Flight Fare Prediction System

Vinod Kimbhaune, Harshil Donga, Asutosh Trivedi,
Sonam Mahajan and Viraj Mahajan

May 19, 2021

FLIGHT FARE PREDICTION SYSTEM

Dr. V. V. Kimbhahune
Department of Computer Engineering
Smt. Kashibai Navale College of Engineering Pune, India

Harshil Donga
Department of Computer Engineering
Smt. Kashibai Navale College of Engineering Pune, India

Ashutosh Trivedi
Department of Computer Engineering
Smt. Kashibai Navale College of Engineering
Pune, India

Sonam Mahajan
Department of Computer Engineering
Smt. Kashibai Navale College of Engineering
Pune, India

Viraj Mahajan
Department of Computer Engineering
Smt. Kashibai Navale College of Engineering
Pune, India

Abstract— Travelling through flights has become an integral part of today's lifestyle as more and more people are opting for faster travelling options. The flight ticket prices increase or decrease every now and then depending on various factors like timing of the flights, destination, duration of flights. various occasions such as vacations or festive season. Therefore, having some basic idea of the flight fares before planning the trip will surely help many people save money and time. In the proposed system a predictive model will be created by applying machine learning algorithms to the collected historical data of flights. This system will give people the idea about the trends that prices follow and also provide a predicted price value which they can refer to before booking their flight tickets to save money. This kind of system or service can be provided to the customers by flight booking companies which will help the customers to book their tickets accordingly.

I. Introduction

This project aims to develop an application which will predict the flight prices for various flights using machine learning model. The user will get the predicted values and with its reference the user can decide to book their tickets accordingly.

In the current day scenario flight companies try to manipulate the flight ticket prices to maximize their profits. There are many people who travel regularly through flights and so they have an idea about the best time to book cheap tickets. But there are also many people who are inexperienced in booking tickets and end up falling in discount traps made by the companies where actually they end up spending more than they should have. The proposed system can help save millions of rupees of customers by proving them the information to book tickets at the right time.

The proposed problem statement is "Flight Fare prediction system".

II. Related Work

Proposed study[1] Airfare price prediction using machine learning techniques, For the research work a dataset consisting of 1814 data flights of the Aegean Airlines was collected and used to train machine learning model. Different number of features were used to train model various to showcase how selection of features can change accuracy of model.

In case study[2] by William groves an agent is introduced which is able to optimize purchase timing on behalf of customers. Partial least square regression technique is used to build a model.

In a survey paper [4] by supriya rajankar a survey on flight fare prediction using machine learning algorithm uses small dataset consisting of flights between Delhi and Bombay. Algorithms such as K-nearest neighbours (KNN), linear regression, support vector machine (SVM) are applied.

Research done by Santos[3] analysis is done on air fare routes from Madrid to London, Frankfurt, New York and Paris over course of few months. The model provides the accepted number of days before buying the flight ticket.

Tianyi wang[5] proposed framework where two databases are combined together with macroeconomic data and machine learning algorithms such as support vector machine, XGBoost are used to model the average ticket price based on source and destination pairs. The framework achieves a high prediction accuracy 0.869 with the adjusted R squared performance metrics

In[6] the research a desired model is implemented using the Linear Quantile Blended Regression methodology for San Francisco–New York course where each day airfares are given by online website. Two features such as number of days for departure and whether departure is on weekend or weekday are considered to develop the model.

III. IMPLEMENTATION

For this project, we have implemented the machine learning life cycle to create a basic web application which will predict the flight prices by applying machine learning algorithm to historical flight data using python libraries like Pandas, NumPy, Matplotlib, seaborn and sklearn. Figure.1 shows the steps that we followed from the life cycle:

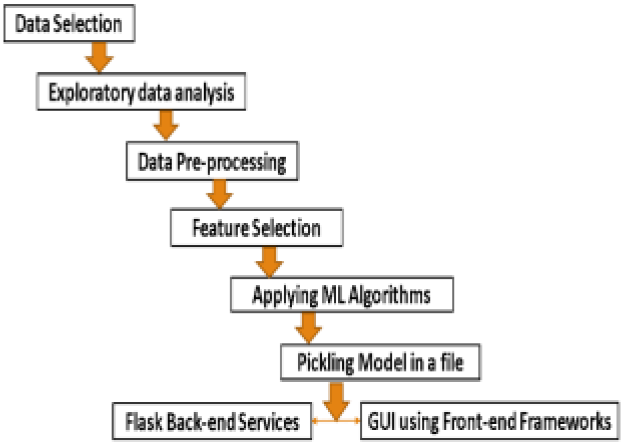

Fig. 1. Machine Learning Life Cycle

Data selection is the first step where historical data of flight is gathered for the model to predict prices. Our dataset consists of more than 10,000 records of data related to flights and its prices. Some of the features of the dataset are source, destination, departure date, departure time, number of stops, arrival time, prices and few more.

In the exploratory data analysis step, we cleaned the dataset by removing the duplicate values and null values. If these values are not removed it would affect the accuracy of the model. We gained further information such as distribution of data.

Next step is data pre-processing where we observed that most of the data was present in string format. Data from each feature is extracted such as day and month is extracted from date of journey in integer format, hours and minutes is extracted from departure time. Features such as source and destination needed to be converted into values as they were of categorical type. For this One hot-encoding and label encoding techniques are used to convert categorical values to model identifiable values.

Feature selection step is involved in selecting important features that are more correlated to the price. There are some be selected and passed to the group of models. Random forest basically uses group of decision trees as group of models. Random amount of data is passed to decision trees and each decision tree predicts values according to the dataset given to it. From the predictions made by the decision trees the features such as extra information and route which are unnecessary features which may affect the accuracy of the model and therefore, they need to be removed before getting our model ready for prediction.

After selecting the features which are more correlated to price the next step involves applying machine algorithm and creating a model. As our dataset consist of labelled data, we will be using supervised machine learning algorithms also in supervised we will be using regression algorithms as our dataset contains continuous values in the features. Regression models are used to describe relationship between dependent and independent variables. The machine learning algorithms that we will be using in our project are:

Linear Regression

In simple linear regression there is only one independent and dependent feature but as our dataset consists of many independent features on which the price may depend upon, we will be using multiple linear regression which estimates relationship between two or more independent variables and one dependent variable.

The multiple linear regression model is represented by:

$$\boldsymbol{Y} = \beta_0 \boldsymbol{x1} + ... + \beta_n \boldsymbol{xn} + \boldsymbol{\varepsilon}$$

$\boldsymbol{Y}$ = the predicted value of the dependent variable

$\boldsymbol{Xn}$ = *the independent variables*

β_n = independent variables coefficients

$\boldsymbol{\varepsilon}$ = *y-intercept when all other parameters are 0*

Decision Tree

Decision trees are basically of two types classification and regression tree where classification is used for categorical values and regression is used for continuous values. Decision tree chooses independent variable from dataset as decision nodes for decision making.

It divides the whole dataset in different sub-section and when test data is passed to the model the output is decided by checking the section to which the datapoint belong to. And to whichever section the data point belongs to the decision tree will give output as the average value of all the datapoints in the sub-section

Random Forest

Random Forest is an ensemble learning technique where training model uses multiple learning algorithms and then combine individual results to get a final predicted result. Under ensemble learning random forest falls into bagging category where random number of features and records will

average value of the predicted values if considered as the output of the random forest model.

Performance Metrics

Performance metrics are statistical models which will be used to compare the accuracy of the machine learning models trained by different algorithms. The sklearn.metrics module will be used to implement the functions to measure the errors from each model using the regression metrics. Following metrics will be used to check the error measure of each model.

MAE (Mean Absolute Error)

Mean Absolute Error is basically the sum of average of the absolute difference between the predicted and actual values.

$$\mathbf{MAE = 1/n[\sum(y-\acute{y})]}$$

y = actual output values,

ý = predicted output values

n = Total number of data points

Lesser the value of MAE the better the performance of your model.

MSE (Mean Square Error)

Mean Square Error squares the difference of actual and predicted output values before summing them all instead of using the absolute value.

$$\mathbf{MSE = 1/n[\sum(y-\acute{y})^2]}$$

y=actual output values

ý=predicted output values

n = Total number of data points

MSE punishes big errors as we are squaring the errors. Lower the value of MSE the better the performance of the model.

RMSE (Root Mean Square Error)

RMSE is measured by taking the square root of the average of the squared difference between the prediction and the actual value.

$$\mathbf{RMSE = \sqrt{}1/n[\sum(y-\acute{y})^2]}$$

y=actual output values

ý=predicted output values

n = Total number of data points

RMSE is greater than MAE and lesser the value of RMSE between different model the better the performance of that model.

R^2 (Coefficient of determination)

It helps you to understand how well the independent variable adjusted with the variance in your model.

$$\mathbf{R^2} = 1 - \frac{\sum(\acute{y}-\bar{y})^2}{\sum(y-\bar{y})^2}$$

The value of R-square lies between 0 to 1. The closer its value to one, the better your model is when comparing with other model values.

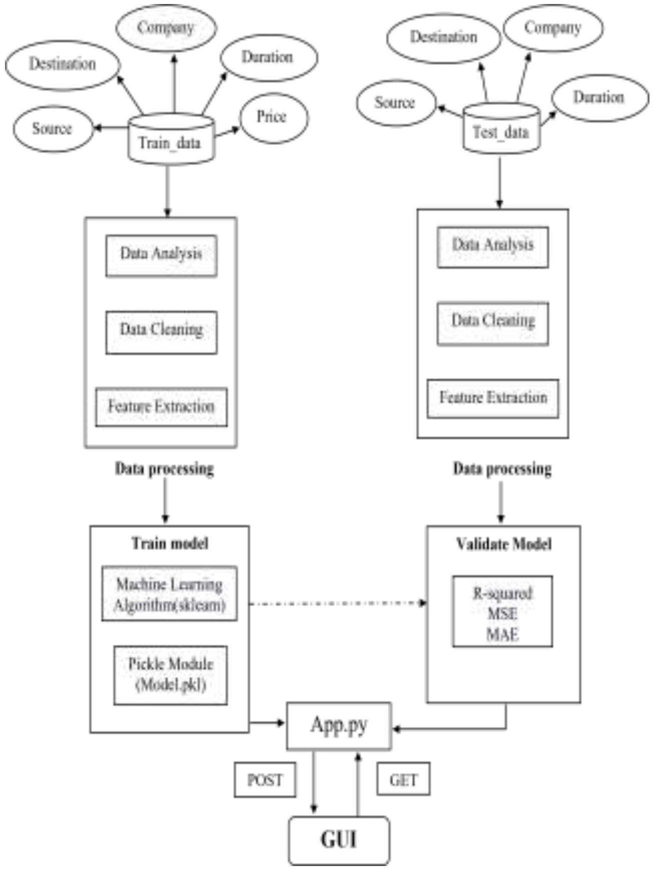

Fig. 2. System Architecture Diagram

There are also different cross-validation techniques such as gridsearchCV and randomizedsearchCV which will be used for improving the accuracy of the model. Parameters of the models such as number of trees in random forest or max depth of decision tree can be changed using this technique which will help us in further enhancement of the accuracy.

The last three steps of the life cycle model are involved in the deployment of the trained machine learning model. Therefore, after getting the model with the best accuracy we store that model in a file using pickle module. The back-end of the application will be created using Flask Framework where API end-points such and GET and POST will be created to perform operations related to fetching and displaying data on the front-end of the application.

The front-end of the application will be created using the bootstrap framework where user will have the functionality of entering their flight data. This data will be sent to the back-end service where the model will predict the output according to the provided data. The predicted value is sent to the front-end and displayed.

IV.Conclusion

A proper implementation of this project can result in saving money of inexperienced people by providing them the information related to trends that flight prices follow and also give them a predicted value of the price which they use to decide whether to book ticket now or later. In conclusion this type of service can be implemented with good accuracy of prediction. As the predicted value is not fully accurate there is huge scope for improvement of these kind of service.

V. Future Scope

Currently, there are many fields where prediction-based services are used such as stock price predictor tools used by stock brokers and service like Zestimate which gives the estimated value of house prices. Therefore, there is requirement for service like this in the aviation industry which can help the customers in booking tickets. There are many researches works that have been done on this using various techniques and more research is needed to improve the accuracy of the prediction by using different algorithms. More accurate data with better features can be also be used to get more accurate results.

Acknowledgment

We feel great pleasure in submitting this project paper on Flight fare prediction. We wish to express true sense of gratitude towards our project guide, Dr. V. V. Kimbhahune who at every stage in the study of this project, contributed his valuable guidance and helped to solve every and every problem. Our great obligation would remain due towards Dr. V. V. Kimbhahune who was a constant inspiration during our project. He provided with an opportunity to undertake the project at Smt. Kashibai Navale College of Engineering, Pune. I feel highly indebted to them who provided us with all our project requirements, and did much beyond our expectations to bring out the best in us. I sincerely thank our respected Head of Department Dr. P. N. Mahalle, he proved to be a constant motivation for the knowledge acquisition and moral support during our course curriculum.

References

[1] K. Tziridis T. Kalampokas G.Papakostas and K. Diamantaras "Airfare price prediction using machine larning techniques" in European Signal Processing Conference (EUSIPCO), DOI: 10.23919/EUSIPCO .2017.8081365L. Li Y. Chen and Z. Li" Yawning detection for monitoring driver fatigue based on two cameras" Proc. 12th Int. IEEE Conf. Intell. Transp. Syst. pp. 1-6 Oct. 2009.

[2] William Groves and Maria Gini "An agent for optimizing airline ticket purchasing" in proceedings of the 2013 international conference on autonomous agents and multi-agent systems.

[3] J. Santos Dominguez-Menchero, Javier Rivera and Emilio Torres-Manzanera "Optimal purchase timing in the airline market".

[4] Supriya Rajankar, Neha sakhrakar and Omprakash rajankar "Flight fare prediction using machine learning algorithms" International journal of Engineering Research and Technology (IJERT) June 2019.

[5] Tianyi wang, samira Pouyanfar, haiman Tian and Yudong Tao "A Framework for airline price prediction: A machine learning approach"

[6] T. Janssen "A linear quantile mixed regression model for prediction of airline ticket prices"

[7] Wohlfarth, T.clemencon, S.Roueff "A Dat mining approach to travel price forecasting" 10th international conference on machine learning Honolulu 2011.

[8] medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e article on performance metrics

[9] www.keboola.com/blog/random-forest-regression article on random forest

[10] https://towardsdatascience.com/machine-learning-basics-decision-tree-regression-1d73ea003fda article on decision tree regression