
# Flight Price Project

## Submitted by:

Sristi Kushwaha

# ACKNOWLEDGMENT

I would like to express my special thanks of gratitude to my SME Sapna Verma as well as our organization who gave me the golden opportunity to do this wonderful project on the topic (Flight Price Project), which also helped me in doing a lot of Research and I came to know about so many new things. I am really thankful to them.

# INTRODUCTION

- **BUSINESS PROBLEM FRAMIMG:**

  Flight ticket prices can be something hard to guess, today we might see a price, check out the price of the same flight tomorrow, it will be a different story. We might have often heard travellers saying that flight ticket prices are so unpredictable. So, we collected the different-2 airlines data which are key part to make a model. And we are going to make a model to predict the future price of the flight.

- **CONCEPTUAL BACKGROUND OF THE DOMAIN PROBLEM:**

  As we know that nowadays, many numbers of people using flights has increased significantly. It is difficult for airlines to maintain prices since prices change dynamically due to different situations. That is why we will try to use machine learning to solve this problem. This can help airlines by predicting how they can maintain their prices. It can also help customers to predict future flight prices so that they can plan their journey accordingly.

- **MOTIVATION FOR THE PROBLEM UNDERTAKEN:**

  We are required to model the price of flight with the available independent variables. This model will then be used by the management to understand how exactly the prices vary with the variables. They can accordingly manipulate the strategy of the airlines and concentrate on areas that will yield high returns. Further, the model will be a good way for the management to understand the pricing dynamics of a new market. And it will be also very useful for customers while booking tickets. That is why we are here to do this project.

# ANALYTICAL PROBLEM FRAMING

- **MATHEMATICAL/ANALYTICAL MODELING OF THE PROBLEM:**

  We start with doing exploratory data analysis with the help of pandas and seaborn. First, we got some data insights from it. We came to know about different types of columns. And with the help of matplotlib and seaborn we did visualizations to analyze the effect of features on our target variable. And then, we did some preprocessing of the data. After above procedures, we splitted the data into training part and testing part and apply different-2 models and for evaluation we used some evaluation matrices.

- **DATA SOURCE AND THEIR FORMATS:**

  We collected the data from the flight booking sites and convert it into csv format. Our dataset has 1710 rows and 9 columns.

- **DATA PROCESSING DONE:**

  Since data is not so clean for that we have to follow some step to clean it. First of all, checked for the null values and its datatype. Some of the columns are not in its right data type so we convert it into appropriate data type and we also clean some unwanted expressions from columns. And we also did some feature engineering as needed.

- **H/W AND S/W REQUIREMENT AND TOOLS USED:**

  We used PCs as hardware tool. For software we used Anaconda prompt in which we used jupyter notebook. In jupyter notebook we used many

libraries like Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn etc. Pandas is used for data structure and their computations, NumPy is used for arrays and some statistical computations, Scikit-learn is used for machine - learning, encoding the data, scaling the data etc. Matplotlib is used for plots and graphs, Seaborn is used for different types of plots advanced version of Matplotlib.

# MODEL DEVELOPMENT AND EVALUATIONS

- **APPROACH FOLLOWED:**

  I followed few steps to solve this problem such as knowing about the data and its features and their stats like mean, median, standard deviation and some plot so that I can identify their distribution along the axis. Next is finding if there are any outliers present or not as well as skewness so that we can remove those problems. We split the data into training part and testing part. We scale the data with Standard Scaler so that data would be ready for model training. And then train the model and compare their accuracy with the help of r2 score and root mean squared error as this was a regression problem. On the basis of that we finalize our best model.

- **TESTING OF IDENTIFIED APPROACH:**

  There are many algorithms which are used for training and testing the data which are listed below.

  - Linear Regression
  - Decision Tree Regressor
  - Adaboost Regressor
  - KNeighbors Regressor
  - Random Forest Regressor
  - XGBRegressor

  Among all Linear Regression, KNeighbors Regressor and AdaBoost Regressor worked very poorly and Decision Tree Regressor, Random Forest Regressor and XGBRegressor worked quite well.

- **RUN AND EVALUATE SELECTED MODELS:**

We trained different-2 models, here we provide model training and its r2 score, cross validation score and root mean squared error to evaluate the model performance.

We did hyper parameter tuning after training all models using grid search cv to improve the accuracy of the models.

After hyperparameter tuning we can see that xgb regressor have highest r2_score and least RMSE value. So, it will be our final model.

- **VISUALIZATIONS:**

We did some different-2 visualizations to know the relationship between independent and dependent variables.

First, we plot count plot to analyze the how the columns are categorizes within its category. And then distribution plot for numerical columns to analyze the distribution of data along the axis.

We plot bar plot between categorical columns and target column to analyze the how the data are affected the target column.

We draw box plot to know about outliers in independent columns, next heatmap to analyze the relationship between independent and dependent variables and of course to know about multicollinearity between independent variables.

- **INTERPRETATION OF THE RESULTS:**

While visualization of our data we try to know what are the relationships among the data according to that we made assumptions also what will be the further process.

While preprocessing the data we try to clean the data so that any unwanted noise will not there which affects our prediction accuracy and also, we analyze some statistical properties of data so that we can understand what type of data we have.

Our final process is modeling where we train our data and test that how much accurate our model can predict. And on the basis of metrices we finalize our best model.

# CONCLUSION

- **KEY FINDING AND CONCLUSION OF THE STUDY:**

  To conclude, the application of machine learning in flight price research is still at an early stage. We hope this study has moved a small step ahead in providing some methodological and empirical contributions to flight prices, and presenting an alternative approach to the valuation of flight prices.

- **LEARNING OUTCOMES OF THE STUDY WITH RESPECT OF DATA SCIENCE:**

  While doing this project I learned so much about handling the real time data. During visualization I understand that how data are dependent or independent to each other and what can be the relationship between them. while cleaning the data, I realize one of the most important point that cleaning is very important if we want better model for accurate prediction, and also learned different-2 technique to clean the data. We used many algorithms such as Linear Regression, Decision Tree Regressor, Random Forest Regressor, XGBoost Regressor, AdaBoost Regressor etc. And we finalize the best algorithm as XGB Regressor.

- **LIMITATIONS OF THIS WORK AND SCOPE FOR FUTURE WORK:**

  Improvement in computing technology has made it possible to examine social information that cannot previously be captured, processed and analyzed. New analytical techniques of machine learning can be used in flight price research.