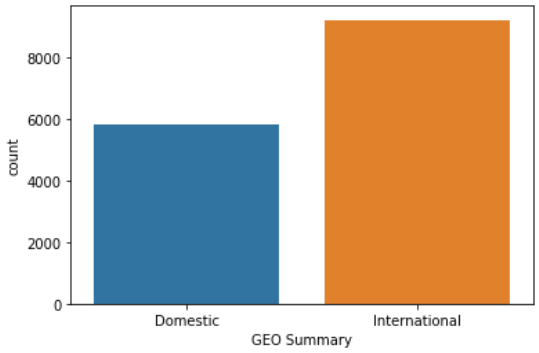

figure:4

- Count Plot of GEO Regions of the given Airlines is shown in figure:5. The figure tells that the data contains more examples of the airlines belonging to the US.

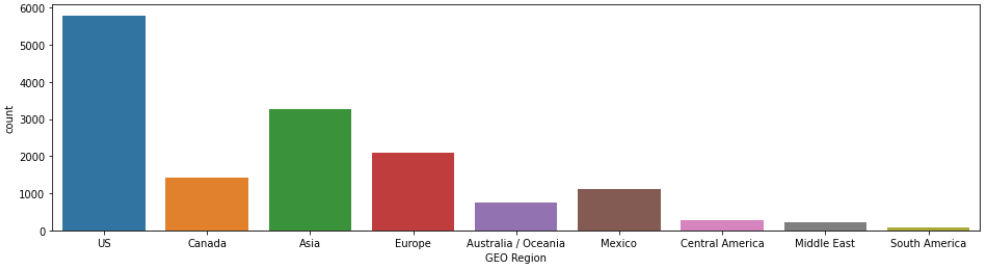

figure:5

- Count Plot of Price Category Code is as shown in figure:6.

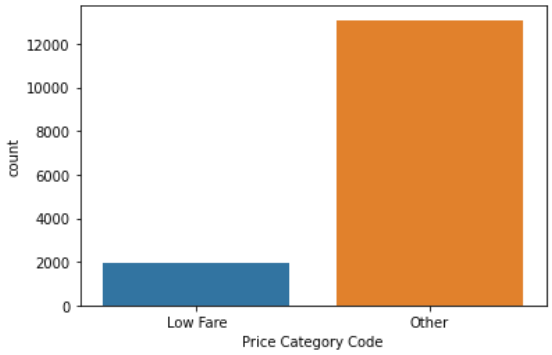

figure:6

➢ From the above visualisations , we can see that most of the data is skewed to any one category of the attribute rather than unbiased distribution. One plausible reason for this might be, as we know that this design of terminals should be taken for larger air ports , so the data is mostly skewed to regions of larger terminals like the US , International flights instead of Domestic, and hence the expensive ticket fares.

- Outliers in the Target feature (Adjusted Passenger Count):
  A box plot depicting the distribution of Adjusted Passenger Count is shown in figure:7.

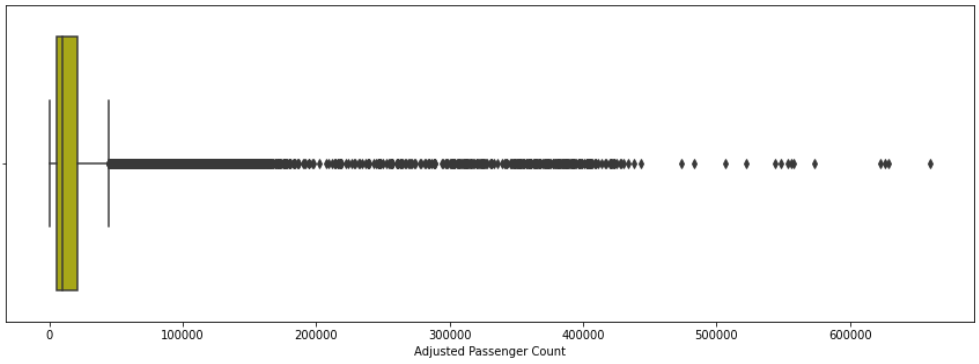

figure:7

➢ There are 322 outliers in the data. And all the outliers are from ‘United Airlines’. As all the outliers belong to the same operating airline, this indicates that these outliers are not fallacy. Eliminating outliers leads to loss of important information. So the best practise to treat outliers is by replacing them with (0.75 quantile + 1.5*(Inter quartile range)) if we chose to use basic models like Linear Regression. But algorithms like Decision Tree regression , Random forest regression are robust to this type of analysis.

Data Preprocessing:

- The ‘Operating Airline IATA Code’ and ‘Published Airline IATA Code’ columns contain null values. These columns are deleted without any loss of information as they give the same meaning as ‘Operating Airline’ and ‘Published Airline’ columns.
- The ‘Passenger Count’ column is removed as it is highly correlated with ‘Adjusted Passenger Count’.
- The ‘Adjusted Activity Type Code’ and ‘Activity Type Code’ columns have the same information. So, ‘Activity Type Code’ column is discarded.
- Treating outliers (if Model==Linear Regression). Boxplot diagram after transforming the outliers is shown in Figure:8.
- Split the data into training and test sets (1:4 ratio respectively).
- The dataset contains heterogeneous types of data (categorical and numerical variables). To make the data fit for the model , the categorical variables are encoded to numerical columns using One Hot Encoding and Label Encoding.

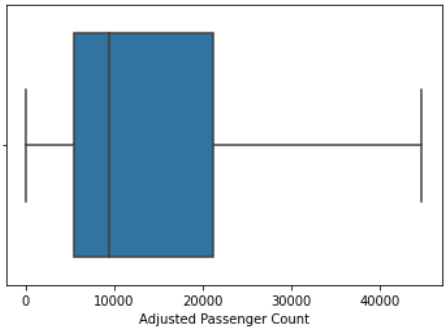

Figure:8

Methodology:

- **<u>Linear Regression:</u>**

  linear regression is a powerful supervised machine learning algorithm that can help us model linear relationships between two variables. Simple linear regression is often a good starting point for exploring our data and thinking about how to build more complex models.

- **Decision Tree Regression:**

  Decision trees build regression or classification models in the form of a tree structure. It **breaks down a dataset into smaller and smaller subsets** while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

- **Random Forest Regression:**

  Random Forest Regression is **a supervised learning algorithm that uses ensemble learning methods for regression**. A Random Forest operates by constructing several decision trees during training time and outputting the mean of the classes as the prediction of all the trees.

Results:

- Linear Regression:

  The data to be fit for the Linear Regression model should be free of outliers as this might lead to predict the results with high error.

- Decision Tree Regression:

  This model does not require any specific preprocessing as it is robust enough to understand 'United Airlines' (outliers in the dataset) as a specific subset.

- Random Forest Regression:

  The number of decision trees is chosen to be 10 , and the output of passenger count is resulted from the average of outputs of the 10 unique decision trees.

Accuracy:

Model	R-Square Error
Linear Regression	0.786
Decision Tree Regression	0.9868
Random Forest Regression	0.987

Table:1

Conclusion and Future work:

In this project the various factors affecting the passenger count at an air terminal are described in detail. From Table:1 we can see that the predictions made by Random Forest Regression outperformed the Linear regression and Decision tree regression models. R-square score of 0.987 is acceptable and hence this model can be used before designing an airport terminal in order to make their passengers more comfortable.

In the future, I would like to explore deep learning and data augmentation. I would also like to see how covid 19 affected the passenger count. I also would like to deploy the best model into the flask application in the future.

This github link consists of all the codes discussed in the report above.