Project – Flight Booking Price Prediction

Problem Statement:

This project's goal is to examine a dataset of flight bookings that was acquired from a well-known website that sells airline tickets. The dataset includes booking-related information, such as passenger, flight, and booking transaction details. To extract significant insights, a thorough Exploratory Data Analysis (EDA), statistical techniques, and machine learning algorithms are to be used. With the use of statistical analysis, predictive modelling, and data exploration, this study seeks to identify important patterns, trends, and relationships in the dataset. The results will offer useful data that can improve the general comprehension of passenger behaviour, streamline the booking procedure, and possibly even improve the platform's user experience.

Project Objective:

- To gain basic understanding of the data.
- To perform EDA for Initial analysis.
- To perform various supervised Machine learning models to predict the booking prices.

Data Description:

```
#   Column            Non-Null Count   Dtype
---  ------            --------------   -----
0   Unnamed: 0        300153 non-null  int64
1   airline           300153 non-null  object
2   flight            300153 non-null  object
3   source_city       300153 non-null  object
4   departure_time    300153 non-null  object
5   stops             300153 non-null  object
6   arrival_time      300153 non-null  object
7   destination_city  300153 non-null  object
8   class             300153 non-null  object
9   duration          300153 non-null  float64
10  days_left         300153 non-null  int64
```

```
11  price              300153 non-null  int64
```

The data set contains some 3 lakhs records with 10 columns**.**

Data Preprocessing:

Data preprocessing is an essential step because it helps us in building more accurate Machine learning model. It improves the model's accuracy to predict the target variable.

Steps in Data Preprocessing:

- Data cleaning
- Data Transformation
- Data reduction

Data cleaning:

Data cleaning includes various tasks like handling missing data points and outlier treatment. uncleaned data set can cause major problems in our machine learning models. For example, extreme outliers can cause our model to give biased results. That's why it is very important to treat any such inconsistencies.

Strategies to handle missing data and outliers:

1. Imputation of missing value with the central values like mean, median and mode.
2. Removing the outliers from the data.
3. Imputing the outliers with lower bound or upper bound value.

For this project we have performed Regression analysis both on base model and dataset without outliers.

Data Transformation:

Data transformation is the process of converting raw data into standard format which helps the machine to easily understand the data. It involves structuring, cleaning, transforming the data to increase the accuracy of the machine learning models.

Data Transformation involves many activities:

1. Changing the data types of the variables.
2. Encoding categorical variables.
3. Normalising the data
4. Aggregating the data

For this project we have performed label encoding to covert categorical variables to int variables.

Data Reduction:

Data reduction is a group of techniques used to select the most important features from the datasets which will contribute to better models building and accuracy.

For this study we have conducted variance inflation factor to check for multicollinearity and we have dropped the feature accordingly.

Choosing the Algorithm for the Project:

Ordinary Least Square Method for Regression analysis:

Ordinary least squares (OLS) regression is an optimization strategy that helps you find a straight line as close as possible to your data points in a linear regression model. OLS is considered the most useful optimization strategy for linear regression models as it can help you find unbiased real value estimates for your alpha and beta.

Correlation Analysis:

Correlation analysis is a statistical technique used to evaluate the strength and direction of the linear relationship between two quantitative variables.

Random Forest Algorithm:

- Random Forest model is an ensemble technique using decision tree as a base.
- Random forest performs bootstrap sampling of the features.
- The bootstrapping is done at the tree level.

Benefits of Random Forest model:

- Works for both regression and classification.
- Gives better accuracy because of the different samples it is considering.
- Handles missing values
- Prevents the overfitting of the model.

Model Evaluation and Technique:

Mean Absolute Error (MAE): Mean Absolute Error is referred to as MAE. It is a figure used to evaluate how well a regression model predicts future events. The average absolute difference (MAE) between the expected and real (also known as ground truth) values is measured.

Formula for calculating MAE:

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Root mean squared error (RMSE): Root mean squared error is used to evaluate Linear regression model. It is a root of mean squared error. Here error is the difference between the actual and predicted values

Formula for RMSE:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Form OLS summary of base-model:

MAE = 4626.1

RMSE = 7005.1

Form OLS summary of the data without outliers:

MAE = 4659.04

RMSE = 7026.8

From the above analysis we can deduct that our base model has less errors compared to the model which we ran without outliers.

From Random Forest Regressor Model:

MAE = 1109.83

RMSE = 2783.05

The random forest model has given us the best results compared to our linear regression models.

Inferences from the Models:

When compared to the model run without outliers, the base model's errors (both MAE and RMSE) are somewhat lower. This shows that in terms of prediction accuracy, the base model—which contains outliers—performs somewhat better. In terms of prediction accuracy, the Random Forest Regressor model performs noticeably better than both OLS models. In comparison to the linear regression models, the Random Forest model yields predictions that are significantly more accurate, as evidenced by the significantly lower MAE and RMSE values.

Conclusion:

Out of all the models that were assessed, the Random Forest Regressor proved to be the most effective one in terms of prediction accuracy. When selecting a model, it's critical to take the problem's context and particular requirements into account. Although Random Forest performs better in this instance, other considerations like interpretability and processing complexity may have an impact on the optimal model selection for implementation. To fully understand the causes influencing the Random Forest model's enhanced performance, more investigation and study could be needed.

References:

https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-linear-regression/

https://towardsdatascience.com/gradient-descent-algorithm-a-deep-dive-cf04e8115f21#:~:text=Gradient%20descent%20(GD)%20is%20an,e.g.%20in%20a%20linear%20regression).

https://builtin.com/data-science/random-forest-python