Project – Flight Price Prediction

Submitted by:

Chaganti Sai Mahathi

ACKNOWLEDGMENT

[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 TorresManzanera "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-ofdetermination-adjusted-r-squared-which-metric-is-bettercd0326a5697e article on performance metrics

[9] www.keboola.com/blog/random-forest-regression article on random forest [10] https://towardsdatascience.com/machine-learning-basics-decisiontree-regression-1d73ea003fda article on decision tree regression

[11]https://www.ijraset.com/research-paper/aircraft-ticket-price-prediction-using-machine-learning

[12] https://medium.com/analytics-vidhya/regression-flight-price-prediction-6771fc4d1fb3

INTRODUCTION

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.

There are some frequent passengers for the airlines, these people can be able to know when the airfare prices will be low and when it goes to high. But there are many people who doesn't know the fluctuations. Airline increase the price of the ticket for their Revenue management. The goal of the career will be to bult it's income, while the person who wants to purchase a ticket will look for the low cost. One of the report says

that India is the third biggest avionics showcase of world in 2020. Here we collect the data from different websites and apply machine learning models. We use different Regression methods to predict fare price at a given time.

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:

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 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: Linear regression is the most common form of regression. Multi linear regression is used to explain the relationship between one continuous dependant variable and two or more explanatory variables. Linear regression has several assumptions such as; Regression residuals must be normally distributed, residuals are homoscedastic and approximately rectangular-shaped, Absence of multicollinearity is assumed in the model.

K Nearest Neighbour Regression: K-Neighbhor's Regressor is a non-parametric lazy learning algorithm. KNN uses a similarity measure such as the Euclidian distance to find a heuristically optimal number of K nearest neighbours based on Root mean square error to label new observations. The output is the property value for the object. This value is the average of the values of its k nearest neighbours. KNN is sensitive to high dimensional data and outliers.

Random Forest Regression: Random Forest is a Supervised Machine

Learning Algorithm that is used widely in Classification and Regression

problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression.

Decision Tree Regression: Decision-tree algorithm falls under the category of supervised learning algorithms. It works for both continuous as well as categorical output variables.

The branches/edges represent the result of the node and the nodes have either:

1. Conditions [Decision Nodes]
2. Result [End Nodes]

Here we use different techniques such as Data Analysis, Data mining and Data Visualisation, Data Pre-processing and model building is done for the prediction.

Steps For Data Collection, Processing, EDA and Modelling the data:

Data collection and Analysis:

Data is collected by using various web scraping techniques using selenium from different websites and creating the data frame for all the data collected. The collected data has many features which gives different information about the fare ticket price of the flight.

FEATURES:

Airline: The name of the airline.

Date_of_Journey: The date of the journey

Source: The source from which the service begins.

Destination: The destination where the service ends.

Route: The route taken by the flight to reach the destination.

Dep_Time: The time when the journey starts from the source.

Arrival_Time: Time of arrival at the destination.

Duration: Total duration of the flight.

Total Stops: Total stops between the source and destination.

Additional_Info: Additional information about the flight

Price: The price of the ticket.

A) Removing null values:

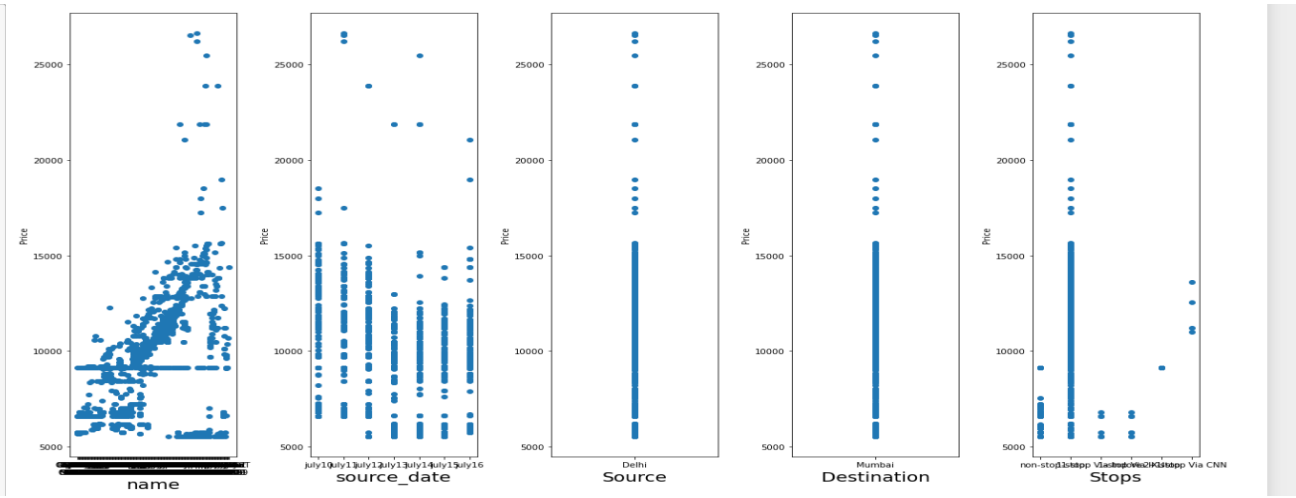
1. After looking in to the data we got the ticket price label with some null values which are removed by replace the comma with space and changing them to float and using np.nan method.

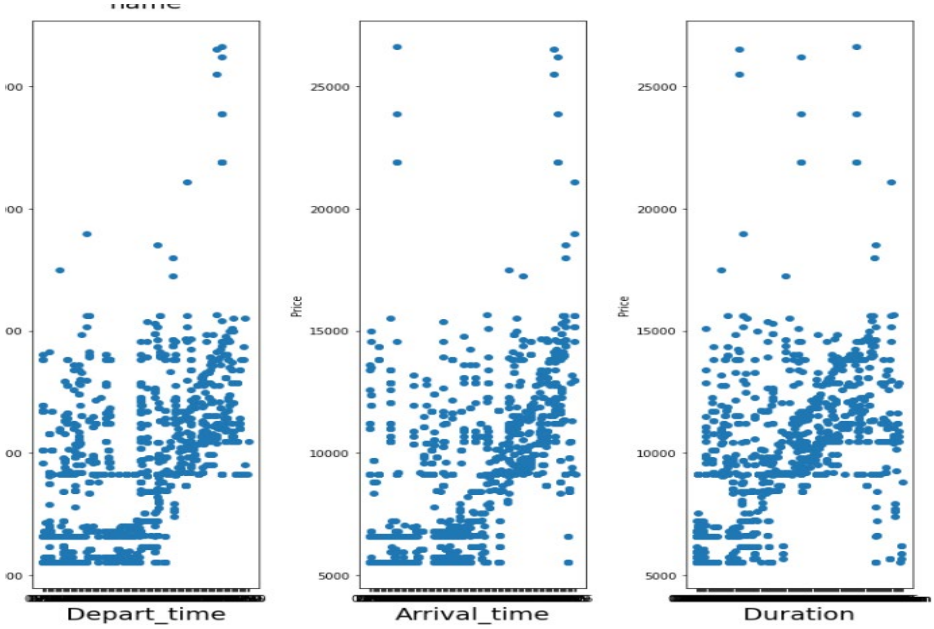
```
df.isnull().sum()

name              0
source_date       0
Source            0
Destination       0
Stops             0
Depart_time       0
Arrival_time      0
Duration          0
Ticket_Price      0
dtype: int64
```

B) Data Visualisation and analysis:

Here we are using scatter plot to visualise all the input features with respect to the Price of the ticket variable.





C) Encoding the data:

TO convert all object type data to numerical of both training and testing data we use Ordinal encoder.And also removing all the unwanted columns.

```
df
```

	name	source_date	Source	Destination	Stops	Depart_time	Arrival_time	Duration	Ticket_Price
0	87.0	0.0	0.0	0.0	5.0	20.0	25.0	2.0	6583.0
1	80.0	0.0	0.0	0.0	5.0	27.0	34.0	3.0	6583.0
2	158.0	0.0	0.0	0.0	5.0	108.0	108.0	3.0	6583.0
3	92.0	0.0	0.0	0.0	5.0	2.0	8.0	4.0	6583.0
4	157.0	0.0	0.0	0.0	5.0	24.0	30.0	4.0	6583.0
...	...	...	...	...	...	...	...	...	...
1654	39.0	6.0	0.0	0.0	0.0	74.0	120.0	51.0	14828.0
1655	17.0	6.0	0.0	0.0	0.0	45.0	120.0	89.0	14828.0
1656	6.0	6.0	0.0	0.0	0.0	36.0	74.0	37.0	15427.0
1657	4.0	6.0	0.0	0.0	0.0	70.0	97.0	16.0	18975.0
1658	22.0	6.0	0.0	0.0	0.0	46.0	97.0	52.0	21075.0

1659 rows × 9 columns

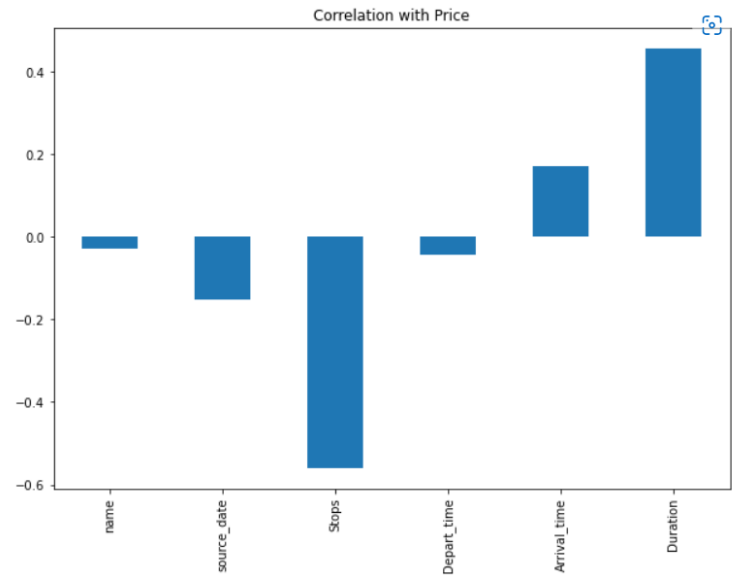
```
df1=df.drop(columns=['Source','Destination'],axis=True)# dropping the unnecessary columns
```

D}Data Pre-Processing:

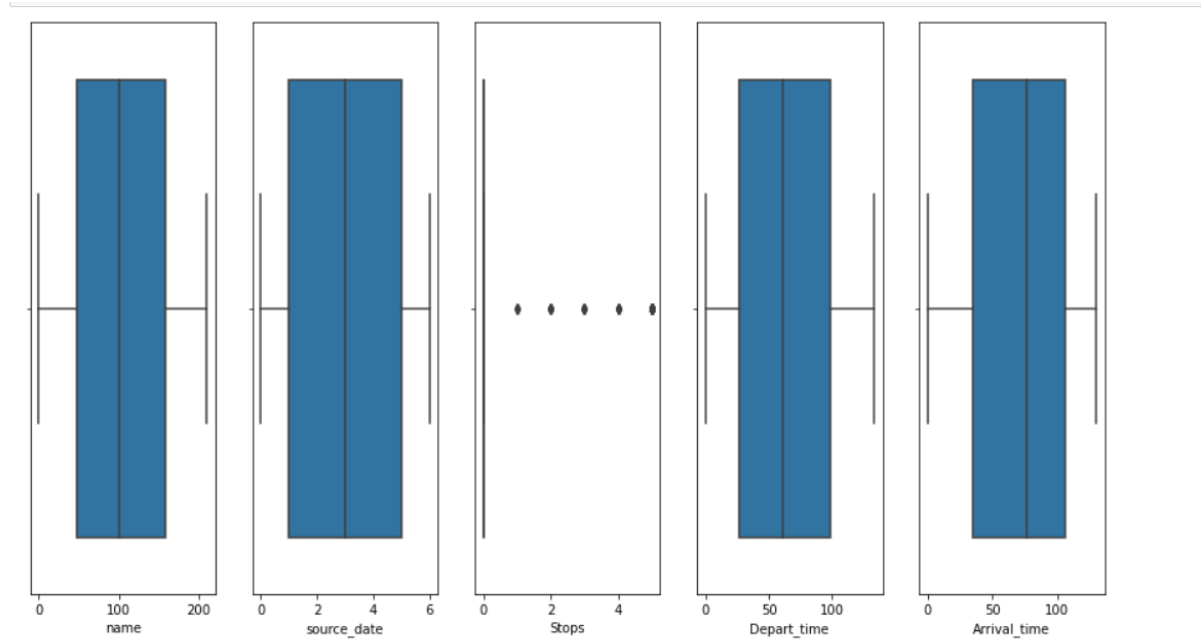
The first step is to Drop all the unwanted column present in the dataset.

Now, since the data sets are ready to process, check the correlation of the training dataset:

```
df1.drop('Ticket_Price',axis=1).corrwith(df1.Ticket_Price).plot(kind='bar',grid=False,figsize=(10,7),title='Correlation with Pri
plt.show()
```



Now lets also check the skewness and outliers if exists lets try to remove using some commonly used techniques.



Removing outliers and skewness is a very difficult process in this case we can use zscore technique with a data loss od 0.8 percent.

Percentage data loss:

```
loss_percent=(1659-1645)/1659*100
loss_percent
```

```
0.8438818565400843
```

Now once the best features are declared we cannot say that we avoided any outliers or skewness completely but now by considering these features let's create a new training and testing data frames.

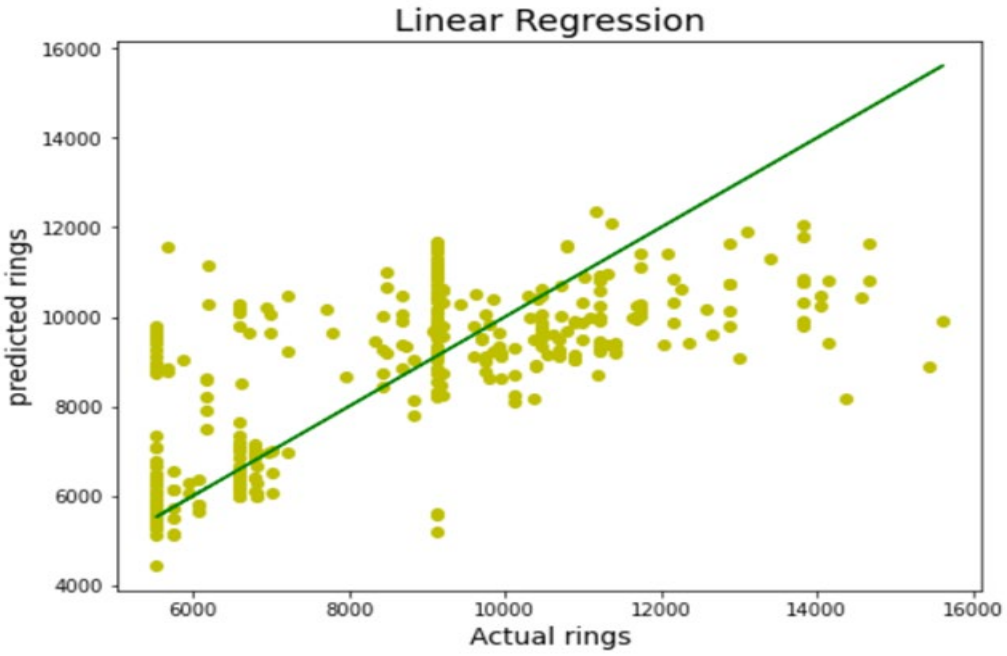
Model Building:

Now, once the data is pre processed and ready for the building, we first Train the data using various models and then use this trained data for testing or predicting the testing model.

First split the feature variables as 'x' data frame and 'y' as target. Then, Transform the 'x' variable using power transformation to remove any skewness present in the data and then normalise the data using Standard Scaler.

Next create a train test split of random state given as 41and test size as 0.20 and then apply linear regression model to test the data the r2 score of the predicted d ata is given as **0.49.**

Linear Regression Curve:



Once the model is linearly regressed then use lasso regularisation technique to g et the maximum best output-**0.4991**
Using the best cross validation score we get the output for this model as

```
0.4159
```

Now we use other Ensemble approaches and further using hyper parameter tuning finding the best parameters gets the better results of

Random Forest Regression:

```
R2 Score:  80.04798220618258
Cross validation Score:  57.00859142355505
```

Decision Tree Regression:

```
R2 Score:  69.79148086298666
Cross validation Score:  -15.122281155692077
```

K-Neighbors Regression:

```
R2 Score:  62.32234763894346
Cross validation Score:  54.39717503179858
```

SVR model:

```
R2 Score:  32.84326348876589
Cross validation Score:  30.728755019352526
```

Out of all models KNN regression is having R2 Score: 62.32234and Cross validation Score: 54.397175. So, this model is finalised and further saved.

Saving the model and Loading the data is the further most step, now is that my model is trained it goes to testing phase

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.

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.