Analysing Airline Dataset using Sentiment analysis and Price prediction in Python

Members : Ayush Kumar Yadav, Rishikesh Dargad, Kritika Jaimini, Sohini Bhadra, Sneha Karmakar

2024-04-06

Executive Summary

In this report, we will perform price prediction and sentiment analysis using R and Python. We will first gather historical price data and then build a predictive model to forecast future prices.

We study different table columns, try to relate them to others and find a relationship between them.

We try to find and analyze critical factors like the class of travel, flight duration, etc., which help us understand the pricing of tickets to plan and schedule our air travel efficiently.

Additionally, we will analyze sentiment data from the obtained dataset to gauge customers' sentiment.

Price Prediction

Table of Contents

3. Data collection and methods

3.1 Data Sources

3.2 Variables Considered

3.3 Data Preprocessing

4. Visualizing the dataset

4.1 Libraries used in the process

4.2 Plots we have used

4.3 Plot of Number of flights of Airlines in India.

4.4 Plot of Ticket Costs

4.5 Plot of Duration of Flights

4.6 Plot on Tickets booked before X days

4.7 Plot of Departure time and Arrival time

4.8 Plot on Airline ticket price based on the source and destination cities.

4.9 Plot of Distribution of most Airlines in economy class

4.10 Plot of Total price by Airline in economy class

4.11 Plot of source city in economy class.

4.12 Plot of most airlines in business class and outcome.

4.13 Plot of cities used business class tickets

4.14 Correlation Heatmap

5. Modeling

6. Conclusion

1. Introduction

1.1 Overview

In this part of the project, we study the data in tabular format using various Python libraries like Pandas, Numpy, Matplotlib, and Seaborn.

We study different table columns, try to co-relate them with others, and find a relation between them.

We try to find and analyze those critical factors like the class of travel, duration of the flight, etc., which helps us understand the pricing of tickets to plan and schedule our air travel in an efficient way

1.2 Questions We imposed

- What is the number of flights operated by each airline?
- What is the price range according to class of travel?
- What is the availability of Tickets according to class of travel?
- What is the price of tickets for different airlines based on the flight duration?
- How do ticket prices vary across different airlines and classes of travel?
- How do airline ticket prices vary depending on when you buy them?
- How does the price of a ticket vary depending on duration?
- How does ticket price vary according to departure time and arrival time?
- How does ticket price vary depending on source and destination?
- How does the price of tickets vary based on no. of stops and airline?

1.3 Scope of Analysis

The study covers detailed exploratory data analysis on different key factors of an Indian Airline Dataset.

The project also enables us to predict the prices of different airlines based on various factors using the LINEAR REGRESSION MODEL.

2. Explaing the dataset

2.1 Overview of Dataset

The dataset is a .csv file format obtained from Google search. The dataset consists of 3000154 rows and 12 columns.

The airline column has six unique airlines: SpiceJet, AirAsia, Vistara, GO_FIRST, Indigo, and Air_India. In source_city & destination_city, there are six unique cities: Delhi, Mumbai, Bangalore, Kolkata, Hyderabad, and Chennai. In the arrival & departure columns, there are six different timings: Night, Morning, Early_Morning, Afternoon, Evening, and Late_Night In a class column, there are two different classes: Economy, Business

2.2 Problems with the Dataset

The main problem with datasets is that the data is two to three years old, so the price prediction might not match the current scenario of airline ticket prices.

3. Data collection and methods

3.1 Data source

Data is being derived from Google search.

3.2 Variables Considered The various features of the dataset are explained below:

- Airline: The name of the airline company is stored in the airline column. It is a categorical feature, having six different airlines.
- Flight: Flight stores information regarding the plane's flight code. It is a categorical feature.
- Source City: The city from which the flight takes off. It is a categorical feature, having six unique towns.
- Departure Time: This derived categorical feature is created by grouping periods into bins. It stores information about the departure time and has six unique time labels.
- Stops: A categorical feature with three distinct values that stores the number of stops between the source and destination cities.
- Arrival Time: This derived categorical feature is created by grouping time intervals into bins. It has six distinct time labels and keeps information about the arrival time.
- Destination City: The city where the flight will land. It is a categorical feature, having six unique towns.
- Class: A categorical feature that contains information on seat class; it has two distinct values: Business and Economy.
- Duration: A continuous feature that displays the time it takes to travel between cities in hours. 10)Days Left: This is a derived characteristic calculated by subtracting the trip date from the booking date.
- Price: Target variable stores information on the ticket price.

3.3 Data Preprocessing

For model fitting, data is preprocessed using LabelEncoder from sklearn—preprocessing module.

- we have to transform the airline, 'source_city,' 'destination_city,' 'departure_time,' 'arrival_time,' 'stops,' 'class' column of data frame 'defusing 'fit_transform' method involving fitting a LabelEncoder on data and then transforming it
- We have dropped the column 'unnamed' from data frame 'df'
- We have dropped the column 'flight' from a dataset that is useless to us in modeling.

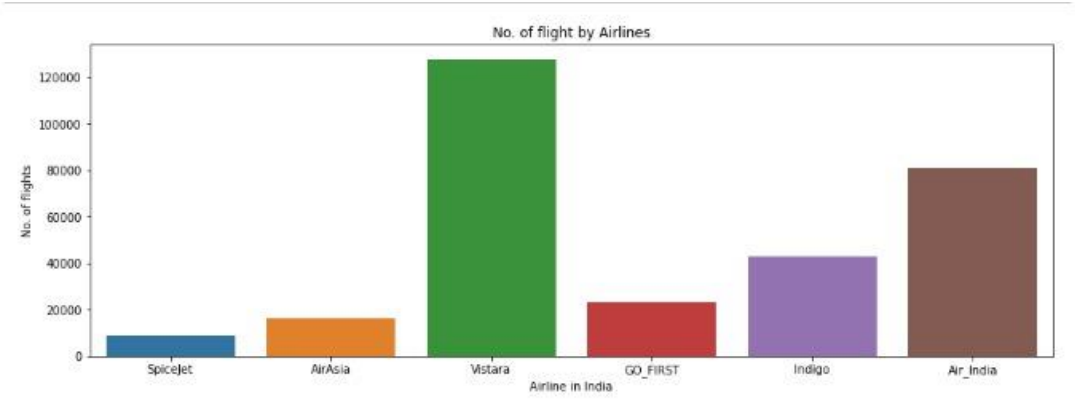
4. Visualizing the dataset

4.1 Libraries used in the process - matplotlib - seaborn

4.2 Plots we have used - bar graphs - line graphs - pie charts - strip plots - scatter plot

4.3 Plot of Number of flights of Airlines in India.

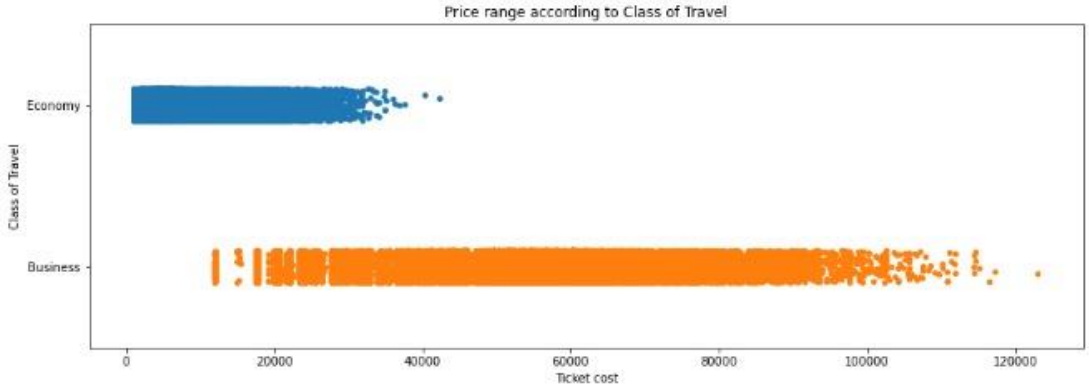
We draw this plot using sns—counterplot for the 'airline' column.



- From the plot, we can see 'Vistara' has the maximum no. of fights followed by 'Air India' while 'Spice Jet' has the most diminutive no. of flights

4.4 Plot of Ticket Costs

The seaborn. stripplot() method is used to draw a scatter plot where one of the variables passed as argument is a categorical variable. Here, the categorical variable is “class of travel.” From the above figure, we can see ‘Economy’ class tickets usually cost between 2500 - 22500 while ‘Business’ class tickets typically cost between 25000 - 95000



4.5 Plot of Duration of Flights

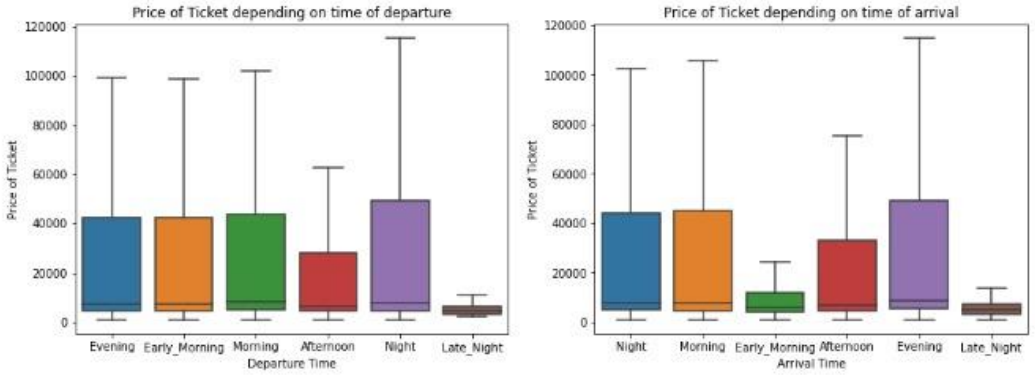
This shows that the distribution of ticket prices varies with the duration of the flight. The green and brown points in the figure are explained by the fact that ‘Vistara’ and ‘Air India’ have a maximum no. of flights. Indigo has more short-duration flights at comparatively cheaper prices.

4.6 Plot on Tickets booked before X days

This plot concludes that ticket prices rise slowly till 20 days from the flight date, then rise sharply till the last day while dramatically reducing just one day before the flight date. This can be explained by the fact that people usually buy flight tickets within 2-3 weeks of flight, generating more profits for airlines. On the last day, prices show dramatic reduction as airlines hope to fill the flight due to an increase in the load factor and a decrease in the operational cost per passenger.

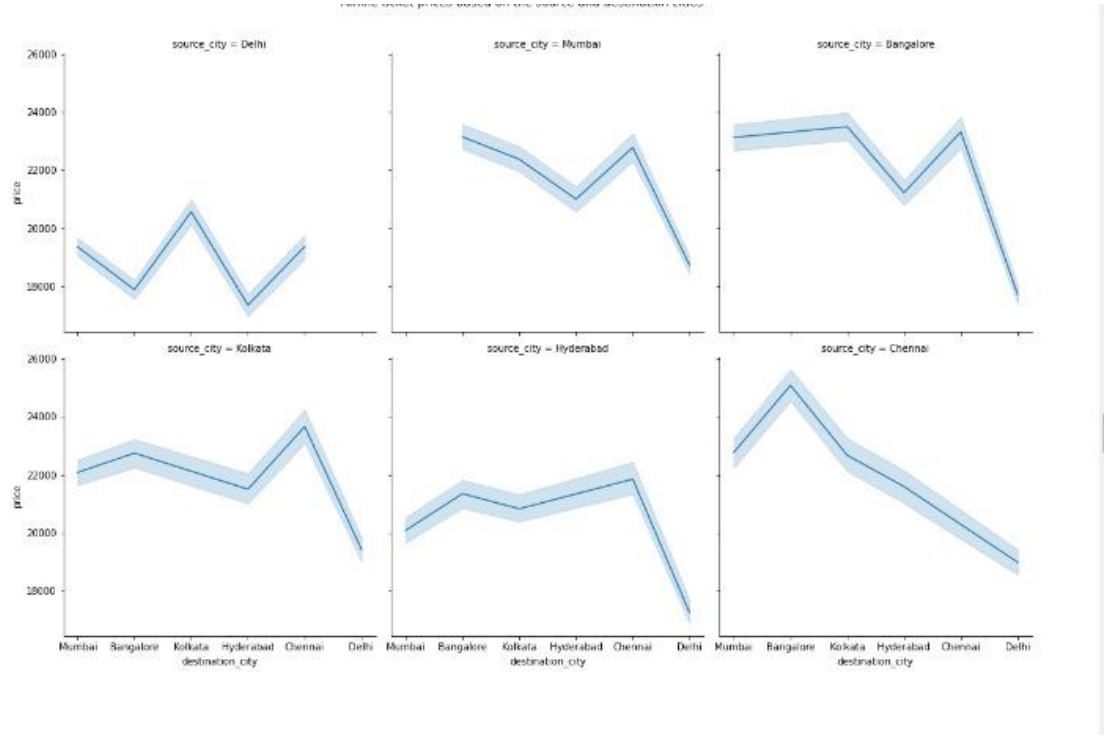
4.7 Plot of Departure time and Arrival time

This plot concludes that flights departing late at night are the cheapest, while those arriving early in the morning and late at night are cheap, too. Flights departing in the afternoon are relatively inexpensive as well.



4.8 Plot on Airline ticket price based on the source and destination cities.

This plot shows that flights departing from Delhi are usually cheaper, which can be explained by the fact that Delhi, the capital, has robust connectivity with every other city and more frequencies, resulting in more affordable ticket prices. Chennai-Bangalore is the most expensive flying route, while Hyderabad is the city's most expensive.

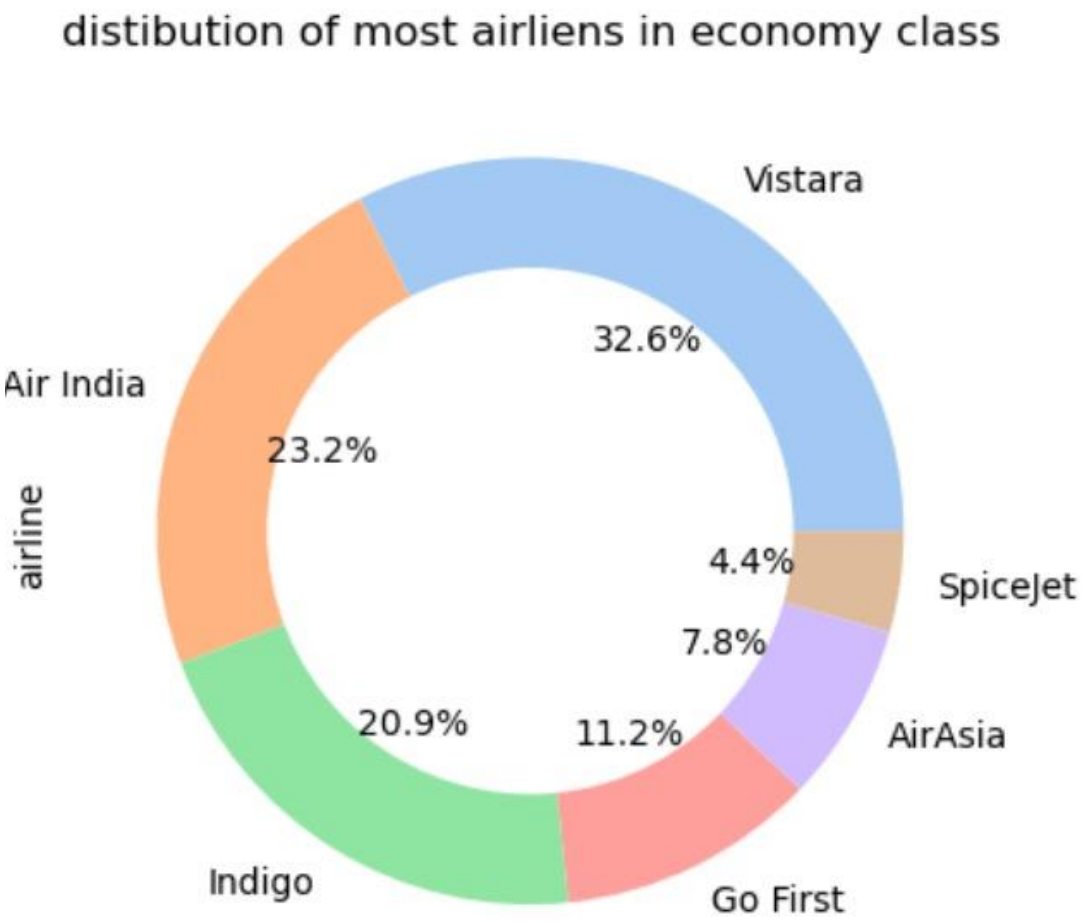


4.9 Plot of Distribution of most Airlines in economy class

This plot gives us insight into most Airlines' distribution in economy class.

Distribution of Airlines in economy class is as follows:

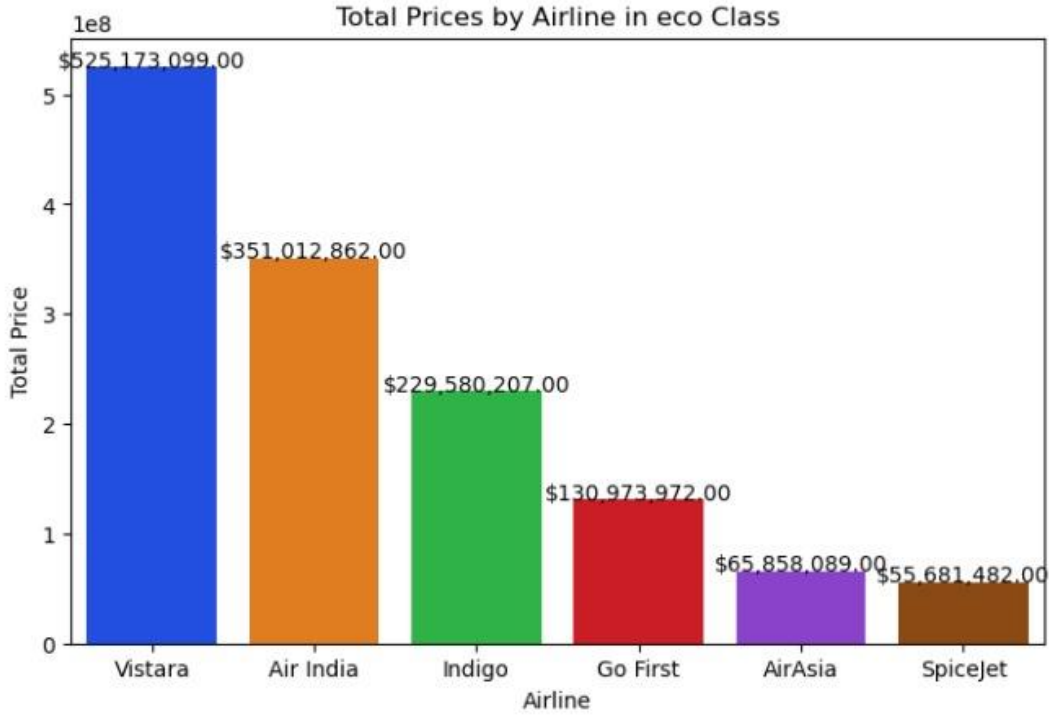
- Air India: 23.2%
- Indigo: 20.9%
- Go First: 11.2%
- Air Asia: 7.8%
- Spicejet: 4.4%
- Vistara: 32.6%



4.10 Plot of Total price by Airline in economy class

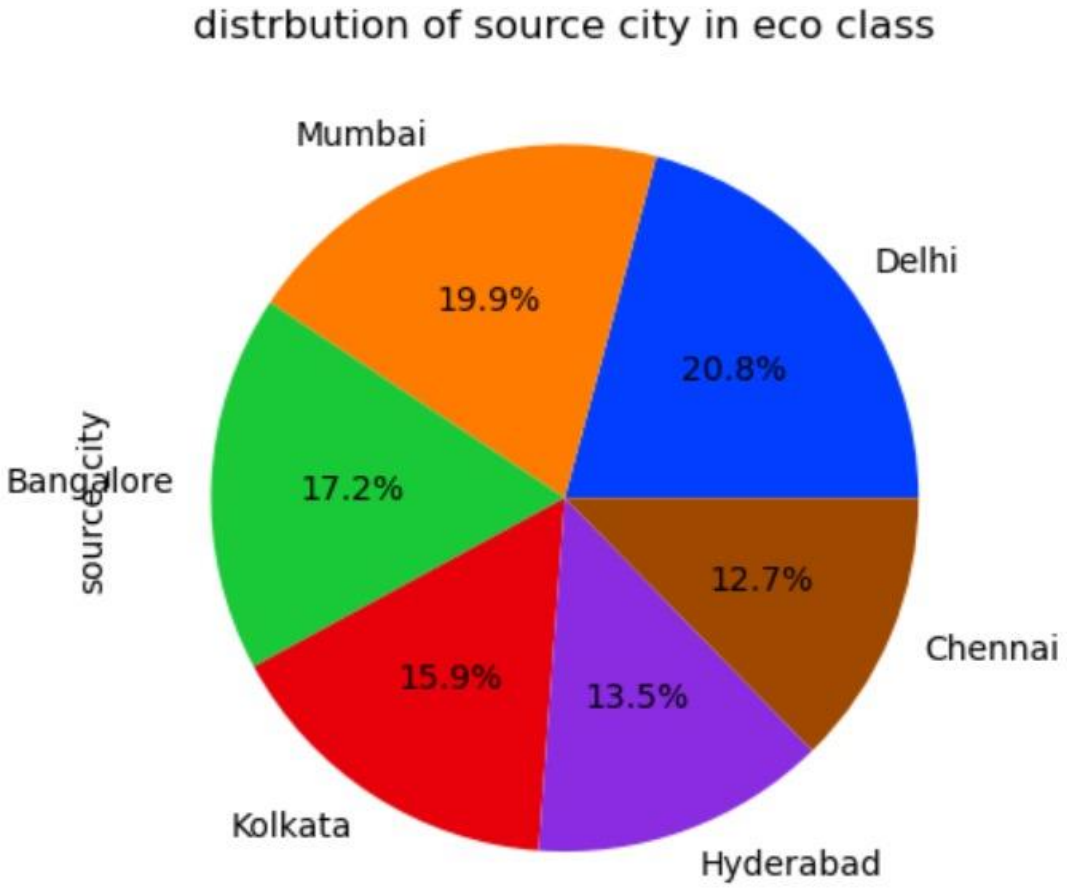
This barplot shows the total price by the Airline.

-Vistara : $525,173,099.00 -Air India: $351,012,862.00 -Indigo : $229,580,207.00 - Go First : $130,973,972.00 - Air Asia : $65,858,089.00 -Spicejet : $55,681,482.00

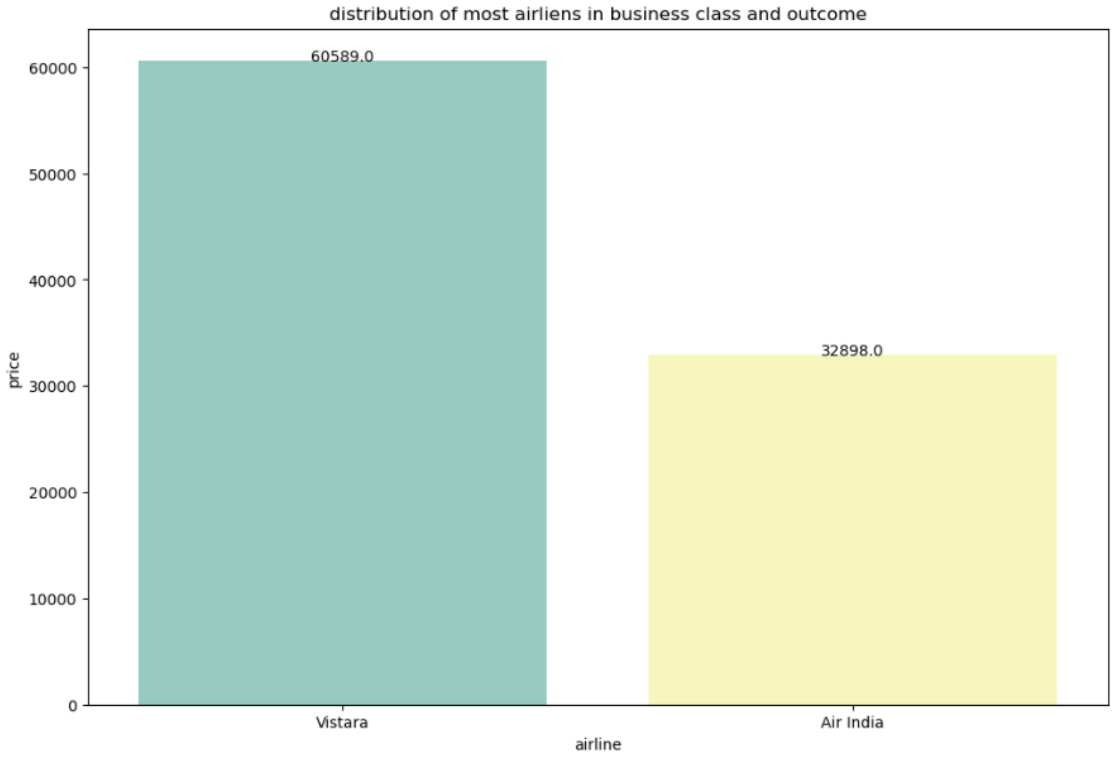


ch city is most used in economic tickets?

4.11 Plot of source city in economy class. The pie chart shows that Delhi is the source city for economy class, followed by Mumbai.

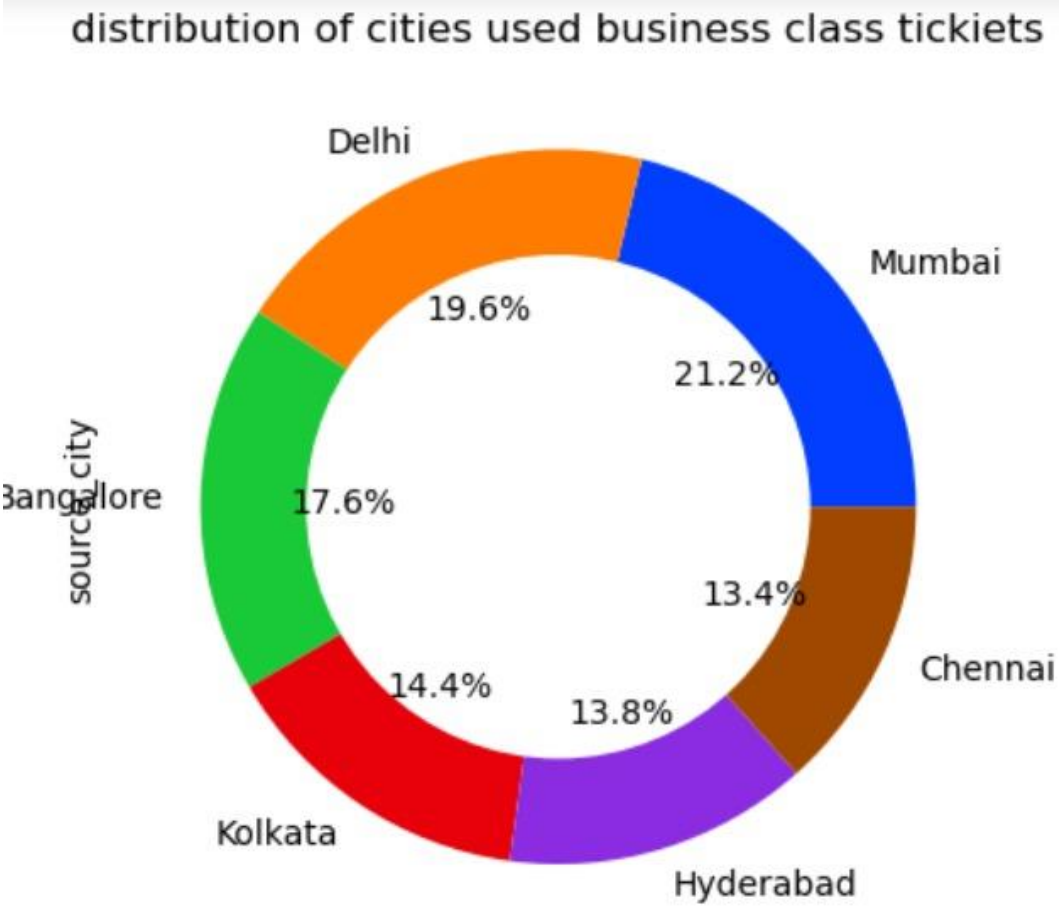


4.12 Plot of most airlines in business class and outcome.



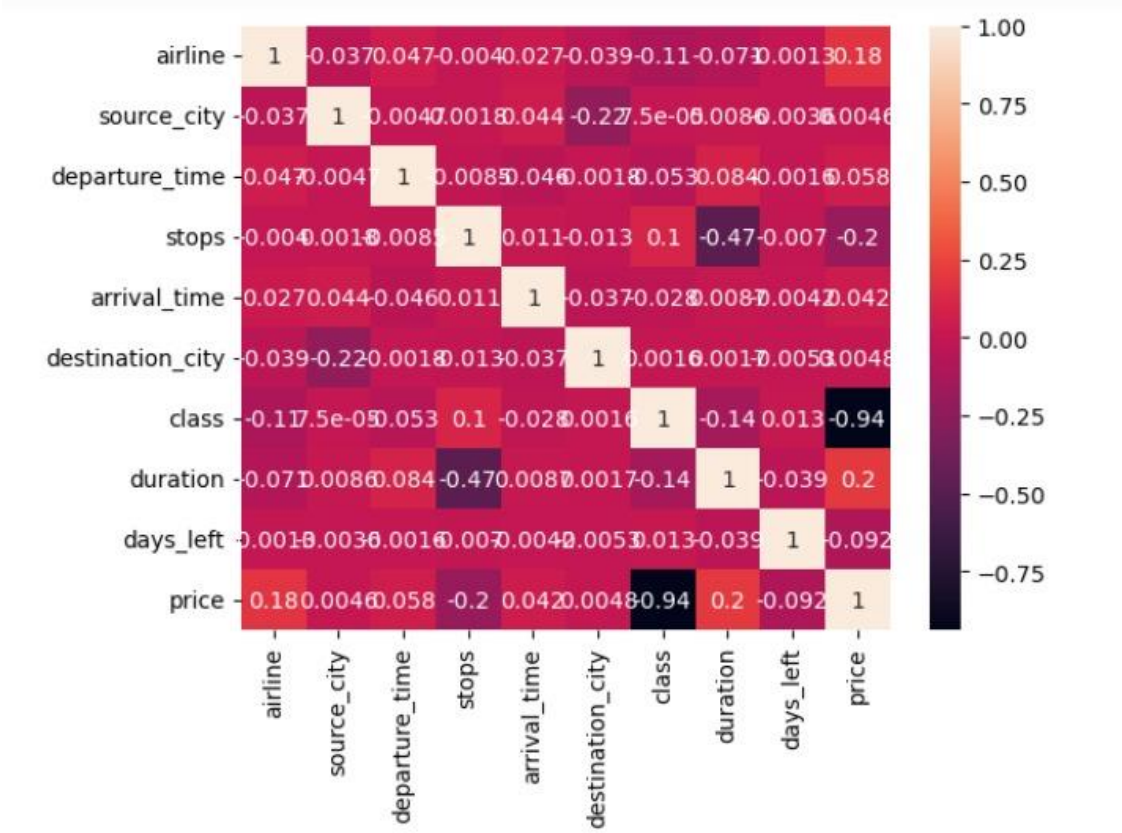
4.13 Plot of cities used business class tickets

The pie chart shows that Mumbai is the source city for business class, followed by Delhi.



4.14 Correlation Heatmap

The correlation heatmap describes the correlation among various features of the dataset.



5. Modeling

The model is trained on a basic Linear Regression model.

We imported test_train_test_split from sklearn.model_selection modeule.

Then, we imported the LinearRegression classifier from linear_model. We have split the dataset for training and testing. 80% of the data is used for training, and 20% is used to test the model.

After fitting the model, we predicted the prices of flight tickets based on testing data. We also evaluated Mean Absolute Error, Mean Square Error, and Mean Absolute Percentage Error.

6. Conclusion

For the linear regression model on the dataset, the following are the conclusions:

-train score 0.9048254651102412

-test score 0.9047473929854847

-r2 score 0.9047473929854847

-MAE 4650.70263783112

-MSE 49101030.427204736

-MAPR 0.443777705784786

Sentiment Analysis

Table of Contents

Overview

1. The project delves into assessing customer sentiment towards SpiceJet by employing two distinct sentiment analysis models, Naive Bayes, and SVC, leveraging the Bag of Words technique for feature extraction.

2. Notably, the Naive Bayes model emerges as the frontrunner, exhibiting enhanced efficacy in identifying positive reviews, as evidenced by its higher F1 score, thereby shedding light on the potential prowess of unsupervised learning paradigms.

3. Furthermore, to bolster the overall accuracy and comprehensiveness of the analysis, proposed enhancements encompass the adoption of advanced text cleaning methodologies, exploration of diverse feature extraction techniques, and meticulous error analysis, all of which are poised to yield more profound insights into customer satisfaction trends and nuances within the SpiceJet ecosystem.

What is Sentiment Analysis?

Sentiment analysis, or opinion mining, is a natural language processing (NLP) technique to determine a text's sentiment or emotional tone. It involves analyzing text data to discern whether the expressed opinion is positive, negative, or neutral. Sentiment analysis algorithms typically classify text into these categories based on the presence of certain words, phrases, or linguistic patterns that convey sentiment. This analysis can be applied to various text data types, such as customer reviews, social media posts, news articles, and more, to gauge public opinion, sentiment trends, and customer satisfaction.

How was sentiment analysis applied to Indian airlines?

The objective of this project was to utilize sentiment analysis as a tool to comprehensively understand and analyze the sentiments expressed by customers towards Indian airlines, thereby providing valuable insights into customer satisfaction, sentiment trends, and potential areas for improvement.

Scope of this Project

This project focuses on sentiment analysis of customer reviews from prominent airlines, including Go Air, Air India, Jet Airways, Indigo, and Spicejet. We consider a wide range of review attributes, including the content of the reviews, ratings, and other relevant factors. However, our analysis is limited to text-based reviews and does not cover different feedback forms, such as ratings or surveys.

Significance of this Analysis

Airlines must understand customer sentiments to enhance passenger experiences, improve service quality, and make informed business decisions. By analyzing customer reviews and classifying sentiments, airlines can identify areas for improvement, address customer concerns, and ultimately increase customer satisfaction and loyalty. Our project aims to provide valuable insights and actionable recommendations for airlines to serve passengers better and stay competitive.

Objectives of the Project

Primary Objectives:

-Perform sentiment analysis on customer reviews from multiple airlines.

-Classify reviews into positive, negative, or neutral sentiments using machine learning techniques.

-Identify common themes and patterns in customer feedback across different airlines.

Secondary Objectives:

-Evaluate the performance of different machine learning models for sentiment classification.

-Compare sentiment expressions across different airlines and analyze factors influencing passenger perceptions.

-Generate actionable insights and recommendations for airlines to improve customer satisfaction and service quality.

Expected Outcomes:

-Development of machine learning models for sentiment classification with high accuracy and performance.

-Insights into customer sentiments and preferences across different airlines.

-Recommendations for airlines to address common pain points and improve overall customer satisfaction.

Libraries Used

JSON: Utilized to read and parse JSON files containing customer reviews from various airlines.

nltk (Natural Language Toolkit): Employed for preprocessing text data, including tokenization, stop word removal and stemming, to prepare it for sentiment analysis.

textblob: Used for sentiment analysis tasks, including polarity analysis, to determine the sentiment expressed in customer reviews.

NumPy: Utilized for numerical computations and array manipulation required for data processing and machine learning tasks.

matplotlib: Utilized to create visualizations such as plots and charts to illustrate the results of sentiment analysis and other data analysis tasks.

Pandas: Used for data manipulation and analysis, including preprocessing structured data such as customer review datasets.

Learn (sci-kit-learn): Employed to implement machine learning algorithms, including support vector machines (SVM) and Naive Bayes classifiers, and for model evaluation and validation.

##Data Cleaning Using a user-defined function, we check for the missing values and remove them, as the scikitlearn package does not work with missing values.

Data PreProcessing

In preprocessing, the primary libraries used were nltk and string. Here's how they were used:

1. nltk (Natural Language Toolkit):

-nltk provides various tools and resources for natural language processing tasks, including tokenization, stopword removal, stemming, and more.

-Tokenization was performed using nltk's word tokenizer (nltk.word_tokenize()), which splits the text into individual words or tokens.

-Stopwords removal was achieved using nltk's stopwords corpus (nltk.corpus.stopwords), which lists common stopwords in multiple languages.

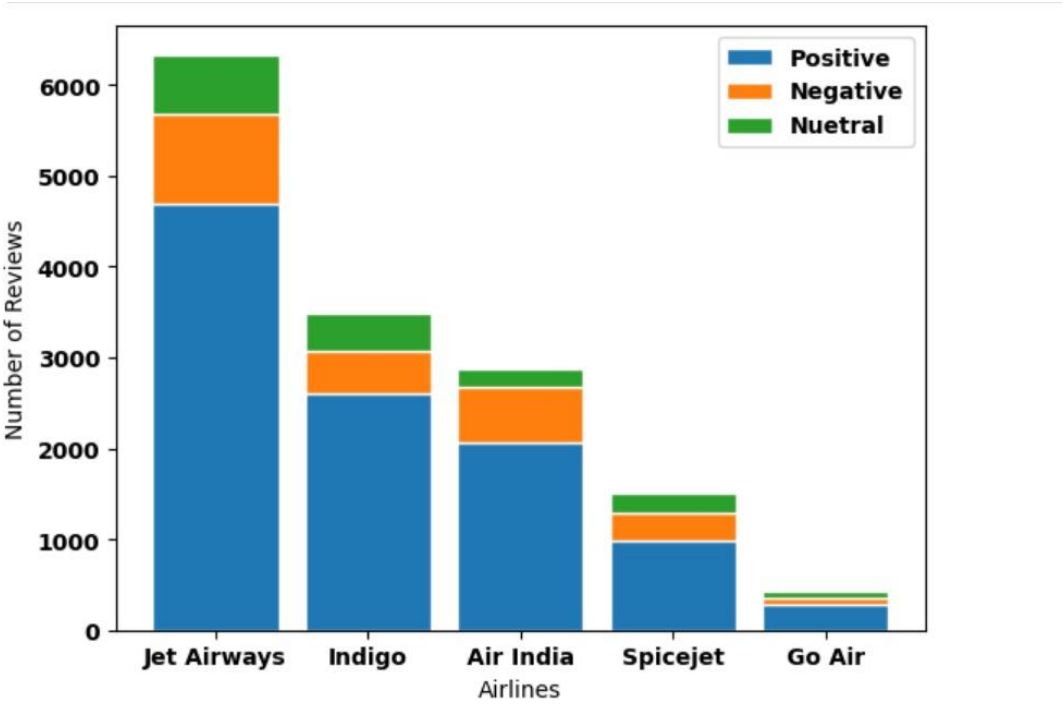
-Stemming was implemented using nltk's Porter stemmer (nltk.stem.porter.PorterStemmer()), which reduces words to their root form.

** -In preprocessing, a user-defined function preprocess is used to remove punctuation marks from the text data.

Data Visualization

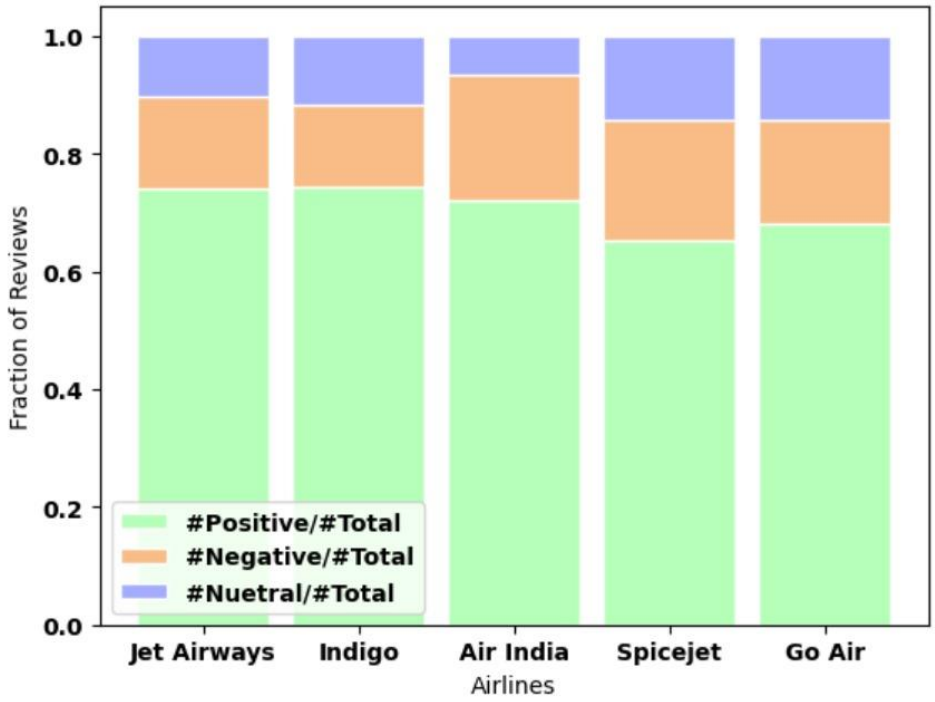
Plot 1

We have plotted subdivided bar plots to show several positive, negative, and neutral reviews for the five airlines using different colors. The number of positive reviews is high for each of the five airlines.



Plot 2

We have observed that the fraction of positive reviews is highest for Indigo and Jet Airways. The fraction of negative reviews is the highest for Air India.



How was SVM used

Support Vector Machines (SVM) were employed as the machine learning model for sentiment classification of customer reviews in this project. SVM was chosen for the following reasons:

1. Classification Task: SVM is well-suited for classification tasks, making it ideal for categorizing customer reviews into positive, negative, or neutral sentiments.

2. High-Dimensional Data: Customer reviews often contain many features (words or tokens), and SVM can effectively handle high-dimensional data, making it suitable for sentiment analysis tasks.

3. Non-linear Separability: SVM can capture complex patterns and relationships in the data, including non-linear relationships between features (words or tokens), which is beneficial for sentiment analysis tasks where relationships may not be linear.

4. Margin Maximization: SVM maximized the margin between the decision boundary and the closest data points, leading to better generalization performance and reduced overfitting. This is crucial for sentiment analysis tasks requiring robust models.

5. Performance: SVM is known for its high accuracy and robustness in classification tasks, making it a suitable choice for sentiment classification in this project.

Use of Naive Bayes

Naive Bayes played a pivotal role in this project for the sentiment classification of customer reviews. Here's why it was chosen:

1. Text Classification Suitability: Naive Bayes is well-suited for classifying customer reviews into sentiment categories (positive, negative, or neutral) due to its effectiveness in text classification tasks.

2. Probabilistic Framework: It operates within a probabilistic framework, estimating the probability of each sentiment class based on observed features (words or tokens) in the reviews, aligning seamlessly with the project's classification needs.

3. Efficiency and Scalability: Naive Bayes offers computational efficiency and scalability, making it suitable for handling large volumes of textual data and crucial for analyzing numerous customer reviews.

4. Robustness to Feature Independence: Despite assuming feature independence, Naive Bayes often performs well, even with noisy or irrelevant review features, ensuring reliable sentiment classification.

5. Baseline Model and Interpretability: It establishes initial performance benchmarks as a reliable baseline model. Its transparent nature allows stakeholders to understand sentiment classification decisions, fostering insight into review analysis.

In summary, Naive Bayes was strategically chosen for sentiment classification due to its alignment with project objectives, efficacy in text classification, computational efficiency, robustness, and interpretability. Leveraging Naive Bayes facilitated accurate sentiment classification, enabling insightful analysis of customer reviews.

How is Linear Algebra and Probability applied here

Linear algebra and probability play essential roles in various aspects of the project, particularly in implementing machine learning algorithms and data analysis. Here's how they are applied:

1. *Linear Algebra*:
    - *Vectorization*: In natural language processing (NLP), text data is often represented as numerical vectors. Linear algebra concepts such as vectors and matrices are fundamental for efficiently representing and manipulating these numerical representations.
    - *Feature Representation*: Techniques like CountVectorizer and TfidfVectorizer, which convert text data into numerical feature vectors, rely on linear algebra operations such as dot products and matrix multiplications. These operations transform raw text data into high-dimensional feature spaces suitable for machine learning algorithms.
    - *Support Vector Machines (SVM)*: SVM is inherently a linear algebra-based algorithm. It works by finding the optimal hyperplane separating different feature space classes. The optimization problem involved in training an SVM model often requires linear algebra techniques such as matrix inversion, dot products, and solving systems of linear equations.
2. *Probability*:
    - *Naive Bayes Classifier*: The Naive Bayes classifier is based on Bayes' theorem, a fundamental concept in probability theory. The classifier assumes that features are conditionally independent given the class label, allowing it to compute class probabilities efficiently.
    - *Sentiment Analysis*: Sentiment analysis tasks involve predicting text data's sentiment (positive, negative, neutral) based on probabilistic models. Probability distributions, likelihood estimation, and Bayesian inference techniques compute the likelihood of different sentiment labels given the observed data.
    - *Model Evaluation*: Probability is crucial in evaluating machine learning models. Performance metrics such as accuracy, precision, recall, and F1-score are computed based on the probabilities predicted by the models. Additionally, cross-validation involves probabilistic sampling and averaging to estimate the model's generalization performance.

Overall, linear algebra and probability form the theoretical foundations of many machine learning algorithms and techniques used in the project. They are essential for

understanding, implementing, and evaluating these algorithms, ultimately contributing to the success of the sentiment analysis task and extracting insights from customer reviews.