

Flight Price Prediction

Submitted by:

HARINATH MALLELA

# ACKNOWLEDGMENT

A unique opportunity like this comes very rarely. It is indeed a pleasure for me to have worked on this project. The satisfaction that accompanies the successful completion of this project is incomplete without the mention of the people whose guidance has made it possible for me to complete this project.I am grateful to my internship company **Flip Robo Technologies** with its ideals and inspiration for providing me with facilities that has made this project a success.

I am grateful to **Vaishali Singh** for providing the opportunity to work as a intern in **Flip Robo Technologies**

I like to acknowledge the effort put in by our SME **Sapna Verma** in helping me understand the relevant concepts related to Data preprocessing and providing guidance when necessary.

**HARINATH MALLELA**

# INTRODUCTION

## Business Problem Framing

* Anyone who has booked a flight ticket knows how unexpectedly the prices vary. The cheapest available ticket on a given flight gets more and less expensive over time. This usually happens as an attempt to maximize revenue based on -

1. Time of purchase patterns (making sure last-minute purchases are expensive)
2. Keeping the flight as full as they want it (raising prices on a flight which is filling up in order to reduce sales and hold back inventory for those expensive last-minute expensive purchases)

* So, you have to work on a project where you collect data of flight fares with other features and work to make a model to predict fares of flights.

## Conceptual Background of the Domain Problem

Airline companies use complex algorithms to calculate flight prices given various conditions present at that particular time. These methods take financial, marketing, and various social factors into account to predict flight prices. Nowadays, the number of people using flights has increased significantly. It is difficult for airlines to maintain prices since prices change dynamically due to different conditions. That’s why we will try to use machine learning to solve this problem. This can help airlines by predicting what prices they can maintain. It can also help customers to predict future flight prices and plan their journey accordingly.

## Review of Literature

Nowadays, airline ticket prices can vary dynamically and significantly for the same flight, even for nearby seats within the same cabin. Customers are seeking to get the lowest price while airlines are trying

to keep their overall revenue as high as possible and maximize their profit. Airlines use various kinds of computational techniques to increase their revenue such as demand prediction and price discrimination. From the customer side, two kinds of models are proposed by different researchers to save money for customers: models that predict the optimal time to buy a ticket and models that predict the minimum ticket price. In this paper, we present a review of customer side and airlines side prediction models. Our review analysis shows that models on both sides rely on limited set of features such as historical ticket price data, ticket purchase date and departure date. Features extracted from external factors such as social media data and search engine query are not considered. Therefore, we introduce and discuss the concept of using social media data for ticket/demand prediction.

## Motivation for the Problem Undertaken

Motivated by previous studies, we can think of various additional useful features from social media that can possibly forecast airlines passenger demand and or ticket prices. For example, sentiment analysis of different twitter hash tags could convey the presence of some event at a flight origin/destination city that improves the prediction of ticket price/demand. This kind of feature extraction might involve searching for special keywords or group of terms, determining the number of times they appear, understanding the location and the date, their context etc.

# Analytical Problem Framing

## Mathematical/ Analytical Modeling of the Problem

* Mathematical Summary:

Dimensions of Dataset: There are 9 columns and 1611 rows in this dataset

Null Values: There are no null values in this dataset Skewness: There is no skewness in columns.

## Data Sources and their formats

Standard deviation is very normal in most of the columns.

There is not much difference between mean and 50th percentile, which means data is skewed

There is not much difference between 75th percentile and max, which means there are outliers

## Data Sources and their formats

All the objects were object data type after scraping before pre- processing.

## Data Pre-processing Done

* Feature Extraction:

At first we had only 9 columns in which Date column consists of Day of the week and Days prior booking. After extracting the required fields using feature extraction we got 11 columns.

## Date Column (before):

Graphical user interface, application

Description automatically generated

After extracting week day and days prior booking: Note: I have scraped data of 13th October.

Graphical user interface, application, Teams

Description automatically generated

1. Extracting Day

A picture containing text

Description automatically generated

1. Extracting the days prior booking

Text

Description automatically generated

1. Converting Duration columns in minutes

Graphical user interface, text, application, email

Description automatically generated

1. Feature extraction of next day arrival column from arrival date

Graphical user interface, text, application

Description automatically generated

1. Extracting date from Arrival date column using Regular expressions

Graphical user interface, text, application

Description automatically generated

1. Feature engineering of Arrival part of the day and departure part of the day

Graphical user interface, text, application, email

Description automatically generated

1. Converting week number to weekday

Text

Description automatically generated with medium confidence

1. Final Data Frame after Data Pre-processing

Table

Description automatically generated

1. Scaling the Input data

Table

Description automatically generated

1. PCA (principle component Analysis)

Chart

Description automatically generated with medium confidence

* Data Inputs- Logic- Output Relationships
* We see good correlation between day’s prior booking and Price as well as weekday and price.
* Airfares change very frequently
* Airfares tend to go down over time

## Hardware and Software Requirements and Tools Used

* Machine: Can use a laptop/desktop.
* Operating system: Windows 8 or 10, Mac OS X 10.9 Mavericks or Higher
* RAM & Processor: 4 GB+ RAM, i3 5th Generation 2.2 Ghz or equivalent/higher
* Tools Used: Jupyter Notebook, Microsoft Excel.
* Libraries Used : Sklearn,Seaborn,Matplotib,Pandas,Numpy

,warnings

# Model/s Development and Evaluation

## Identification of possible problem-solving approaches (methods)

* I started by extracting all the random flights data from Yatra website from five different location which include Delhi, Mumbai, Kolkata and Bangalore.
* After extracting the required information, I had column called Date which consist of three different inputs which days prior booking and day of the week.
* I used Regular Expression and other methods to extract the information from the existing column.
* I then dropped the old columns after extracting the required information.
* Converted Object data type like price to INT after replacing Symbols.
* Testing of Identified Approaches (Algorithms) Below are the algorithms used for the training and testing. Lr = LinearRegression()

dtc = DecisionTreeRegressor()

knn = KNeighborsRegressor(n\_neighbors=5) rf = RandomForestRegressor()

ada = AdaBoostRegressor()

1. Linear Regression

Text

Description automatically generated

Main Code for Running Multiple model at a time

Graphical user interface, text, application, email

Description automatically generated

1. Decision Tree Regressor

Text

Description automatically generated with medium confidence

1. KNN Regressor

Text

Description automatically generated

1. Random Forest Regressor

Text

Description automatically generated

1. AdaBoost Regressor

Text

Description automatically generated

## Run and Evaluate selected models

These are the algorithms used along with the snapshot of their code.

## Key Metrics for success in solving problem under consideration

#### R2 Score:

##### R-squared is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.

Best possible score is 1.0 and it can be negative (because the model can be arbitrarily worse). A constant model that always predicts the expected value of y, disregarding the input features, would get a R2 score of 0.0.

We obtained the r2 score of 91.12 % which is very good.

* **Root Mean Squared Error (RMSE):**

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points. RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit.

The RMSE is the square root of the variance of the residuals. It indicates the absolute fit of the model to the data–how close the observed data points are to the model's predicted values. Whereas R-squared is a relative measure of fit, RMSE is an absolute measure of fit. Lower values of RMSE indicate better fit.

Root Mean Squared of this Random forest model is 1044 which is low compared to others.

* **Mean Squared Error:**

Mean squared error is the average of the squared error that is used as the loss function for least squares regression. It is the sum, over all the data points, of the square of the difference between the predicted and actual target variables, divided by the number of data points.

Mean Squared of this Random forest model is 1090567.711779662 which is low compared to others.

* **Mean Absolute Error:**

In the context of machine learning, absolute error refers to the magnitude of difference between the prediction of an observation and the true value of that observation. MAE takes the average of absolute errors for a group of predictions and observations as a measurement of the magnitude of errors for the entire group.

Mean Absolute error of this random forest model is 651.2223661971831.

* Visualizations

The plots made along with their pictures and the inferences and observations obtained from those.

**Outliers in Target column:**

Chart, box and whisker chart

Description automatically generated

**Distribution of the target column:**

Chart, histogram

Description automatically generated

**Bivariate Analysis with Target Variable:**

Chart, bar chart

Description automatically generated

Chart, bar chart, box and whisker chart

Description automatically generated

Chart, bar chart, box and whisker chart

Description automatically generated

Chart, bar chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart, bar chart, box and whisker chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, scatter chart

Description automatically generated

1. Do airfares change frequently?

Chart

Description automatically generated

2. Do they move in small increments or in large jumps?

Chart

Description automatically generated

3. Do they tend to go up or down over time?

Chart, scatter chart

Description automatically generated

4. What is the best time to buy so that the consumer can save the most by taking the least risk?

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

##### Does price increase as we get near to departure date?

##### Calendar Description automatically generated with medium confidence

6. Is Indigo cheaper than Jet Airways?

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

## Interpretation of the Results

I choose Linear Regression as the final model since it has the best r2 score among all the model of 91%. This model is very less difference between cross validation cross and r2 score. So, this model is not over fitted or under fitted.

**Hyper parameter tuning of Random Forest Regressor**

Graphical user interface, text, application, email

Description automatically generated

**Output of Random Forest Regressor with best parameters obtained from hyper parameter tuning:**

Graphical user interface, text, application, email

Description automatically generated

**Saving the Final model:**

Serialization using joblib.

**Pickled model as a file using joblib:** Joblib is the replacement of pickle as it is more efficient on objects that carry large numpy arrays. These functions also accept file-like object instead of filenames.

**Joblib.dump** to serialize an object hierarchy.

Graphical user interface, text, application, email

Description automatically generated

**Let’s plot y\_test vs predicted:**

Chart, scatter chart

Description automatically generated

**Final Predicted Car Prices:**

Table

Description automatically generated

**CONCLUSION**

* Key Findings and Conclusions of the Study

Though this is the simplest model we’ve built till now, the final predictors still seem to have high correlations. One can go ahead and remove some of these features, though that will affect the adjusted- r2 score significantly (you should try doing that).

Thus, for now, the final model consists of the 11 variables mentioned above.

* Learning Outcomes of the Study in respect of Data Science

I choose Random Forest Regressor algorithm as my final model since it was having least difference between its cross validation score and r2 score.

I learned how to use regular expression to extract the required information from the existing columns

## Limitations of this work and Scope for Future Work

##### Yes there is still room for improvement, like doing a more

Web scraping from different websites, extensive feature engineering, by comparing and plotting the features against each other and identifying and removing the noisy features. The values of R-squared obtained from the algorithm give the accuracy of the model. In the future, if more data could be scraped such as the business class, so predicted results will be more accurate.