Flight Fare Prediction Using Machine Learning

Article · May 2024

CITATION
1

READS
4,457

3 authors, including:

Tjcse Dheemansh Publication Hub
The Journal of Computational Science and Engineering
31 PUBLICATIONS **16** CITATIONS

SEE PROFILE

The Journal of Computational Science and Engineering.
ISSN: 2583-9055

Flight Fare Prediction Using Machine Learning

Shikha Gupta[1], Nishi Gupta[2]

[1]Department of Computer Science and Engineering,
Maharaja Agrasen Institute of Technology, Rohini, Delhi, India
[2]Department of Computer Science and Engineering,
The NorthCap University, Gurugram, Haryana, India
Shikha.gpt1@gmail.com[1], nishigupta99@gmail.com[2]

Keyword:
Flight Price Estimation, Travel Cost Analysis, Historical Price Trends, Airfare Forecasting

ABSTRACT

The "Flight Fare Prediction" project aims to develop an advanced predictive model leveraging machine learning algorithms to estimate and forecast airfare prices accurately. The unpredictable and dynamic nature of flight ticket pricing poses a significant challenge for travelers in planning and budgeting for their trips. This research project seeks to alleviate this challenge by harnessing the power of machine learning to provide reliable and real-time predictions of flight fares. The study initiates by meticulously examining the myriad factors impacting flight fares, encompassing departure and arrival locations, booking timings, seasonal variations, airline preferences, and historical pricing trends. Through meticulous data collection and preprocessing, pertinent features are identified and subjected to a thorough analysis to discern their influence on ticket prices.

Corresponding Author: Email: nishigupta99@gmail.com[2]

INTRODUCTION

The aviation industry, characterized by its dynamic and often unpredictable nature, experiences frequent fluctuations in airfare prices influenced by a myriad of factors. Travelers, seeking cost-effective and efficient means of transportation, are confronted with the challenge of navigating through the complexities of fare variations. Simultaneously, airline companies strive to optimize revenue through strategic pricing strategies, making airfare prediction a critical aspect of modern aviation. Traditional approaches to airfare pricing have often relied on historical data and manual analysis, proving insufficient in capturing the intricate patterns and nuances of the evolving market. The integration of machine learning algorithms, including regression models, time series analysis, and ensemble learning, empowers the FFPS to discern complex relationships within vast datasets. This enables the system to adapt to changing market dynamics, providing real time and accurate predictions for a diverse range of flight routes. The significance of this research extends beyond the realm of operational efficiency. By offering travelers reliable insights into future airfare trends, the FFPS contributes to a more consumer-friendly aviation landscape. Furthermore, airline companies can benefit from improved revenue management, aligning pricing strategies with demand fluctuations and optimizing yield.

In the contemporary landscape of air travel, where dynamic market conditions and diverse influencing factors contribute to the volatility of airfare prices, the need for an accurate

and adaptive Flight Fare Prediction System (FFPS) is paramount. This research endeavors to introduce a sophisticated predictive model leveraging a combination of decision tree algorithms, k Nearest Neighbors (KNN), and linear regression to enhance the precision and versatility of airfare estimations.

The inclusion of the k-Nearest Neighbors algorithm adds a dynamic and adaptive dimension to our FFPS. KNN relies on the proximity of data points in a multidimensional space, allowing the system to consider the similarities between a target flight and its neighboring instances. This technique proves invaluable in capturing localized patterns and trends, particularly relevant when predicting airfares for specific routes or during distinct time periods.

Linear regression, a widely employed algorithm in predictive modeling, is seamlessly integrated into our system to capture linear relationships between independent variables and airfare prices. By understanding the linear dependencies within the data, the FFPS enhances its predictive accuracy, especially in scenarios where factors exhibit a clear and direct impact on pricing dynamics. This comprehensive fusion of decision tree algorithms, k-Nearest Neighbors, and linear regression forms the backbone of our research, aimed at developing a robust FFPS. The amalgamation of these algorithms seeks to overcome the limitations of individual models, providing a more holistic and accurate representation of the intricate interplay of factors affecting airfare prices.

The motivation behind the development of a Flight Fare Prediction system stems from the desire to empower travelers with the ability to make informed decisions, optimize their travel budgets, and navigate the intricacies of the air travel market. By harnessing the advancements in machine learning, this project aims to revolutionize the way individuals plan and book their flights, offering a data-driven approach to predicting airfare prices with increased accuracy.

LITERATURE SURVEY

Flight fare prediction is a critical aspect of travel planning, alleviating uncertainties related to fluctuating ticket prices. Various methodologies have been explored in the literature to develop accurate prediction models. This literature review provides a comprehensive overview and analysis of existing research endeavors in the domain of flight fare prediction using machine learning techniques. An overview and findings of some of the recent researches are presented here.

In 2017, K. Tziridis, T. Kalampokas, G. Papakostas, and K. Diamantaras presented a study on airfare price prediction titled "Airfare price prediction using machine learning techniques" at EUSIPCO. The research utilized a dataset of 1814 Aegean Airlines flight records, employing methods like MLP, Regression Neural Network, ELM, and Random Forest Regression Tree. The study explored various models, observing optimal results with the Bagging regression tree. [8]

In 2013, William Groves and Maria Gini introduced an agent for optimizing airline ticket purchasing. The study focused on developing a model using partial least square regression for optimizing purchase time on behalf of consumers. Feature selection strategies such as Feature Extraction, Lagged Feature Computation, Regression Model Construction, and Optimal Model Selection were employed. Through trials, the study explored the actual costs of implementing prediction models, with the lag scheme technique proving effective for various machine learning algorithms. PLS regression emerged as the most effective method, attributed

to its natural resistance to irrelevant and collinear factors. [12]

In 2014, J. Santos Dominguez-Menchero, Javier Rivera, and Emilio Torres Manzanera conducted a study titled "Optimal purchase timing in the airline market." The authors explored the general price behavior of airlines and developed a technique for analyzing various routes and/or carriers. The research aimed to provide clients with information to choose the optimal time for ticket purchase, considering both cost savings and time constraints. The study highlighted the effectiveness of non-parametric isotonic regression approaches compared to 5 traditional parametric methods. The findings included insights into the margin of time customers could postpone their purchase without a significant price rise, the economic loss for each day of delay, and whether waiting until the last day was preferable for making a purchase. [DomínguezMenchero, J. S., Rivera, J., & Torres-Manzanera, E. (2014). Optimal purchase timing in the airline market. Journal of Air Transport Management, 40, 137-143.] [13]

In June 2019, Supriya Rajankar, Neha Sakhrakar, and Omprakash Rajankar conducted a study titled "Flight fare prediction using machine learning algorithms", published in the International Journal of Engineering Research and Technology (IJERT). The research focused on predicting flight fares using machine learning algorithms with a limited dataset of flights between Delhi and Bombay. Various algorithms, such as K-nearest neighbors (KNN), linear regression, and support vector machine (SVM), were employed to obtain diverse results and conduct a comprehensive study. [Supriya Rajankar, Neha Sakhrakar, & Omprakash Rajankar. (June 2019). Flight fare prediction using machine learning algorithms. International Journal of Engineering Research and Technology (IJERT).] [14]

In 2015, Michael I. Jordan and Thomas M. Mitchell authored the paper titled "Machine learning: Trends, perspectives, and prospects," published in the journal Science (Vol. 349, No. 6245, pp. 255-260). The paper explores the trends, perspectives, and prospects in the field of machine learning, providing valuable insights into the evolving landscape of this discipline. [Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.] [1]

In April 2023, N. S. S. V. S. Rao and S. J. J. Thangaraj presented a paper titled "Flight Ticket Prediction using Random Forest Regressor Compared with Decision Tree Regressor" at the Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM). The study compares the predictive performance of Random Forest Regressor with Decision Tree Regressor for flight ticket prediction, providing valuable insights into the effectiveness of these algorithms in the context of airfare forecasting. [Rao, N. S. S. V. S., & Thangaraj, S. J. J. (2023, April). Flight Ticket Prediction using Random Forest Regressor Compared with Decision Tree Regressor. In 2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM) (pp. 1-5). IEEE.] [4]

In May 2023, a team of researchers including Joshitta S. M., Sunil M. P., A. Bodhankar, C. Sreedevi, and R. Khanna presented a paper titled "The Integration of Machine Learning Technique with the Existing System to Predict the Flight Prices" at the 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE). The study explores the integration of machine learning techniques into existing systems for predicting flight prices, contributing to the advancement of technologies in the field. [Joshitta, S. M., Sunil, M. P., Bodhankar, A., Sreedevi, C., & Khanna, R. (2023, May). The Integration of Machine Learning Technique with the Existing System to Predict the Flight Prices. In 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) (pp. 398-402). IEEE.] [11].

PROPOSED METHODOLOGY

A. Methodology

1. *Data Collection and Preparation:*
- Libraries: pandas
- Historical flight pricing data is gathered from multiple sources. The data is pre-processed using pandas and NumPy to handle missing values and inconsistencies.

2. *Analyzing Flight Timing:*
- Libraries: matplotlib, seaborn
- Data analysis techniques are applied to determine peak flight times, aiding in fare prediction accuracy.

3. *Analyzing Flight Duration vs. Price:*
- Libraries: matplotlib, seaborn
- Visualization and analysis of how flight duration influences flight fares.

4. *Data Analysis and Visualization:*
- Libraries: Pandas, Matplotlib, Seaborn
- Analyzing the impact of the number of stops on price using a scatter plot.
- Visualizing the distribution of prices across different airlines using a boxplot.
- Implementing one-hot encoding for categorical variables.
- Determining mean prices for each airline.

5. *Outlier Detection and Handling:*
- Libraries: Matplotlib, Seaborn
- Creating a custom function for plotting distribution and box plots to identify outliers.
- Utilizing the IQR (Interquartile Range) method to detect outliers in the 'Price' column.
- Replacing extreme values (greater than 35K) with the median of the 'Price' column.

6. *Feature Selection and Model Training:*
- Libraries Used: Scikit-learn
- Calculating mutual information between the target variable ('Price') and other features.
- Implementing Random Forest Regressor for training the machine learning model.
- Predicting prices for the test data and analyzing predictions.

7. *Model Evaluation and Prediction Analysis:*
- Libraries: Scikit-learn, Seaborn
- Defining a function to evaluate the model's performance metrics.
- Assessing the model's training score, predictions, R-squared score, MAE, MSE, RMSE, and MAPE.
- Visualizing the distribution of the difference between predicted and actual prices.

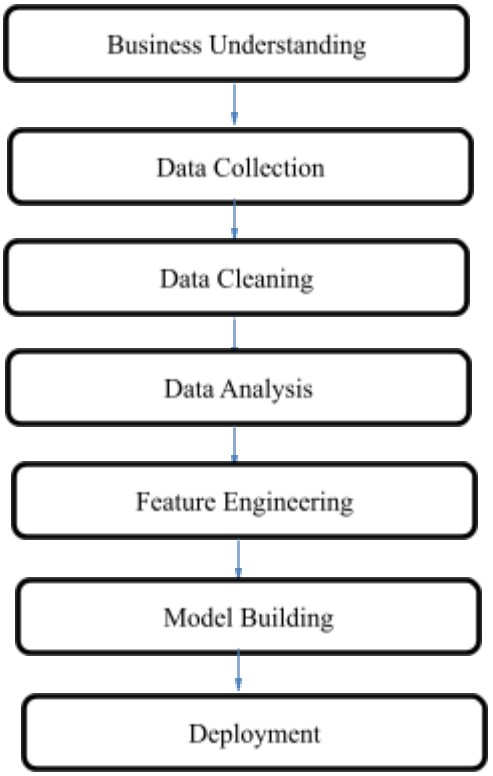

Figure 1: Flowchart for the proposed method

B. Algorithms

1. Random Forest Regressor:

The Random Forest Regressor is an ensemble learning algorithm that builds a multitude of decision trees during training and outputs the average prediction of the individual trees for regression tasks. In this project, the Random Forest Regressor is used to predict flight prices based on various features.

a) *r2 score (Coefficient of Determination):* The r2 score measures the proportion of the variance in the dependent variable (flight prices) that is predictable from the independent variables. An r2 score of 0.81 indicates that the model explains about 81% of the variability in flight prices, showcasing a good fit.
b) *MAE (Mean Absolute Error):* The MAE represents the average absolute difference between predicted and actual flight prices. With an MAE of 1179.82, the model's predictions are, on average, off by approximately 1179.82 INR.
c) *MSE (Mean Squared Error) and RMSE (Root Mean Squared Error):* These metrics quantify the average squared and square root of the differences between predicted and actual prices. The RMSE of 1932.77 indicates the average magnitude of errors in predicting flight prices.
d) *MAPE (Mean Absolute Percentage Error):* The MAPE measures the average percentage difference between predicted and actual prices. With a MAPE of 13.18%, the model's predictions have an average relative error of around 13.18%.

2. *Decision Tree Regressor:*

The Decision Tree Regressor is a decision tree-based algorithm for regression tasks. It works by recursively partitioning the data into subsets based on the values of the features.

a) *r2 score:* The Decision Tree Regressor achieves an r2 score of 0.67, indicating that it explains about 67% of the variability in flight prices. While a respectable score, it is lower than the Random Forest Regressor.
b) *MAE:* The MAE for the Decision Tree Regressor is 1426.62, suggesting slightly higher prediction errors compared to the Random Forest Regressor.
c) *MSE and RMSE:* The Decision Tree Regressor's MSE and RMSE are higher than those of the Random Forest Regressor, indicating a larger spread in prediction errors.
d) *MAPE:* The MAPE of 15.73% suggests a slightly higher relative error compared to the Random Forest Regressor.

The Random Forest Regressor generally outperforms the Decision Tree Regressor in terms of predictive accuracy. The ensemble nature of the random forest helps mitigate overfitting and provides a more robust model. However, both models demonstrate decent performance in predicting flight prices, with the random forest being the preferred choice due to its superior results. Further fine-tuning and optimization could potentially enhance the models' performance.

EXPERIMENTAL RESULTS

A. Data Collection and Preparation

In the initial phase of our project, we undertook the crucial task of data collection and preparation, pivotal for the success of our flight fare prediction model. The dataset, a comprehensive compilation of historical flight pricing data, was obtained from Kaggle—a renowned platform for datasets and data science competitions. Kaggle, a hub for data enthusiasts and practitioners, provided us with the dataset titled "Flight Fare Prediction This dataset encompassed various attributes such as departure and arrival locations, travel dates, airlines, and ticket prices, forming the foundation for our predictive model.

Table 1: Data Collection and preparation

	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info
2	IndiGo	24/03/2019	Banglore	New Delhi	BLR → DEL	22:20	01:10 22 Mar	2h 50m	non-stop	No info
3	Air India	1/05/2019	Kolkata	Banglore	CCU → IXR → BBI → BLR	05:50	13:15	7h 25m	2 stops	No info
4	Jet Airway	9/06/2019	Delhi	Cochin	DEL → LKO → BOM → COK	09:25	04:25 10 Jun	19h	2 stops	No info
5	IndiGo	12/05/2019	Kolkata	Banglore	CCU → NAG → BLR	18:05	23:30	5h 25m	1 stop	No info
6	IndiGo	01/03/2019	Banglore	New Delhi	BLR → NAG → DEL	16:50	21:35	4h 45m	1 stop	No info
7	SpiceJet	24/06/2019	Kolkata	Banglore	CCU → BLR	09:00	11:25	2h 25m	non-stop	No info
8	Jet Airway	12/03/2019	Banglore	New Delhi	BLR → BOM → DEL	18:55	10:25 13 Mar	15h 30m	1 stop	In-flight meal not included
9	Jet Airway	01/03/2019	Banglore	New Delhi	BLR → BOM → DEL	08:00	05:05 02 Mar	21h 5m	1 stop	No info
10	Jet Airway	12/03/2019	Banglore	New Delhi	BLR → BOM → DEL	08:55	10:25 13 Mar	25h 30m	1 stop	In-flight meal not included
11	Multiple ca	27/05/2019	Delhi	Cochin	DEL → BOM → COK	11:25	19:15	7h 50m	1 stop	No info
12	Air India	1/06/2019	Delhi	Cochin	DEL → BLR → COK	09:45	23:00	13h 15m	1 stop	No info
13	IndiGo	18/04/2019	Kolkata	Banglore	CCU → BLR	20:20	22:55	2h 35m	non-stop	No info
14	Air India	24/06/2019	Chennai	Kolkata	MAA → CCU	11:40	13:55	2h 15m	non-stop	No info
15	Jet Airway	9/05/2019	Kolkata	Banglore	CCU → BOM → BLR	21:10	09:20 10 May	12h 10m	1 stop	In-flight meal not included
16	IndiGo	24/04/2019	Kolkata	Banglore	CCU → BLR	17:15	19:50	2h 35m	non-stop	No info
17	Air India	3/03/2019	Delhi	Cochin	DEL → AMD → BOM → COK	16:40	19:15 04 Mar	26h 35m	2 stops	No info
18	SpiceJet	15/04/2019	Delhi	Cochin	DEL → PNQ → COK	08:45	13:15	4h 30m	1 stop	No info
19	Jet Airway	12/06/2019	Delhi	Cochin	DEL → BOM → COK	14:00	12:35 13 Jun	22h 35m	1 stop	In-flight meal not included
20	Air India	12/06/2019	Delhi	Cochin	DEL → CCU → BOM → COK	20:15	19:15 13 Jun	23h	2 stops	No info
21	Jet Airway	27/05/2019	Delhi	Cochin	DEL → BOM → COK	16:00	12:35 28 May	20h 35m	1 stop	In-flight meal not included
22	GoAir	6/03/2019	Delhi	Cochin	DEL → BOM → COK	14:10	19:20	5h 10m	1 stop	No info
23	Air India	21/03/2019	Banglore	New Delhi	BLR → COK → DEL	22:00	13:20 19 Mar	15h 20m	1 stop	No info
24	IndiGo	3/04/2019	Banglore	Delhi	BLR → DEL	04:00	06:50	2h 50m	non-stop	No info
25	IndiGo	1/05/2019	Banglore	Delhi	BLR → DEL	18:55	21:50	2h 55m	non-stop	No info
26	Jet Airway	6/05/2019	Kolkata	Banglore	CCU → BOM → BLR	18:55	08:15 07 May	13h 20m	1 stop	In-flight meal not included
27	Jet Airway	9/06/2019	Delhi	Cochin	DEL → IDR → BOM → COK	21:25	12:35 10 Jun	15h 10m	2 stops	No info
28	IndiGo	1/06/2019	Delhi	Cochin	DEL → LKO → COK	21:50	03:35 02 Jun	5h 45m	1 stop	No info
29	GoAir	15/05/2019	Delhi	Cochin	DEL → BOM → COK	07:00	12:55	5h 55m	1 stop	No info

B. Handling missing values and inconsistencies, ensuring that the data is ready for analysis.

The meticulous handling of missing values and inconsistencies played a pivotal role in ensuring the dataset's readiness for analysis. This crucial step involved addressing gaps and irregularities in the dataset, which could potentially skew the accuracy and reliability of our flight fare prediction model shown in figure 2. Missing values, a common occurrence in real-world datasets, were carefully addressed to prevent any distortion in the analytical process. This entails employing strategies such as imputation or removal of rows with incomplete information, depending on the nature and significance of the missing data. By systematically handling missing values, we aimed to enhance the completeness and integrity of the dataset. Overall, the conscientious handling of missing values and inconsistencies reflects our commitment to data quality, laying the groundwork for accurate and reliable insights in our flight fare prediction model.

```
### getting all the rows where we have missing value

train_data[train_data['Total_Stops'].isnull()]
```

Python

	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info	Price
9039	Air India	6/05/2019	Delhi	Cochin	NaN	09:45	09:25 07 May	23h 40m	NaN	No info	7480

Figure 2: Handling missing values and inconsistencies

C. Analyzing when will most flights take off

The analysis of when most flights take off is a pivotal aspect of understanding temporal patterns and trends in the flight dataset and is shown in figure 3. By examining the distribution of flight departure times, we aimed to identify the peak periods when a significant number of flights are scheduled to depart. This analysis involves creating visual representations, to showcase the frequency distribution of departure times. Insights gained from this examination can provide valuable information about the preferred times for flight departures, potential rush hours, and variations in travel patterns throughout the day.

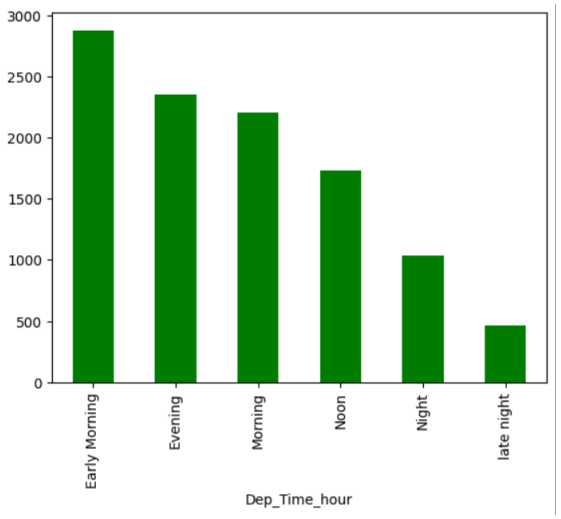

Figure 3: Analyzing when will most flights take off

D. Analyzing whether flight duration impacts price or not

Analyzing whether flight duration impacts price is a critical investigation to understand the relationship between the duration of a flight and its corresponding ticket price. It is shown in figure 4. This analysis aims to uncover patterns and trends that indicate how the length of a flight influences the cost of airfare.

The approach involves visualizing the correlation between flight duration and ticket prices, often through scatter plots or regression analysis. By examining these relationships, we gain insights into whether longer flights tend to have higher or lower prices and whether there are any specific trends in the pricing structure based on the duration of the journey.

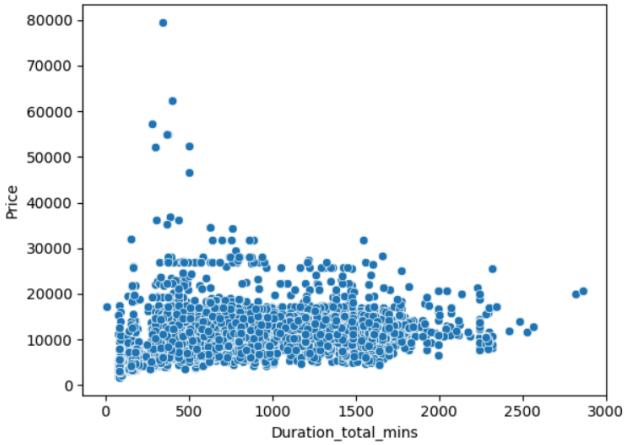

Figure 4: Analyzing whether flight duration impacts price or not

E. Analyzing how the no. of stops impacts price

Analyzing how the number of stops impacts the price of flights is a crucial exploration aimed at understanding the correlation between the number of layovers in a journey and the corresponding ticket prices. This analysis is pivotal for both airlines and passengers, providing insights into pricing dynamics based on the convenience or inconvenience of non-stop or multi-stop flights.

The methodology involves examining the relationship between the number of stops and flight prices, typically through visual representations like scatter plots or statistical

analyses. It is shown in figure 5.

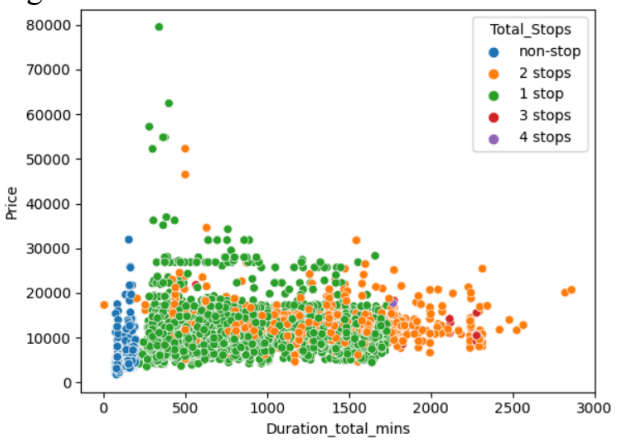

Figure 5: Analyzing how the no. of stops impacts price

F. Finding which airline has the highest price

Finding which airline has the highest price involves a comprehensive examination of the pricing variations among different airlines and is shown in figure 6. This analysis is crucial for both passengers and the airline industry, shedding light on the competitiveness of various carriers and helping travelers make informed decisions based on their budget constraints and preferences.

The methodology typically employs data visualization tools such as box plots, which provide a clear overview of the price distributions for different airlines. By displaying the data in this way, it becomes apparent which airlines tend to have higher or lower ticket prices. This insight is valuable for both airlines and passengers.

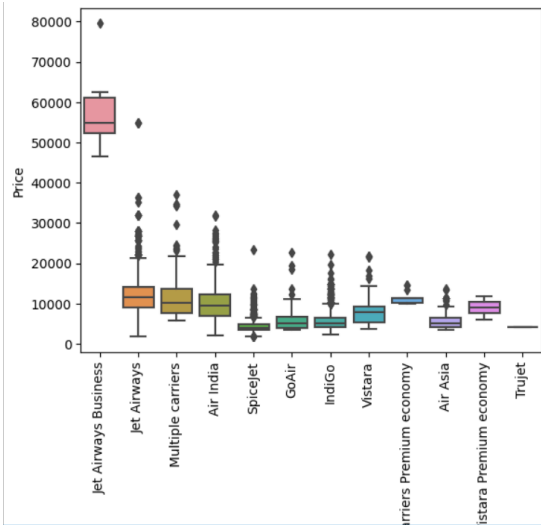

Figure 6: Finding which airline has the highest price

G. Applying one-hot Encoding:

Applying one-hot encoding is a crucial step in preparing categorical data for machine learning algorithms. In the context of this project, it involves transforming categorical variables into a numerical format that can be easily utilized by machine learning models.

Categorical variables, such as the airline, source, destination, and other factors, are initially

represented as labels or names. However, machine learning models typically require numerical input.

H. Finding mean prices for airlines:

Finding the mean prices for airlines involves analyzing and summarizing the average ticket prices for each airline in the dataset. It is represented in figure 7. This step provides valuable insights into the pricing strategies of different airlines, helping users make informed decisions based on historical pricing trends.

In the context of this project, the code snippet data.groupby(['Airline'])['Price'].mean().sort_values() is utilized. Here, the dataset is grouped by the 'Airline' column, and the mean price for each airline is calculated. Sorting the results allows for a clear comparison of average prices across different airlines.

The resulting information is particularly useful for travelers seeking cost-effective options or those interested in understanding the price variations among different airlines. By identifying airlines with lower or higher average prices, users can tailor their choices based on budget constraints or preferences.

```
data.groupby(['Airline'])['Price'].mean().sort_values()
```

```
Airline
Trujet                            4140.000000
SpiceJet                          4338.284841
Air Asia                          5590.260188
IndiGo                            5673.682903
GoAir                             5861.056701
Vistara                           7796.348643
Vistara Premium economy           8962.333333
Air India                         9612.427756
Multiple carriers                10902.678094
Multiple carriers Premium economy    11418.846154
Jet Airways                      11643.923357
Jet Airways Business             58358.666667
Name: Price, dtype: float64
```

Figure 7: Finding mean prices for airlines

I. Plotting graphs to spot outliers

The code snippet for plotting graphs to spot outliers involves visualizing the distribution of the 'Price' column to identify potential outliers. This step is crucial in understanding the data's spread and detecting any extreme values that might significantly impact the accuracy of the machine learning models. The function plot(df, col) is defined to create three subplots: a distribution plot, a box plot, and a histogram. These visualizations help in assessing the central tendency, spread, and skewness of the 'Price' column. Figure 8 shows plots to identify outliers.

Distribution Plot

The distribution plot provides an overview of the data's distribution, highlighting the concentration of prices and potential deviations from a normal distribution.

Box Plot

The box plot displays the interquartile range (IQR), indicating the spread of prices and helping identify any data points beyond the IQR, which could be considered outliers.

Histogram

The histogram complements the distribution plot by providing a visual representation of the frequency of different price ranges. The absence of a smooth curve (kernel density estimate) in this case aids in focusing on individual data points.

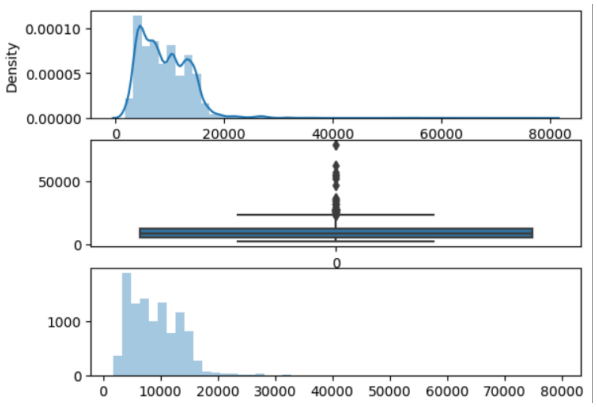

Figure 8: Plotting graphs to spot outliers

J. Outlier Detection Using IQR Method for 'Price' Column

Calculate Quartiles and IQR:
The first quartile (Q1) and third quartile (Q3) are calculated using the quantile method on the 'Price' column as shown in figure 9. The Interquartile Range (IQR) is computed as the difference between Q3 and Q1.

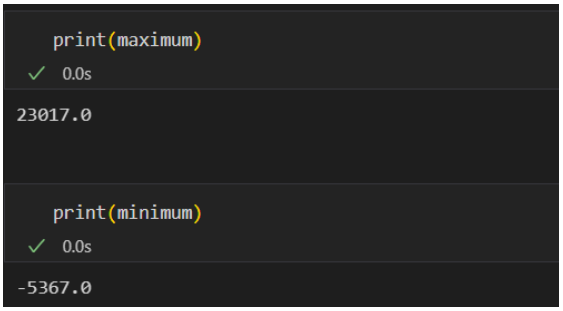

Figure 9: Outlier Detection Using IQR Method for 'Price' Column

Define Thresholds:
The thresholds for identifying outliers are determined by extending beyond the IQR.
The maximum and minimum values beyond which data points are considered outliers are calculated.

K. Plotting graphs after altering the Price column

Identify Outliers:
Data points with prices above the upper threshold or below the lower threshold are considered outliers.
Handle Outliers:
The identified outliers can be handled, for example, by replacing them with a central tendency measure like the median, shown in figure 10.
This process helps in ensuring that extreme values in the 'Price' column, which could potentially impact the performance of machine learning models, are appropriately addressed. It contributes to refining the dataset for more accurate predictions.

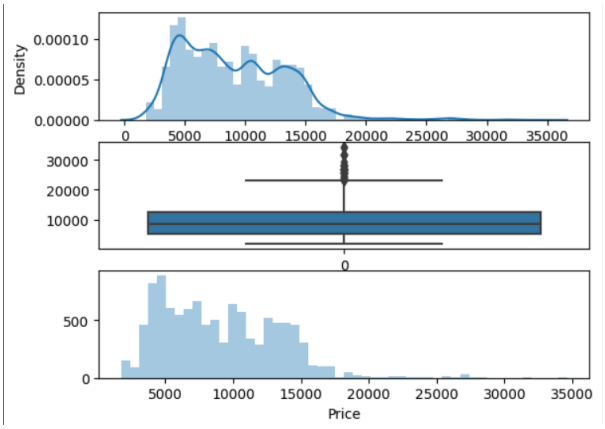

Figure 10: Plotting graphs after altering the Price column

L. Calculating mutual information between the target variable "Price" and other features

In this step, the code calculates the mutual information between the target variable, which is the flight ticket price ("Price"), and the other features present in the dataset, shown in fugure 11. Mutual information is a statistical measure that quantifies the level of dependence between two variables. In the context of this project, it helps assess how much information each feature carries about the variation in flight prices.

The mutual information scores are computed using the mutual_info_regression function from scikit-learn. These scores indicate the amount of shared information between each feature and the target variable. A higher mutual information score suggests a stronger relationship between a feature and the flight price, making it a valuable predictor.

By analyzing these scores, the project gains insights into which features play a significant role in influencing the variation in flight prices. This information is crucial for feature selection and model training, contributing to the overall accuracy and effectiveness of the flight fare prediction model.

```
imp
✓ 0.0s
array([1.3189775 , 1.06205482, 0.7804578 , 0.37177855, 0.63077249,
       0.93249179, 0.75818062, 1.14948805, 0.90579211, 1.11789445,
       0.67773981, 0.95880631, 0.68793412, 0.39284877, 0.45123736,
       0.52368471, 0.13814344, 0.19933225])
```

Figure 11: Calculating mutual information between the "Price" and other features

M. Training the model and predicting price for the test data:

In this phase, the machine learning model is trained using historical flight data to learn patterns and relationships between various features and the target variable, which is the flight ticket price ("Price"). It is shown in figure 12. The code utilizes the Random Forest Regressor model for this task. Random Forest Regressor is an ensemble learning algorithm that builds multiple decision trees and merges their predictions to improve accuracy and robustness.

The training is performed on a subset of the dataset, typically referred to as the training set (X_train, y_train). Once the model is trained, it is then used to predict the flight prices for a separate set of data, known as the test set (X_test). The predicted prices (y_pred) are obtained, and these predictions can be compared with the actual prices (y_test) to evaluate the model's performance.

```
array([16808.43,  5538.36,  8810.65, ...,  3498.38,  6205.67,  6818.43])
```

Figure 12: Training the model and predicting price for the test data

N. Model Evaluation and Prediction Analysis:

In this stage, the code focuses on evaluating the performance of the machine learning model that has been trained on historical flight data. The Random Forest Regressor model is employed for this purpose. The evaluation is crucial to assess how well the model generalizes to new, unseen data and makes predictions on flight prices. It is shown in figure 13.

The code calculates various evaluation metrics to gauge the model's accuracy and effectiveness. These metrics include:

Training Score: The score achieved by the model on the training data, indicating how well it fits the training set.

Predictions: The code generates predictions for the flight prices in the test set using the trained model.

R-squared Score: A measure of how well the model's predictions match the actual flight prices, with a higher R-squared score indicating a better fit.

$$R^2 = 1 - \frac{RSS}{TSS}$$

$R^2 = coefficient\ of\ determination$
$RSS = sum\ of\ squares\ of\ residuals$
$TSS = total\ sum\ of\ squares$

Mean Absolute Error (MAE): The average absolute difference between the predicted and actual prices.

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

$MAE = mean\ absolute\ error$
$y_i = prediction$
$x_i = true\ value$
$n = total\ number\ of\ data\ points$

Mean Squared Error (MSE): The average of the squared differences between predicted and actual prices.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

$MSE = mean\ squared\ error$
$n = number\ of\ data\ points$
$Y_i = observed\ values$
$\hat{Y}_i = predicted\ values$

Root Mean Squared Error (RMSE): The square root of the MSE, providing a measure of the average magnitude of the errors.

The sum of the squared differences between the predicted and observed values is divided by

the number of observations, and the square root of the result is taken to yield the RMSE.

$$RMSE = \sqrt{\left[\frac{\left(\Sigma\left(P_i - O_i\right)^2\right)}{n}\right]}$$

$RMSE = root\ mean\ squared\ error$
$n = number\ of\ data\ points$
$P_i = predicted\ values$
$O_i = observed\ values$

Mean Absolute Percentage Error (MAPE): The average percentage difference between predicted and actual prices.

$$M = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

$M = mean\ absolute\ percentage\ error$
$n = number\ of\ times\ the\ summation$
$iteration\ happens$
$A_i = actual\ value$
$F_t = forecast\ values$

Additionally, the code generates a distribution plot to visualize the differences between the predicted and actual flight prices. This analysis aids in understanding the strengths and limitations of the model, guiding potential adjustments for future enhancements.

```
Training score : 0.9513767685922624
predictions are : [16789.81  5404.71  8789.86 ...  3511.14  6337.6   6868.64]


r2 score : 0.8116751783166001
MAE : 1180.2396909077447
MSE : 3666232.235595947
RMSE : 1914.740775038738
MAPE : 13.220929084951152
```

Figure 13: Model Evaluation and Prediction Analysis

CONCLUSION

The flight fare prediction method proposed in this paper establishes the potential and success of employing machine learning algorithms for forecasting airline ticket prices, providing a reliable tool for travelers. The utilization of an ensemble of algorithms, specifically Random Forest and Decision Trees, showcases the versatility and effectiveness of combining different models to achieve accurate and robust predictions, catering to diverse scenarios. Feature importance analysis not only enhances the transparency of the model but also empowers users with valuable insights into the key factors influencing flight prices, fostering a deeper understanding of the prediction process. The paper's extensive exploration and analysis, including data collection, preprocessing, and various machine learning model implementations, contribute to a comprehensive framework for future research and advancements in the domain of flight fare prediction. The successful handling of missing values, outlier detection, and data cleaning in the preprocessing phase demonstrates the paper's

commitment to ensuring high-quality data for training and evaluation. The comparison and evaluation of machine learning models, provide a comprehensive view of their respective performances, offering flexibility and adaptability for different prediction scenarios. The exploration of additional factors, such as the impact of the number of stops, flight duration, and airline choices on pricing, enriches the understanding of the complexities influencing ticket fares. The paper's systematic approach, from problem definition to model evaluation, serves as a valuable reference for future researchers and practitioners entering the field of predictive analytics in the airline industry.

FUTURE SCOPE

The current implementation doesn't explore the incorporation of external factors such as economic indicators, geopolitical events, and market dynamics. These factors will be incorporated in the future to improve the model's adaptability to real-world scenarios.Advanced hyper-parameter tuning techniques and optimization algorithms will be investigated and integrated in the current scheme to streamline the process of fine-tuning the ensemble models, reducing computational requirements. It is also planned to develop mechanisms for dynamic data updates, allowing the model to adapt to changing market conditions and ensuring consistently accurate predictions. A feedback mechanism can also be integrated that allows users to provide feedback on predicted fares, enabling continuous improvement of the model based on user experience.

REFERENCES

1. Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.
2. Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer.
3. Valiant, L. G. (1984). A theory of the learnable. Communications of the ACM, 27(11), 1134-1142.
4. Rao, N. S. S. V. S., & Thangaraj, S. J. J. (2023, April). Flight Ticket Prediction using Random Forest Regressor Compared with Decision Tree Regressor. In 2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM) (pp. 1-5). IEEE.
5. Burger, B., & Fuchs, M. (2005). Dynamic pricing—A future airline business model. Journal of Revenue and Pricing Management, 4(1), 39-53.
6. Malighetti, P., Paleari, S., & Redondi, R. (2010). Has Ryanair's pricing strategy changed over time? An empirical analysis of its 2006–2007 flights. Tourism management, 31(1), 36-44.
7. Liu, T., Cao, J., Tan, Y., & Xiao, Q. (2017, December). ACER: An adaptive context-aware ensemble regression model for airfare price prediction. In 2017 International Conference on Progress in Informatics and Computing (PIC) (pp. 312-317). IEEE.
8. Tziridis, K., Kalampokas, T., Papakostas, G. A., & Diamantaras, K. I. (2017, August). Airfare prices prediction using machine learning techniques. In 2017 25th European Signal Processing Conference (EUSIPCO) (pp. 1036-1039). IEEE.
9. Can, Y. S., & Alagöz, F. (2023, October). Predicting Local Airfare Prices with Deep Transfer Learning Technique. In 2023 Innovations in Intelligent Systems and Applications Conference (ASYU) (pp. 1-4).

IEEE.

10. Malkawi, M., & Alhajj, R. (2023, August). Real-time web-based International Flight Tickets Recommendation System via Apache Spark. In 2023 IEEE 24th International Conference on Information Reuse and Integration for Data Science (IRI) (pp. 279-282). IEEE.
11. Joshitta, S. M., Sunil, M. P., Bodhankar, A., Sreedevi, C., & Khanna, R. (2023, May). The Integration of Machine Learning Technique with the Existing System to Predict the Flight Prices. In 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) (pp. 398-402). IEEE.
12. Groves, W., & Gini, M. (2013, May). An agent for optimizing airline ticket purchasing. In Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems (pp. 1341-1342).
13. Domínguez-Menchero, J. S., Rivera, J., & Torres-Manzanera, E. (2014). Optimal purchase timing in the airline market. Journal of Air Transport Management, 40, 137-143.
14. Sarao, P., & Samanta, P. (2022). Flight Fare Prediction Using Machine Learning. Available at SSRN 4269263.
15. Python Software Foundation. Python 3 Documentation: https://docs.python.org/3/ - The official documentation for Python 3, pivotal in the development and implementation of machine learning models for flight price prediction.
16. NumPy Community. NumPy Documentation: https://numpy.org/doc/stable/ - NumPy, a fundamental package for scientific computing with Python, played a crucial role in handling numerical operations and data manipulation in the flight price prediction project.
17. Pandas Development Team. Pandas Documentation: https://pandas.pydata.org/pandas-docs/stable/ - Pandas, a powerful data manipulation library, was used extensively for data preprocessing, cleaning, and analysis in the flight price prediction project.
18. Scikit-learn Developers. Scikit-learn Documentation: https://scikit-learn.org/stable/documentation.html - Scikit-learn provided a wide range of machine learning tools and models, including Random Forest, employed for building predictive models in the project.
19. Matplotlib Development Team. Matplotlib Documentation: https://matplotlib.org/stable/contents.html - Matplotlib, a popular data visualization library, played a key role in creating visualizations and plots to analyze trends and patterns in flight price data.
20. Seaborn Development Team. Seaborn Documentation: https://seaborn.pydata.org/ - Seaborn, built on top of Matplotlib, was utilized for enhancing the aesthetics of visualizations and facilitating data exploration.
21. Kaggle. Flight Price Prediction Dataset: https://www.kaggle.com/nikhilmittal/flight-fare-prediction-mh - Kaggle provided the dataset used in the project, offering a valuable resource for historical flight pricing data.
22. Stack Overflow: https://stackoverflow.com/ - A community of developers where technical questions related to coding, data analysis, and machine learning were addressed.