APPLYING MACHINE LEARNING TO FLIGHT PRICE PREDICTION

Student 1, Student 2, Student 3

[1, 2, 3] **Students, Computer Science and Engineering department, <YOUR INSTITUTE NAME>, <PLACE>**

----------***----------

ABSTRACT

The recent global situations had a huge impact on the aviation sector due to many reasons. This impact has two category people, the first is business perspective and the second is the customer's perspective. As safety is the major reason for such impact on the aviation sector, the governments around the world amended different rules to their respective airlines companies. These restrictions had made the availability of the flights and their attendee capacity less. Taking all these factors in consideration the cost of the flight tickets has increased and vary from one place to the other. Booking a flight ticket has split into two, one is the online and the other is the offline bookings. Both these have their respective criteria for cost of the ticket, one such example is the server load and the number of booking requests. In this machine learning implementation, we will see various factors that impact the cost of the flight ticket and predict the appropriate price of the ticket.

1. Introduction

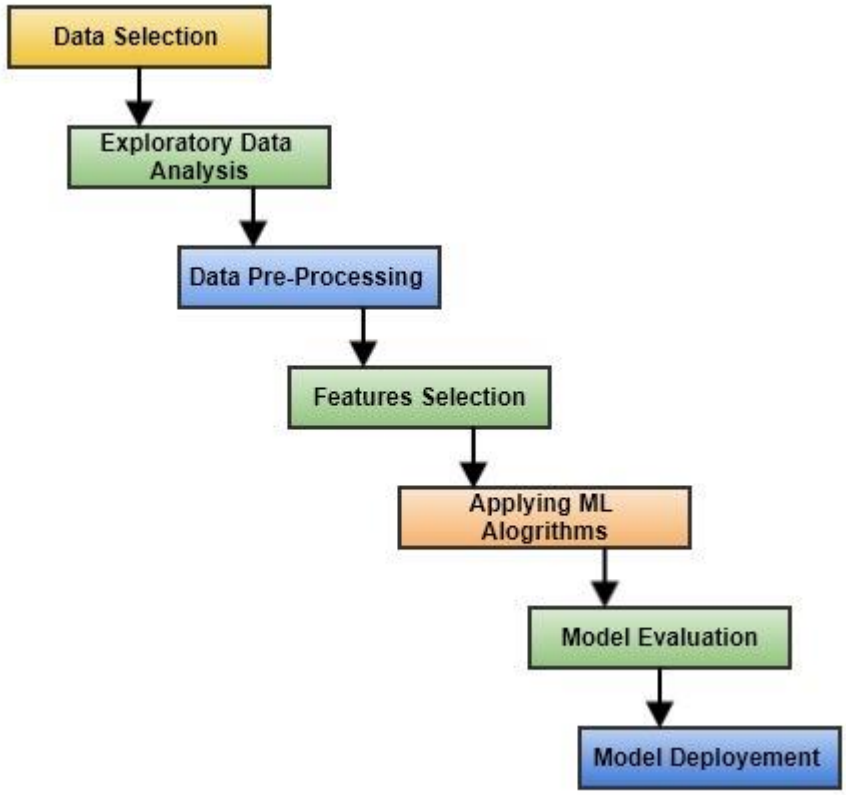
This project aims to develop an application which can predict the flight prices for various flights using different machine learning techniques. The user will get the expected values and with its reference the user can plan to book their tickets suitably. At this time, carrier ticket costs can shift powerfully and fundamentally for an identical flight, in any event, for accessible seats inside the identical cabin. Clients are attempting to urge the foremost minimal cost while Airlines companies try to stay their general income as high as could reasonably be expected and boost their benefit

Airlines utilize different computational methods to extend their income, as an example, demand forecast and value segregation. The proposed system can help save immeasurable rupees of shoppers by proving them the knowledge to book tickets at the correct time. Parameters on which fares are calculated

- Airline
- Date of Journey
- Source
- Destination
- Departure Time
- Duration
- Total Stops
- Weekday/Weekend

2. Proposed Methodology

For this project, we've implemented the machine learning life cycle to make a basic web application which is able to predict the flight fare by applying machine learning algorithms on historical flight data using some python libraries like Pandas, NumPy, Matplotlib, seaborn, and sklearn. Below image shows the number of steps that we followed from the life cycle



Data selection is that the initial step where historical data of flight is assembles for the model to predict prices. Our dataset consists of quite 10,000+ records of information associated with flights and costs. A number of the features of the dataset are source, destination, departure date, point, and number of stops, point in time, prices. Within the exploratory data analysis step, we cleaned the dataset by removing the duplicate values and null values. If the null values aren't removed, the accuracy of the model are affected. Next step is data pre-processing where we\ observed that almost

all of the information was present in string format. Data from each feature is extracted i.e., day and month is extracted from date of journey in integer format, hours and minutes is extracted from time of departure. Features like 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 wont to convert categorical data into the integer data

Feature selection step is involved in selecting important features that are more correlated to the value. So, some features to be selected and passed to the group of models. Random forest is an ensemble learning method that basically uses group of decision trees as group of models. Random amount of knowledge is passed to decision trees and every decision tree predicts values in line with the dataset given to that. From the predictions made by the choice trees the features like extra information and route which are unnecessary features which can affect the accuracy of the model and so, they have to be removed before getting our model ready for prediction

3. Technical Requirements

There are no hardware requirements required for using this application, the user must have an interactive device which has access to the internet and must have the basic understanding of providing the input. And for the backend part the server must run all the software that is required for the processing the provided data and to display the results.

2.1 Tools Used

- Python 3.7 is used as the programming language and frame works like Numpy, pandas, SK-Learn and other modules for building the model.
- Jupyter Notebook is used as IDE.
- For visualizations Seaborn and parts of Matplotlib are being used.
- GitHub is used for version control.

4. Data Requirements

The data requirement is completely based on the problem statement. And the data set is available on the Kaggle in the form of excel sheet (.xlsx). As the main theme of the project is to get the experience of real time problems, we are again importing the data into the Cassandra data base and exporting it into csv format.

3.1 Constraints

The flight fare prediction answer should be user friendly, as automatic as attainable and also the user should not be needed to understand any of the operations.

3.2 Assumptions

The most objective of the project is to implement the utilization cases for the new dataset that the user provides through the programme. Machine learning

Model is employed to process the on top of a computer file. It's additionally assumed that each one aspect of this project has the flexibility to figure along within the approach the designer is expecting.

3.3 Data Gathering from Main Source

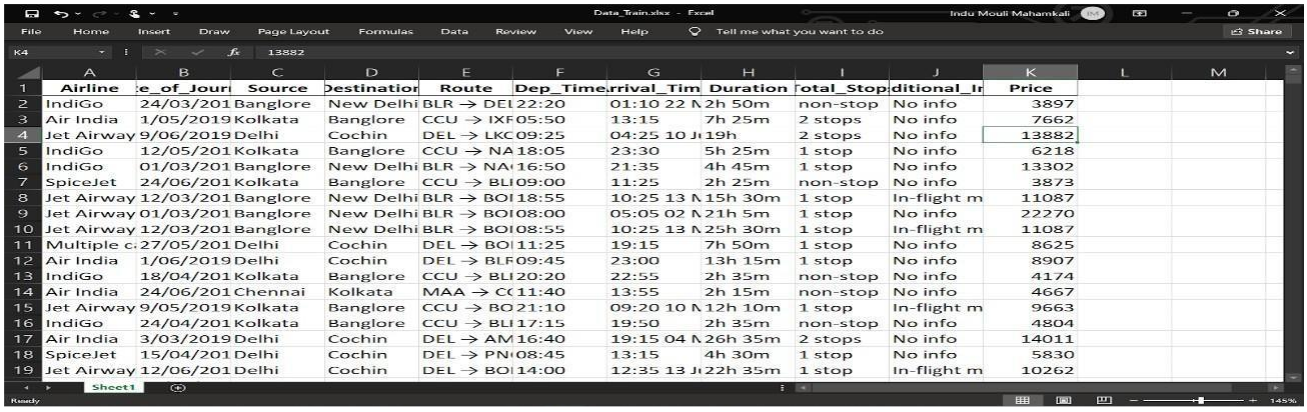
The data for the current project is being gathered from Kaggle dataset, the link to the data is: https://www.kaggle.com/nikhilmittal/flight-fare-prediction-mh

5. Data Description

We have 2 datasets here: training set and test set. The training set contains the features, along with the prices of the flights. It contains value. Following is the description of features available in the dataset 10683 records, 10 input features and 1 output column 'Price'. The test set contains 2671 records and 10 input features. The output 'Price' column needs to be predicted in this set.

We will use Regression techniques here, since the predicted output will be a continuous

There are about 10k+ records of flight information such as airlines, data of journey, source, Destination, departure time, arrival time, duration, total stops, additional information, and price. A Glance of the dataset is shown below:



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Airline	e_of_Jouri	Source	Destinatior	Route	Dep_Time	rrival_Tim	Duration	otal_Stop	ditional_In	Price		
2	IndiGo	24/03/201	Banglore	New Delhi	BLR → DEL	22:20	01:10 22 M	2h 50m	non-stop	No info	3897		
3	Air India	1/05/2019	Kolkata	Banglore	CCU → IXR	05:50	13:15	7h 25m	2 stops	No info	7662		
4	Jet Airway	9/06/2019	Delhi	Cochin	DEL → LKO	09:25	04:25 10 Ju	19h	2 stops	No info	13882		
5	IndiGo	12/05/201	Kolkata	Banglore	CCU → NA	18:05	23:30	5h 25m	1 stop	No info	6218		
6	IndiGo	01/03/201	Banglore	New Delhi	BLR → NA	16:50	21:35	4h 45m	1 stop	No info	13302		
7	SpiceJet	24/06/201	Kolkata	Banglore	CCU → BLI	09:00	11:25	2h 25m	non-stop	No info	3873		
8	Jet Airway	12/03/201	Banglore	New Delhi	BLR → BOI	18:55	10:25 13 M	15h 30m	1 stop	In-flight m	11087		
9	Jet Airway	01/03/201	Banglore	New Delhi	BLR → BOI	08:00	05:05 02 M	21h 5m	1 stop	No info	22270		
10	Jet Airway	12/03/201	Banglore	New Delhi	BLR → BOI	08:55	10:25 13 M	25h 30m	1 stop	In-flight m	11087		
11	Multiple c	27/05/201	Delhi	Cochin	DEL → BO	11:25	19:15	7h 50m	1 stop	No info	8625		
12	Air India	1/06/2019	Delhi	Cochin	DEL → BLF	09:45	23:00	13h 15m	1 stop	No info	8907		
13	IndiGo	18/04/201	Kolkata	Banglore	CCU → BLI	20:20	22:55	2h 35m	non-stop	No info	4174		
14	Air India	24/06/201	Chennai	Kolkata	MAA → CC	11:40	13:55	2h 15m	non-stop	No info	4667		
15	Jet Airway	9/05/2019	Kolkata	Banglore	CCU → BO	21:10	09:20 10 M	12h 10m	1 stop	In-flight m	9663		
16	IndiGo	24/04/201	Kolkata	Banglore	CCU → BLI	17:15	19:50	2h 35m	non-stop	No info	4804		
17	Air India	3/03/2019	Delhi	Cochin	DEL → AM	16:40	19:15 04 M	26h 35m	2 stops	No info	14011		
18	SpiceJet	15/04/201	Delhi	Cochin	DEL → PN	08:45	13:15	4h 30m	1 stop	No info	5830		
19	Jet Airway	12/06/201	Delhi	Cochin	DEL → BO	14:00	12:35 13 Ju	22h 35m	1 stop	In-flight m	10262		

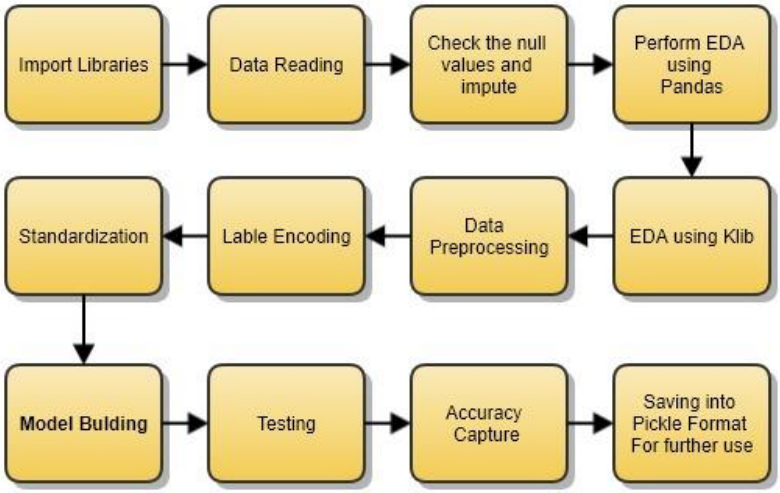
Following is the description of features available in the dataset

- 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

6. Data Pre-Processing

- The data types are being checked and found only the price column is of type Integer.
- Checked for null values as there are few null values, those rows are dropped.
- Hours and Minutes is extracted from departure time
- Extracted the components from Date Column like Day, Month, and Year.
- Type casted all the required columns into the date time format.
- Performed one-hot encoding for the required categorical columns.
- Applied Scaling techniques on required data
- Performed Label Encoding and hot-Encoding to convert Categorical values to Model Identifiable values and, the info is prepared for passing to the machine Learning formula.

7. Design Flow



6.1 Logging

Each step is being logged within the system that runs internally, that shows the date time and therefore the process that has been performed, work is completed in several layers as information, DEBUG, ERROR, WARNINGS. This provides the US the perception of the logged info.

6.2 Error Handling

Once a slip has occurred, the reason is logged in its several log files, in order that the developer will rectify the error.

8. Modelling

The pre-processed data is then visualized and all the required insights are being drawn.

Although from the drawn insights, the data is randomly spread but still modelling is performed with Different machine learning algorithms to make sure we cover all the possibilities. And finally, as Expected random forest regression performed well and further hyper parameter tuning is done to increase the model's accuracy.

After selecting the features which are more correlated to price the next step involves applying machine algorithms 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 relationships between dependent and independent variables. The machine learning algorithms that we will be using in our project are:

1. Linear Regression
2. Decision Tree
3. Random Forest

7.1 Linear Regression

In simple linear regression there's only 1 independent and one dependent feature but as our dataset consists of the many independent features on which the worth may rely on, we are going to be using multiple linear regression which estimates relationship between two or more independent variables and one dependent variable. The multiple linear regression models are represented by:

$$y = m_1x_1 + m_2x_2 + \cdots + m_nx_n + C$$

Where, y = the predicted value of the dependent variable

X_n= the independent variables

m= independent variables coefficients

C= y-intercept x

7.2 Decision Tree

There are basically of two type's Decision tree i.e., classification and regression tree where classification is employed for categorical values and regression is employed for continuous values. Decision tree chooses independent variable from dataset as decision nodes for decision making. It divides the entire dataset in several sub-section and when test data is passed to the model the output is determined by checking the section to which the info point belongs to. And to whichever section the info point belongs to the choice tree will give output because the average value of all the information points within the sub-section

7.3 Random Forest

Random Forest is an ensemble learning technique where training model uses multiple learning algorithms then combine individual results to urge 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 expected values if considered because the output of the random forest model

9. Performance analysis

9.1 Reusability

Elements of the code written are accustomed to different applications and therefore the rest is changed and reused

9.2 Application Compatibility

The various parts for this project are exploitation python as an associate interface between them. Every element can have its own tasks to perform, and it's the work of the python to make sure the transfer of data.

9.3 Resource Utilization

Once any task is performed, it'll doubtless use all the process power offered till that performance is finished

9.4 Deployment

The model can be deployed in Heroku or any other cloud instance such as EC2 or GCP Compute/App Engine.

10.Performance Metrics

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

9.1 MAE (Mean Absolute Error)

Mean Absolute Error is basically the sum of average of the

Absolute difference between the expected and actual values.

MAE = 1/n [∑(y-ý)]

y = actual output values,

ý = predicted output values

n = Total number of data points

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

9.2 MSE (Mean Square Error)

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

$$MSE = 1/n\,[\sum (y - y')^2]$$

y=actual output values

y'=predicted output values

n = Total number of data points

Lower the value of MSE better the performance of the mode

9.3 RMSE (Root Mean Square Error)

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

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

y=actual output values

y'=predicted output values

n = Total number of data points

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

11.RESULTS

We had used many algorithms for prepared our ML model i.e., Linear Regression, Decision tree, Random Forest. From all of these, Random Forest gives us the most accurate Predictions. Models' performance using a few metrics are given below:

A. RANDOM FOREST:
   - R2 SCORE : 0.8598264560438941
   - MAE : 1164.5042333744152
   - MSE : 4043823.6841106885
   - RMSE : 2010.3049101804945

By observing all of the performance metrics here the Random Forest will give us accurate and better results. So, we use Random Forest model for the deployment.

CONCLUSION

This project can result in saving money of inexperienced people by providing them the information related to trends of the flight prices and also give them a predicted value of the price which they use to decide whether to book ticket now or later. On working with different models, it was found out that Random Forest algorithm gives the highest accuracy in predicting the output.

REFERENCES

[1] K. Tziridis T. Kalampokas G.Papakostas and K. Diamantaras "Airfare price prediction using machine learning 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 multiagent systems.

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

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

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

[6] medium.com/analytics-vidhya/mae-mse-rmsecoefficient-ofdetermination-adjusted-r-squared-whichmetric-is-bettercd0326a5697e article on performance metrics