National College of Ireland

Predict the Prices of Airfares In India

MSc Research Project
Data Analytics

Vishan Lal
Student ID: x21120803

School of Computing
National College of Ireland

Supervisor: Prof. Cristina Muntean

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Vishan Lal
Student ID:	X21120803
Programme:	MSc in Data Analytics **Year:** 2022-23
Module:	Research Project
Supervisor:	Prof. Cristina Muntean
Submission Due Date:	15th December 2022
Project Title:	Predict Price of Airfares in India
Word Count:	8508 **Page Count:** 22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Vishan Lal

Date: 15th December 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Predict Price of Airfares in India

Vishan Lal
X21120803

Abstract

The number of passengers traveling out using flight in India are increasing, and so are price changes. Seasonal and special event fluctuations are always seen in Indian airfares in multiple periods of the year. The proposed study will explore several data points related to forecasting airfares for different flights in India and use automated methods based on various machine learning models to predict airfares based on multiple characteristics. This research provides various exploratory analysis and insights to the customers which help them purchase their tickets at an optimal cost by exploring various factors, this research also showcases basic and advanced regression models that have been used to accurately predict the price of the airline for the customer. The basic and advanced regression models used in this research are Linear Regression, Decision tree as a regressor, Random Forest as a regressor, XG Boost regressor, K-neighbors regressor, Bagging Regressor, Extra trees regressor, Ridge regression and Lasso regression. Further, the models have been evaluated based on the Mean Average Error (MAE), Root Mean Square Error (RMSE), and adjusted R-Square values.

1 Introduction

Currently, airlines use sophisticated tactics and procedures to allocate ticket pricing in a dynamic manner. These techniques take into consideration a number of financial, marketing, commercial, and societal elements that have a direct impact on the ultimate price of flight. Due to the tremendous complexity of the pricing methods used by airlines, it is very difficult for a passenger to acquire an airline ticket at the lowest price, since the price fluctuates constantly.

This research aims to address about how can Machine learning models like Linear Regression, Decision Tree Regressor, Bagging Regressor, XG Boost Regressor, K neighbours Regressor, Extra Trees Regressor, Ridge Regressor and Lasso Regressor are able to accurately predict the price of the airfare for a specific departure date, and how does flight price in India vary based on various factors that determines the flight price using the different exploratory data analysis techniques to make a better customer experience in booking the flight at the cheapest cost.

Recently, a number of algorithms have been presented that may offer the customer with the optimal time to acquire an airline ticket by forecasting the airfare price (Groves & Gini, 2013). The bulk of these technologies use complex prediction models from the subfield of artificial intelligence research known as Machine Learning (ML).

Specifically, (Groves & Gini, 2013)used the PLS regression model to optimize the purchase of airline tickets with an accuracy of 75.3%. (acc.). Using Ripple Down Rule Learner (74.5% acc.), Logistic Regression (69.9% acc.), and Linear SVM (69.9% acc.), (Papadakis,2014) predicted whether the price of the ticket would reduce in the future as a classification job. (Janssen, 2014) suggested a linear quantile mixed regression model to estimate plane ticket

prices with acceptable performance several days before departure for inexpensive tickets. (Ren, Yang, and Yuan, 2015) investigated the accuracy of Linear Regression (77.02%), Naive Bayes (73.06% accuracy), SoftMax Regression (76.82% accuracy), and SVM (80.6% accuracy for two bins) models in forecasting the price of airline tickets.

The airline industry is particularly vulnerable to the impacts of seasonality. It tends to be greater in the summer compared to the winter since vacationers and leisure travelers like to travel and take vacations during the summer when schools are out of session and the weather is nicer. (Merkert & Webber, 2018). Owing to numerous events that have occurred during the year as well as due to consistent seasonality throughout the year, airlines have also consistently changed their patterns in real time. According to (Puller & Taylor, 2012), the supply and demand dynamics of ticket sales on different days of the week often influence the costs of airlines. By analyzing historical datasets for the variation in travel prices, this study will improve the customer experience while purchasing a journey at a very low cost.

The airfares of Indian airlines varies on a lot of factors, such as, the name of the airline company, City from which the flight takes off, Departure Time of the flight, Airline Stops, Arrival Time of the flight, Destination City, Class, Duration of Flight, month of flight booking, seasonality and Days Left before departure which makes it hard for the passenger to predict the prices of the flight and make a booking at an optimal fare.
The traditional approaches that were being used to determine the flight price prediction failed to predict the prices of the flights for the passengers in order to get the cheapest fare for the passengers. The objective of this research is (1) To determine the movement of the flight prices on various factors, (2) Determine the number of days in advance for a passenger to make the flight booking, (3) Explore various features of the flight pricing data and generate exploratory data analysis, (4) identify the impact of different features and parameters which have the significant impact in driving the variability of flight fares and (5) To predict the flight prices for the passengers so that the passenger can make a booking at the optimal cost.

Additionally, during the COVID-19 pandemic, the seasonality in airfares was seen to have a particularly dynamic character, which shows that real-time attitudes have a substantial influence on the demand for air travelers and, therefore, have an impact on the variability of the airfares (Wozny, 2022).

The outcome of this project was to demonstrate the movement pattern of the flight fares based on various parameters, showcase how machine learning techniques can help in predicting the prices of the flights and produce the distribution of flight price prediction of the flight fares of prior to 50 days before departure. After applying various machine learning techniques, some of the models had produced the predicted prices of the flights which were quite close to the actual prices and also these models produced accurate movement for flight prices prior to 50 days before departure. This research can be efficiently applied in the airlines industry and field and can be used by customers to book their flight tickets at the most optimal cost.

2 Literature Review

Researchers have been attempting to estimate the cost of airfares in recent years by examining several factors that affect flight costs. It's difficult to predict which strategies will be more successful. Many researchers have investigated how passengers behave while making advance

purchases. The goal of performing the literature research is to demonstrate how an existing machine learning algorithms were used to forecast flight ticket pricing behaviour. There were many literature reviews that have been studied in order to carry this research.

2.1 Empirical approach to determine the movement of the airfares and customer behaviour to purchase tickets

According to study by (Puller & Taylor, 2012), price variation based on the day of the week a ticket is purchased is one form of differential pricing employed by airlines. Same-airline, same-route tickets purchased on different days of the week using unique transaction data after adjusting for travel day were compared, ticket limitations, flight demand, and the number of days in advance. It was shown by (Puller & Taylor, 2012) that purchasing tickets on a weekend yields a 5% savings. It is believed that this constitutes a kind of price discrimination. When the mix of consumer purchases leads to more flexible demand on pricing, airlines may offer discounted rates on weekends. If they believe weekend travellers are budget-conscious travellers.

As a consequence of demand shocks and distorting fluctuations in willingness to pay, airfares alter throughout time. (Williams, 2020) claims that a reliable flight pricing model was developed and forecast using recent flight-level data.
Using the model estimates, it was possible to distinguish important relationships between the arrival patterns of different client categories and the lack of capacity as a consequence of demand parametric uncertainty. It has been shown that variable flight pricing boosts output by lowering the prices demanded by price-conscious customers who buy in advance. Furthermore, it ensures seats for those latecomers with the greatest willingness to pay (such as business travellers), who are subsequently charged premium fees. Furthermore, it was shown that dynamic airline pricing increases total welfare in comparison to a more constrained pricing structure.

2.2 Statistical approach used to determine the movement of the airfares

Using artificial intelligence models, (Groves & Gini, 2013) coupled the PLSR (Partial Least Square Regression) model to get the largest presentation for the least cost of purchasing airplane tickets, with an accuracy of 75.3%. (Janssen, 2014) introduced a direct quantile mixed relapse model to forecast the price of airline tickets for inexpensive tickets many days before the day of flight. (Ren, 2015), took into consideration the application of models such as linear regression (with a precision of 77.06%), naive bayes (with a precision of 73.06%), SoftMax regression (with a precision of 76.84%), and support vector machine (with a precision of 80.6%) to estimate the cost of airline tickets. Papadakis (2014) anticipated that the price of the ticket would decrease in the future by accepting the problem as a grouping problem with the assistance of Ripple Down Rule Learner (74.5% exactness.), Logistic Regression with 69.9% precision and Linear SVM with the (69.4% exactness) Machine Learning models.

(Groves & Gini, 2013)used the Partial Least Square Regression (PLSR) technique in order to construct a model that was capable of predicting the ideal time to buy airline tickets. The information was obtained from the most popular online travel booking services between the dates of 22 February 2011 and 23 June 2011. Additional data were also obtained, and these data are now being put to use in order to validate the comparisons of the final model's performances.

(Janssen, 2014)developed an expectation model for the San Francisco to New York route by applying the Linear Quantile Blended Regression technique. The model was constructed using the current daily airfares that were provided by www.infare.com. The model took into account two key aspects, namely the number of days remaining before the departure date and whether or not the flight date falls at the beginning or the end of the week. The model's prediction of airfare is accurate for the days that are quite a ways out from the departure date; but, for a significant amount of time that is relatively near to the departure day, the forecast is not convincing.

This study (Dutta & Santra, 2017) looked at the domestic airline market's dynamic price dispersion. The average power divergence statistic (PDS) and average airfare movement were computed for each route. The impact of Average PDS was assessed based on a few selected route characteristics and market structure parameters. The research (Dutta & Santra, 2017)demonstrates that prices increase as the departure date draws near if both full service airlines and low-cost carriers are present in the same market. Revenue Management and dynamic pricing are often used in the domestic aviation industry in India. Additionally, it suggests that a route's characteristics may affect how and where airfares are distributed.

To investigate the price dispersion of U.S. airlines, (Zhai, 2015) offers a theoretical model for the best design of ticket alternatives under capacity restriction.
The model predicts that as seats become more in demand, price dispersion would fluctuate and finally decline. Original data, which includes seat availability and high-frequency pricing information for differential tickets, reveals that the price difference decreases gradually before becoming smaller as the departure date approaches.

Seasonality and its impact on forecasting are well-known in the aviation service industries.
In order to help with pricing and/or capacity management during periods of strong seasonality, (Merkert & Webber, 2018) suggest a model of rational airline seasonal behaviour. The model was developed, calibrated, and tested using airlines from the Asia-Pacific region, Europe, North America, and South America. The statistics show that seat factors and airline costs fluctuate seasonally. Contrary to logic, seasonal fluctuation in average airfare is greater than that in seat factor. The high-to-low average price ratio need to be less than the high-to-low seat factor ratio. Like airfare, seat factor should be seasonal. Pricing has long been the main topic of competitive strategy and revenue management literature.

Price factors for airline tickets include flight distance, purchasing window, fuel cost, and others. Each carrier has unique algorithms and price-setting standards. These principles may be inferred using recent developments in AI and ML, which can also mimic price movement. An innovative application is provided by (Wang et al., 2019) and is based on the ACS and DB1B public air transportation datasets (T-100). The suggested approach combines the two databases with macroeconomic data, applies machine learning methods, and uses origin and destination pairings to estimate the quarterly average ticket price. With an adjusted R squared of 0.869, the method correctly predicts the testing dataset.

2.3 Supervised Machine learning used to determine the movement of the airfares

The research (Wozny, 2022) examines machine learning predictions for hypothetical scenarios involving air travel. As an example, the impact of COVID-19 policy changes on 2020 airfare was predicted. Studying airfares is crucial since air travel is essential for mobility. Results indicate that approach influences the accuracy of predictions. OLS had a significant prediction

error for counterfactual scenarios, while ML outperformed it. Prior to 2020, forecasts made by ML models were precise. The RMSE of airfares was 5 between 2012 and 2019, fluctuating between USD 242 and 134. The discrepancy between actual and anticipated airfare in 2020 was 25 USD, which was three times more than in previous years. 2020 airfares would be higher without COVID-19.

Uncertain decision-making is a challenge for modern computer-aided approaches and applications. Here, Bayesian prediction techniques are helpful. (Boruah et al., 2019) used the Kalman filter, a Bayesian estimate method, to anticipate flight prices. The linear Kalman Filter model is used in this technique. Based on previous fares, this model predicts future flight prices. The linear model receives observed data as a matrix, then computes an expected ticket price for an upcoming flight.
It is challenging to forecast airline ticket prices and demand since both internal and external factors are subject to rapid change. To assist consumers or airlines in estimating demand, (Abdella et al., 2019) have created alternate ticket price/demand prediction algorithms. Customer-side and airline-side prediction models looked into it.
Both sides' models depend heavily on historical ticket prices, purchase dates, and departure dates. Search engine results and information from social media are not regarded as external components.

Regular attendees will have opportunity to compare prices. The management of airline income involves changing the cost of tickets. If demand outpaces capacity, the airline may raise prices. A specific airway's departure, arrival, and airways were recorded over time to establish the minimum flight rate (Parbat et al., 2021). Hybrid machine learning (ML) models has been utilized to gather information from these factors in order to understand customer perception. Machine learning regression techniques for pricing are presented in this work. This research uses the algorithm and shows how accurate it is (98 percent).

The airline Optimal Ticket Purchasing Decision-Support Service (OTPS) that (Xu & Cao, 2017) suggests identifies the optimal window of opportunity prior to departure. The DPLP technique, which considers variations in airline fare, is the foundation of OTPS.
To increase the generality and reliability of the OTPS, parameter values are route-specific and often modified. The real-world ticket price dataset for many routes is utilized for in-depth analyses. Experiments show that OTPS outperforms other state-of-the-art solutions.

3 Research Methodology

In order to carry out more research and development for this investigation, the research methodology is shown below. The research methodology can be broken down into a total of six distinct stages, each of which is explained in turn below (as shown in Fig 1).

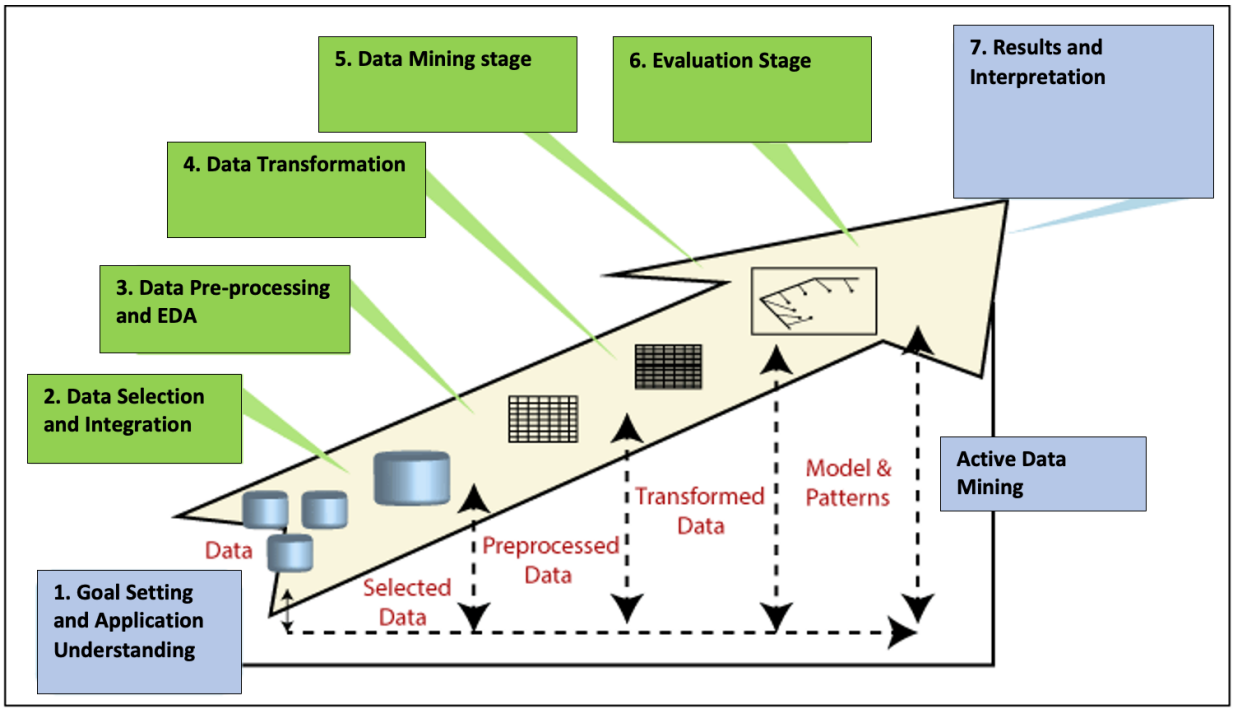

Fig 1: Research Methodology

3.1 Goal Setting and Application Understanding

Variability in the airfares can be analysed effectively by using various Machine Learning techniques. The airfares movement are being judged manually based on the human sentiments based on their experiences, which lacks to take into account the other factors which affect the variability in the airfares and also this manual process takes a lot of time to identify the right price for the flight for a specific departure date for a customer. Recognizing that traditional techniques might, at times, be inaccurate and realizing that using modern ways will result in findings that are more accurate and produced more quickly. This research can enhance in automating the customer's experience to make the booking of the flight at the most optimal cost. The passengers will be able to make flight booking at an optimal cost by getting the predictions and also determine the number of days in advance for a passenger to make the flight booking.
This project also focuses on analysing the variability of the flight fares and detect the dates having the most optimal price for a flight for a specific departure date by using different exploratory data analysis techniques.

3.2 Data Selection and Integration

The dataset is being taken from the Kaggle[1] which is a public and open source repository. The dataset consists of 3 sets of data, the following information describes the information and features of the datasets.

- The first dataset which is named as **economy.csv** comprises of the characteristics of economy class of the flight level data with various features such as date of travel,

[1] https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction?select=Clean_Dataset.csv

Source, Destination of the flight, airline company, time of journey, etc, this dataset consists of 206,774 records.

- The second dataset i.e **business.csv** consist of the characteristics of the business class of the flight level data with the similar features as economy class dataset, this dataset consists of 93,487 records.
- The third dataset is constructed by aggregating multiple features of the economy and the business class dataset.
- The Dataset captures the flight level data over the period of 11th February 2022 to 31st March 2022, i.e. for 50 days.
- The dataset collected consists of the flight level data for the top 6 major metro cities of India and also it captures the data for all the major airline companies of India.
- Totally, 300261 datapoints and 11 features have been taken into consideration in the aggregated dataset.

The datasets that are taken into consideration considers the flight level data for 6 major cities of India which is Delhi, Mumbai, Chennai, Hyderabad, Kolkata and Bangalore, and also the flight level dataset captures the data from 6 major airline service providers which are Vistara, Air India, Go First, Indigo, Air Asia and Spice Jet, among all of these airline companies Vistara and Air India are known to be premium airline service providers and Go First, Indigo, Air Asia and Spice Jet are marked as the low cost airline carriers.

3.3 Data Pre processing and Exploratory Data analysis (EDA)

The dataset have undergone through various steps of pre processing stages, the following steps were performed for the pre-processing of dataset:

- Initially all the 3 sets of data i.e. economy.csv, business.csv and flight_cleaned.csv were read into their respective data frames.
- The Unnamed column was dropped from the flight_cleaned.csv dataset
- All the 3 sets of data were diagnosed for the null values and no null values were found in any of the columns of the 3 datasets.
- The date object was transformed into datetime field in all the 3 datasets and the price object was converted to Integer field in all the 3 sets of data.
- The distribution of the numerical variables were seen by box plots to check for the outlier values in the datasets.
- The day_of_week feature was generated from the date column for all the sets of data to get the weekday on which flight is departing or arriving at any major cities of India.
- After applying the pre-processing steps the Exploratory data analysis were carried out to gain insights from the pre-processed datasets.

Some of the EDA activities that is carried for drawing insights are listed below

3.3.1. Correlation matrix for all the features vs the flight price

From the below correlation matrix (as shown in fig 2) it can be inferred that class of airline and airline brand plays a significant role in driving the prices of then airline tickets. The duration of the flight also has a moderately significant impact in driving the prices of the airline tickets.
On the contrary, the source and the destination city had the least significant impact on the pricing of the airline tickets.

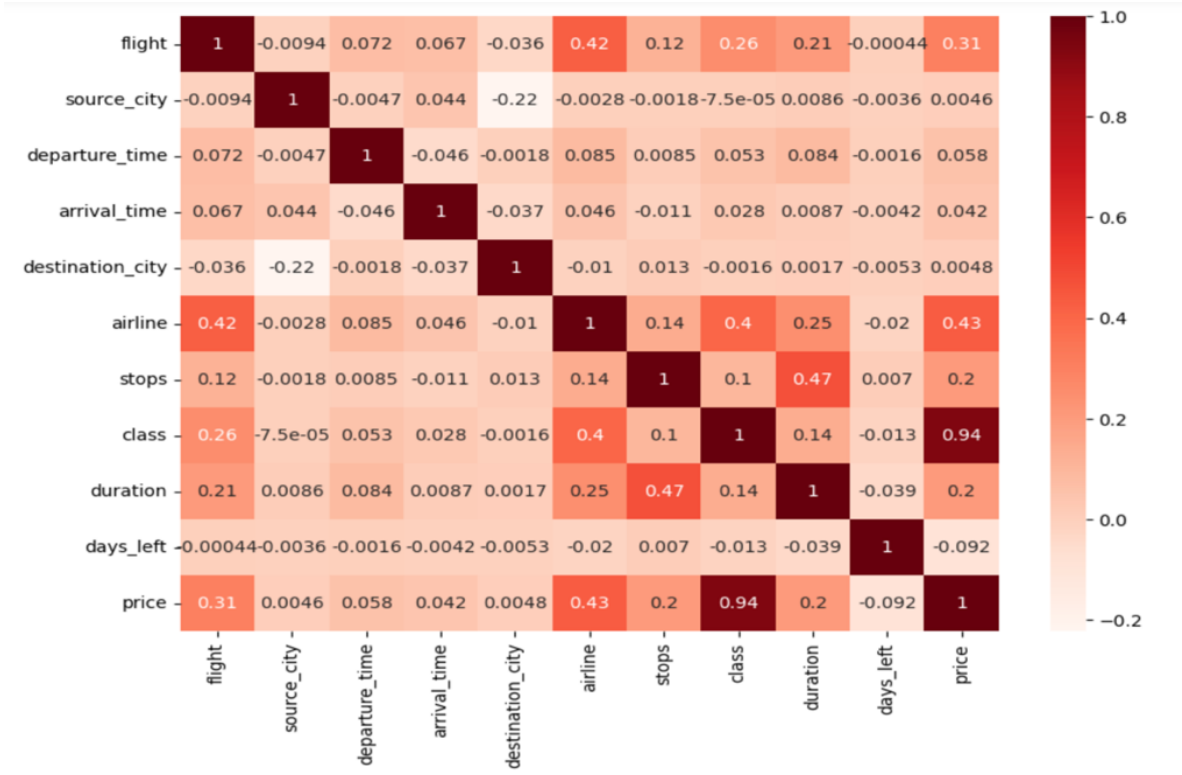

Fig 2: Correlation matrix for all the features of flight dataset

3.3.2. Impact of Major source and destination on the pricing of the airline fare

The below graphs (in fig 3 and fig 4) shows how does the average Airline prices change based on the different source and destination cities. The following can be inferred from the above graphs(fig 3 and fig 4) :

- The source or destination at Delhi remarks to be the cheapest place to depart with an average price less than 20,000 INR(Indian national Rupees) or to arrive there with the average price of 17,000 INR approximately.
- On the contrary, The most expensive place to depart is Kolkata with the average ticket price as 22000 INR and the most expensive place to arrive is found to be Chennai, with the average ticket fare of 23,000 INR.
- The average prices to depart from Chennai is 21,000 INR approximately which is lesser than to arrive at Chennai.
- The average prices to depart and arrive at Mumbai is almost the same with an average price value of 20,500 INR.
- The average price to depart from Hyderabad is 19,000 INR(approximately) which is lesser than the average price to arrive at Hyderabad with an average cost of 20,000 INR.
- The average cost to depart from Bangalore is 21,000 which is slightly lesser than the average price to arrive there which 22,000 INR

These differences and variation in the average prices of the flight might be because of the passenger volume and the airport rentals and tariffs that the flight companies might have to bear in order to keep up there operations in various cities of India.

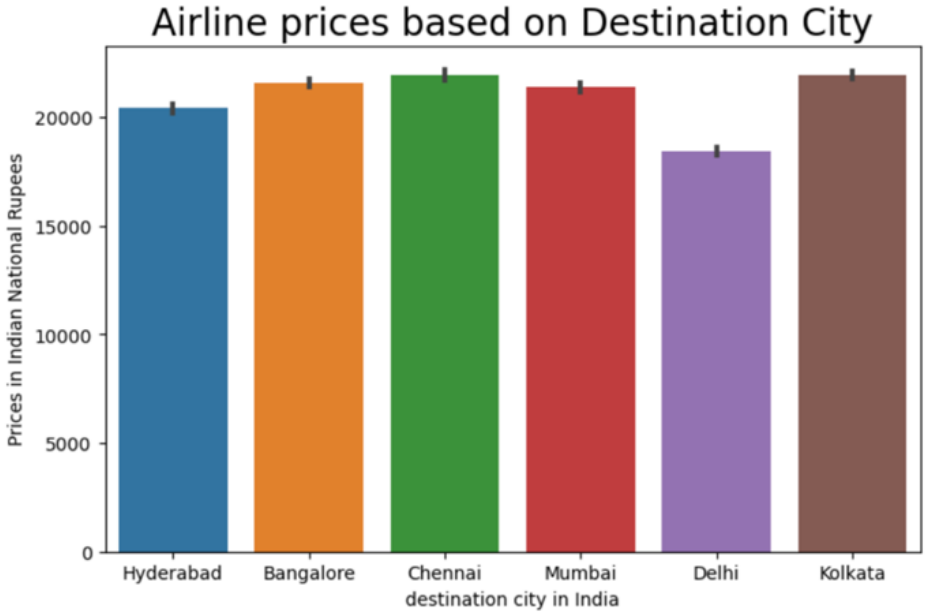

Fig 3: Airline Prices based on Destination cities

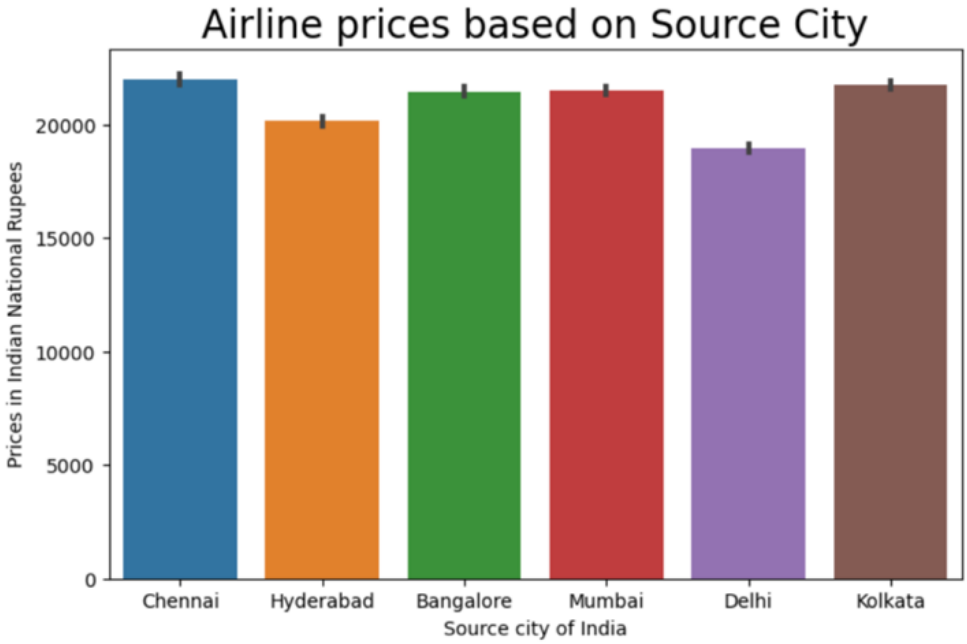

Fig4: Airline Prices based on source cities

3.3.3. Impact of different times of the day on the pricing of the flight to all the major cities of India

The below figures (as shown in fig 5 and 6) showcases the variability of the flight fares on the different times during the arrivals and departures respectively of flight in a day. The following can be inferred from the above figures:

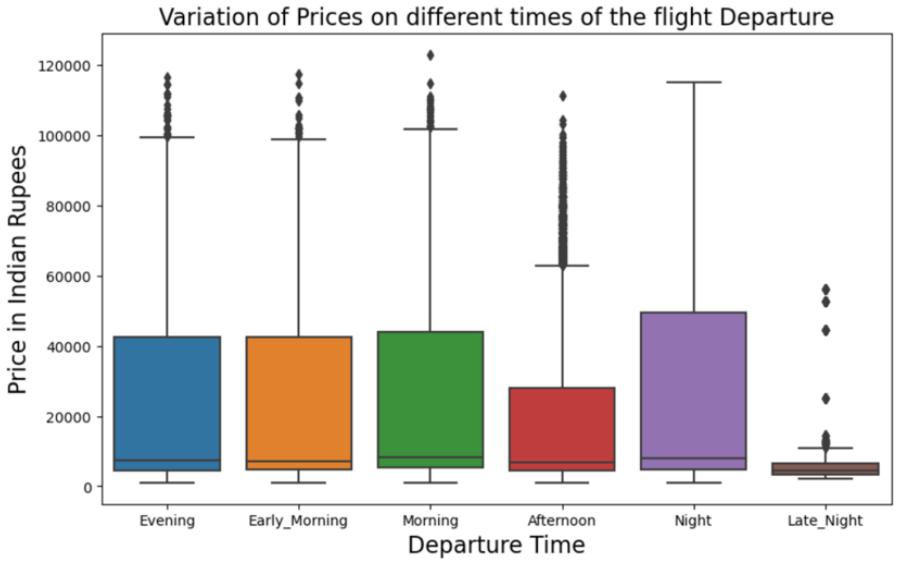

Fig 5: Variation of Prices on different times of the flight Arrivals

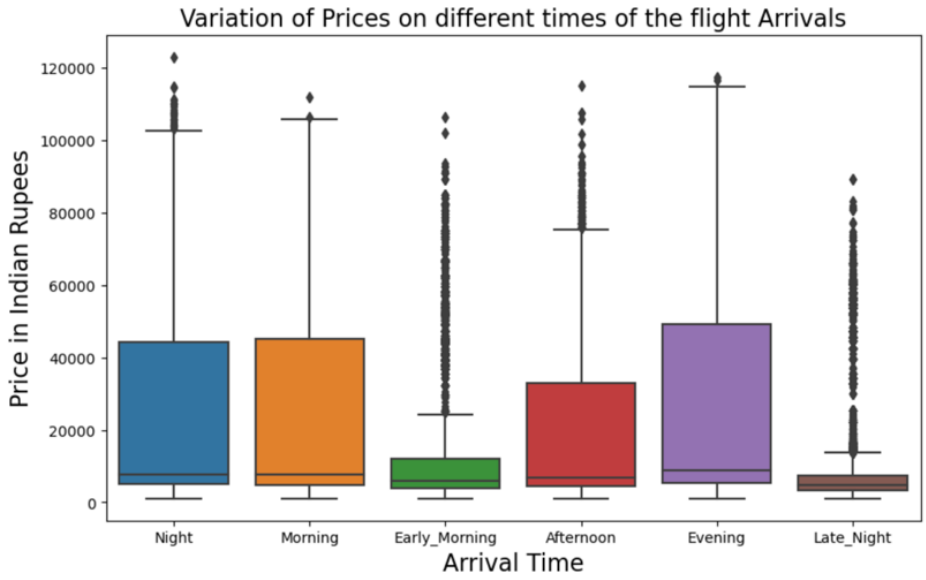

Fig 6: Variation of Prices on different times of the flight Departures

- When flights depart at night, the cost of the tickets increases.
- The cost of a ticket is almost the same for flights with early-morning, morning, and evening departure times.
- Low-cost tickets are available for flights departing late at night.
- When the arrival time is in the evening, the cost of the flight tickets increases.
- Flight tickets cost about the same. Arrival times are in the morning and evening.
- The cost of a ticket is low for flights with late-night arrival times that match departure times.

3.3.4. Impact of the departure date based on the weekday vs the price of the flight

The below figure (as shown in fig 7.) shows that the average price of the flights on all the weekdays remain the same which is approximately around 7,000 Rupees and the maximum prices are the highest on Sunday and Wednesday which 40,000 and 42,000 Rupees respectively.

On the other hand, the maximum prices of the airfares are seen to be lowest for the flights flying on Tuesday which is approximately around 31,000 Rupees. Its more favourable for a passenger to fly on Tuesday as it can save the customer more money.

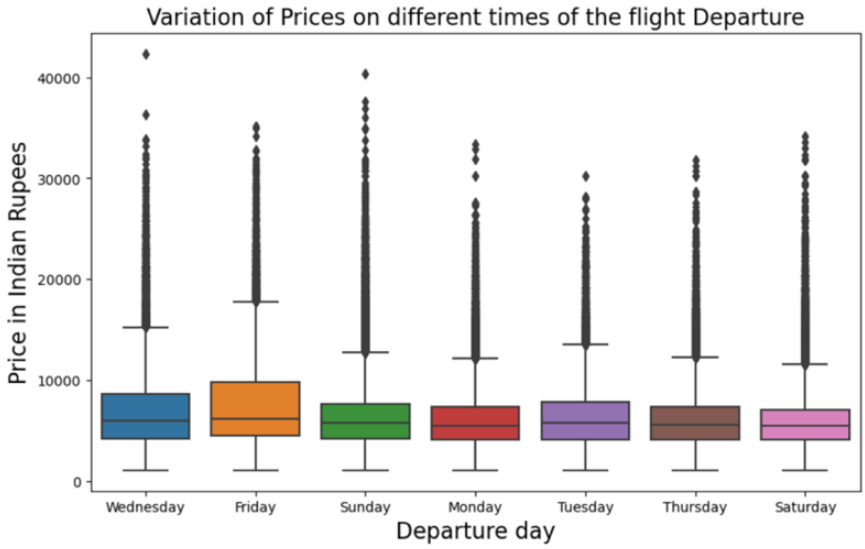

Fig 7: Variation of Prices on different days of the week

3.3.5 Significance of the number of days in advance the tickets were booked vs the price movement of the airfare

The below figure (as shown in fig 8) shows that there is a similar behaviour of the low cost (Go First, Indigo, Air Asia and Spice Jet) and the premium (Air India and Vistara) airlines in terms of number of days before departure and price of the ticket. The ideal time to buy a flight ticket is before 20 days of departure as the price rises significantly after that. When the number of days left before departure are between than 20 and 10 the increased price amounts to 5000 Rupees approximately and 7000 Rupees for the premium airline carriers as compared the flight prices after 20 days of departure.

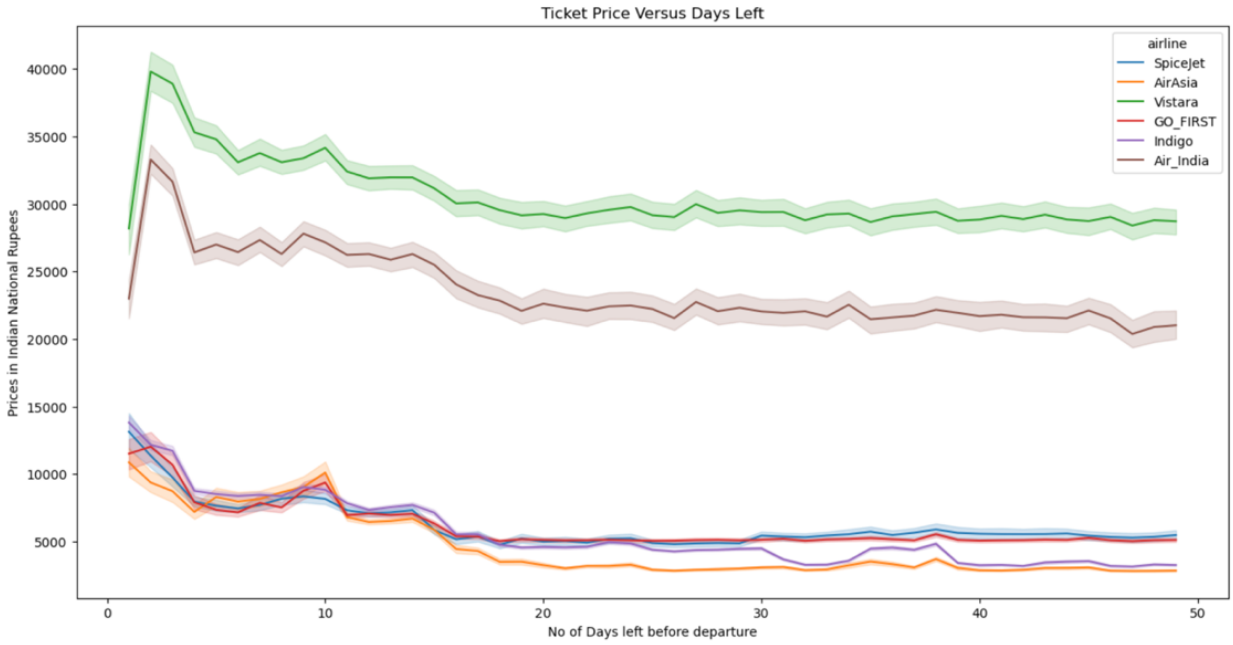

Fig 8: Flight price ticket vs the number of days left

3.3.6 Price movement based on the duration of the flight

It can be inferred from the above figure (as shown in fig 9) that average price of most of the flights are low when the flight duration is between 2-5 hours which amounts to an average price of 7,000 Rupees. The average price of most of the flights are relatively higher when the flight duration is between 15-25 hours which varies between 18,000 to 30,000 Rupees, and flights have the highest variation in the average prices when the duration is between 30-35 hours which ranges between 8,000 to 54,000 Rupees.

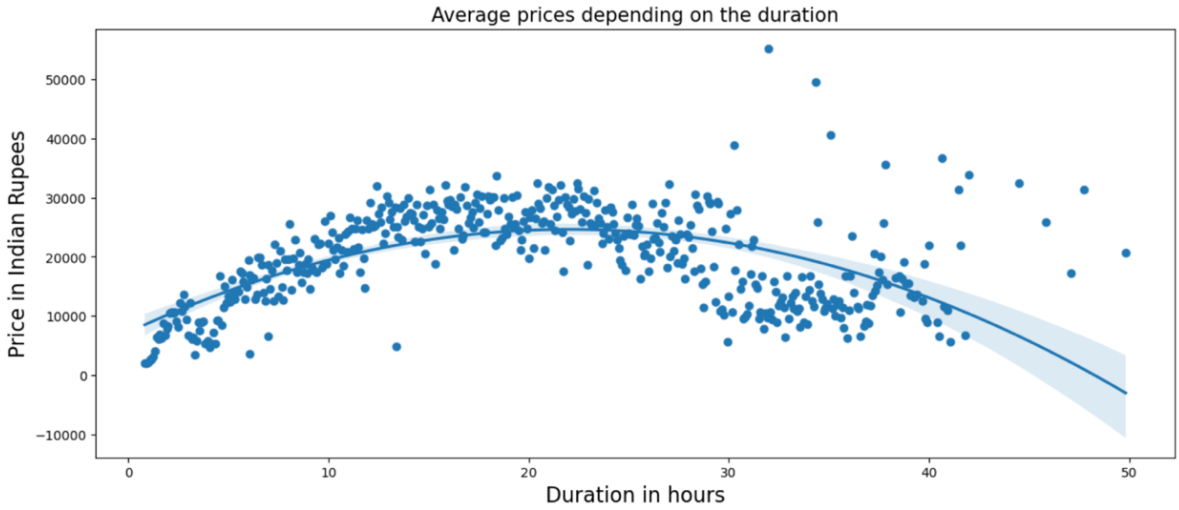

Fig 9: Average prices of the flight based on the duration

3.4. Data Transformation

The process of changing data into the right form needed by mining procedures is known as data transformation.
The below steps are used to transform the dataset to fit into the model:

- Firstly, the dummy variables for all the categorical data in the pre-processed dataset is generated using the one hot encoding technique.
- The correlation matrix is then plotted for all the dummy variables and the original feature with the price of the airline to check the impact of all the variables on the price of the airline.
- Split is made between the training and testing sets in the ratio of 70:30 (i.e. train : 70% and test : 30%) to apply various machine learning models on the pre-processed dataset and after that the target variable i.e the price of the airline is being separated with the features set.
- The scaling Min max scaling transformation was applied to the train and the test sets
- The train and test split sets undergo the min max scaling to have a better fit into the machine learning models

3.5 Data Mining Stage

After the train and test sets of the data are obtained, multiple machine learning models have been applied on the training dataset to predict the flight price. The models implemented to predict the flight price are listed below:

3.5.1 Multiple Linear regression:

A statistical supervised learning method called linear regression uses a linear connection with one or more independent factors to predict the quantitative variable. It assists in determining whether independent variable(s) significantly influences the prediction of the dependent variable, as well as if an independent variable is effective in predicting the dependent variable. The most popular kind of linear regression analysis is multiple linear regression. The multiple linear regression is used as a predictive approach to describe the connection between a continuous dependent variable and two or more independent variables The independent variables may be categorical or continuous (dummy coded as appropriate).
The assumptions that need to be justified in the dataset before computing the multiple linear regression are as follows

- Homoscedasticity: The size of our prediction error is not considerably affected by the value of the independent variable.
- Independence of observations: Observations in the dataset were gathered using statistically acceptable sampling techniques, and there are no hidden correlations between variables.
- Normality: The data follows a normal distribution.
- Linearity: The best-fitting line through the data points is a straight line, as opposed to a curve or other grouping factor.

3.5.2 Decision tree Regressor

One of the most popular and useful methods for supervised learning is the decision tree. Both Regression and Classification problems may be solved with it, while Classification is more often utilized in real-world settings.
It was the way the tree asked the correct questions at the right node in Decision Trees for Classification to provide precise and effective classifications. Entropy and Information Gain are the two metrics used in Classification Trees to accomplish this. However, because we are making predictions about continuous variables, we are unable to compute the entropy and follow the same procedure, for the continuous target variable the mean square error(MSE) is a measurement that indicates how much our projections stray from the initial goal.

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

We only care about how far the forecast deviates from the goal; Y is the actual value, and Y hat is the prediction. not which way around. In order to divide the total amount by the total number of records, we square the difference.
We use the same procedure as with classification trees in the regression tree approach. However, rather than focusing on entropy, we strive to lower the Mean Square Error.

3.5.3 K-Neighbours Regressor

K-Neighbours regression is a non-parametric technique that, by averaging the data in the same neighbourhood, intuitively approximates the relationship between independent variables and the continuous result. The analyst must decide on the neighbourhood's size, or cross-validation may be used to determine the size that minimizes mean-squared error.
To forecast the values of any new data points, the KNN algorithm makes advantage of "feature similarity." In other words, the value given to the new point depends on how much it matches the points in the training set. According to our study, flight ID1 has a duration and a pattern of stops that are comparable to those of flights ID11 and ID10, therefore we can assume that the cost will be around the same as well.

If there had been a classification issue, the mode would have served as the ultimate forecast. Let's use the example of two pricing values in this situation, 7200 and 7700. The final forecast is assumed to be the average of the values.
The algorithm for K-Neighbours regressor works as follows:

- Calculate the Mahalanobis or Euclidean distance between the labelled instances and the query example.
- Rank the samples with labels in order of increasing distance.
- Determine a k of closest neighbours that is heuristically optimum based on RMSE. Cross validation is used to accomplish this.
- Using the k-nearest multivariate neighbours, calculate an inverse distance weighted average.

3.5.4. Extra Trees Regressor

Similar to the random forests technique, the Extra Trees Regression algorithm generates several decision trees, but the sampling for each tree is random and without replacement. As a result, each tree gets its own dataset with distinct samples. For each tree, a certain amount of features from the whole collection of features are also randomly chosen. The extra tree class implements a meta estimator that employs averaging to increase prediction accuracy and reduce overfitting. The meta estimator fits a number of randomized decision trees (also known as extra-trees) on different sub-samples of the dataset.

3.5.5 XG boost and Gradient Boosting as a regressor

A strong method for creating supervised regression models is XGBoost. By understanding this statement's (XGBoost's) objective function and base learners, one may deduce its correctness. A regularization term and a loss function are both part of the objective function. It provides information on the discrepancy between actual and expected values, or how much the model's predictions depart from the actual values. Reg:linear and reg:logistics are the two most used loss functions in XGBoost for regression and binary classification, respectively. XGBoost is one of the ensemble learning techniques, which entails training and integrating several independent models (sometimes referred to as base learners) to produce a single prediction. In order for poor predictions to cancel out and better ones to add up to final positive predictions, XGBoost anticipates having base learners that are consistently awful at the rest.

The gradient boosted trees approach is implemented using the open-source software known as XGBoost, which stands for extreme gradient boosting. It is a supervised learning technique that may be used to classification or regression problems.
Boosting is an ensemble approach, which means it combines predictions from several models into a single forecast. It does so by modelling each prediction progressively dependent on the mistake of its previous. The following steps are to be taken to implement the XG boost and gradient boosting:

- Create a first model utilizing the initial information.
- Using the residuals from the first model, fit a second model.
- Create a third model using the sum of models 1 and 2

3.5.6. Bagging Regressor

An ensemble meta-estimator known as a bagging regressor fits base regressors to individual random subsets of the original dataset, and then combines each prediction (either by voting or by averaging) to get the final prediction. By adding randomization to the process of building a

black-box estimator (such a decision tree), a meta-estimator of this kind can often be used to lower the variance of the estimator.

3.5.7 Ridge regressor

By establishing a ridge regression estimator, ridge regression was devised as a potential remedy to the imprecision of least square estimators when linear regression models contain certain multicollinear (highly correlated) independent variables (RR). Given that its variance and mean square estimator are often lower than the least square estimators previously computed, this gives a more accurate ridge parameters estimate.
In the simplest scenario, the near-singular moment matrix issue is solved by increasing the number of positive diagonal elements, which lowers the condition number of the matrix display style (XTX).

3.5.8. Lasso Regressor

The L1 regularization process used by Lasso regression results in a penalty proportional to the absolute size of the coefficients. A sparse model with few coefficients may be produced by this kind of regularization; certain coefficients may go to zero and be removed from the model. Greater penalties lead to coefficient values that are closer to zero, which is great for creating more straightforward models.
The goal of the algorithm is to minimize:

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

3.6 Evaluation Stage

After all the models were built the metrics such as RMSE, MAE, Adjusted R-Square were evaluated and compared for all the models. Graphical analysis on the performance of all the models were obtained and the model with lowest error terms was taken into consideration to fit the test data and get the flight price prediction on the testing data.
The evaluation metrics are explained below:

3.6.1 MAE(Mean absolute error)

The average of the absolute difference between the dataset's actual and anticipated values is represented by the Mean Absolute Error. It calculates the dataset's residuals' average.

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}|$$

Where,
$\hat{y}$ – *predicted value of y*
$\bar{y}$ – *mean value of y*

3.6.2. Mean Squared Error (MSE)

The mean of the squared difference between the data set's original and forecasted values is known as the mean squared error. It calculates the residuals' variance.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2$$

3.6.3. Root mean squared error (RMSE)

The square root of Mean Squared Error is called Root Mean Squared Error. It calculates the residuals' standard deviation.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

3.6.4 R-Squared and adjusted R- Square

The percentage of the dependent variable's variation that the linear regression model can explain is shown by the coefficient of determination, also known as R-squared. Since the score is scale-free, it will always be less than one regardless of how big or tiny the numbers are.

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

Adjusted,The modified form of R square, will always be less than or equal to R2, and it is adjusted for the number of independent variables in the model. The numbers n and k in the formula below represent the number of observations and independent variables, respectively, in the data.

$$R^2_{adj} = 1 - \left[\frac{(1-R^2)(n-1)}{n-k-1}\right]$$

4. Design Specifications

In order to successfully predict the prices of the airfares it becomes important to comprehend the description of the design approach that I will be following to predict the cost of the air fares. The airfare prediction starts by collecting the data from Kaggle which is a public open source repository and then followed by the pre-processing various features of the data and after getting the pre-processed data, the data is split into train and testing sets and then various regression based models are applied on the training data, followed by the testing of the model on the test data and finally a predicted output of flight price is generated for the test data. The above diagram (as shown in fig 10) explains about the same.

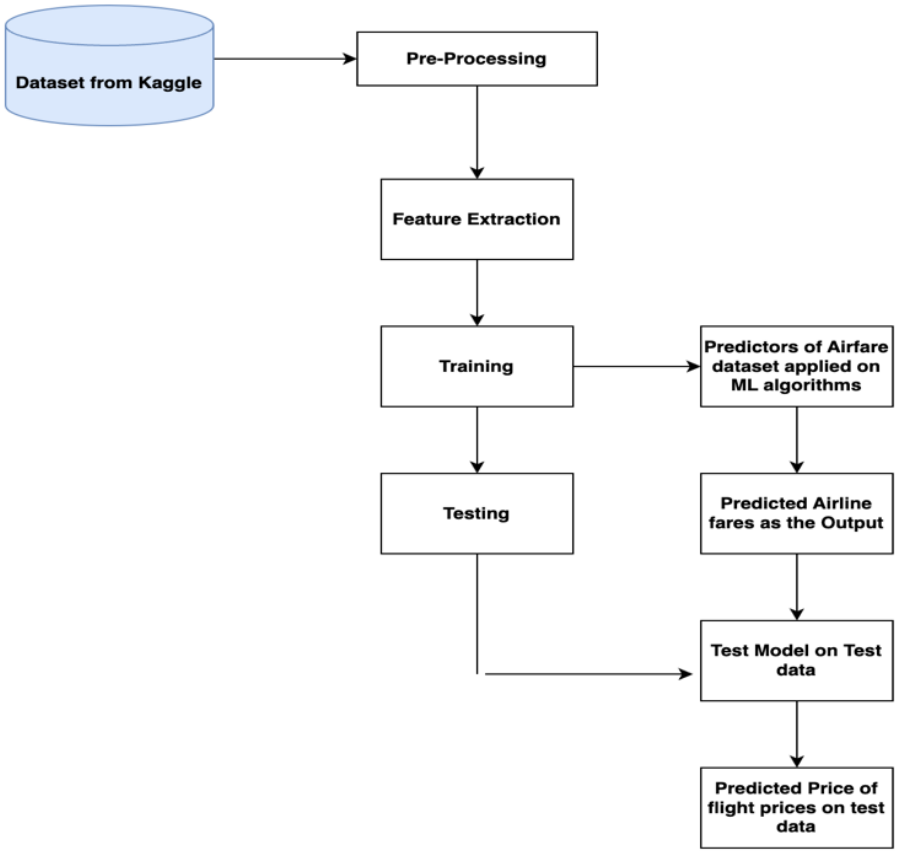

Fig 10: Design specification for predicting the airfares using ML models

5. Implementation

The models described in the methodology have been implemented as follows to predict the flight prices in India.

5.1. Linear Regression

The following parameters were tuned which are also the default values of the parameters in order to implement the linear regression model to predict the prices of flight tickets:

- Fit_intercept= True, this means that, whether the model's intercept should be calculated. If False, the intercept will not be used in computations.
- Copy_X=True,
- N_jobs=None
- Positive=false, This means that it will force the coefficients to be positive.

5.2. Decision Tree Regressor

The following parameters were tuned which are also the default values of the parameters in order to implement the Decision Tree regression model to predict the prices of flight tickets:

- Criterion=Squared Error, The function that is used to evaluate a split's quality.
- Splitter=best, To find the method for selecting the split at each node.
- Max_depth=None, The maximum depth of the tree

- Min_samples_split=2, Minimum necessary number of samples to split an internal node
- Max_features=None, The amount of features to consider while searching for the ideal split
- random_state=None
- max_leaf_nodes=None

5.3 K-Neighbours Regressor

The implementation parameters that were considered in building K-Neighbours model is described as follows, the n_neighbours value was taken to be 5 which is the default value , the weights that were chosen was taken to be uniform which means, Each neighbourhood's points are given equal weight, the algotithm on the auto mode that means, will make an effort to choose the best algorithm based on the values supplied to the fit method, the leaf_size parameter was taken to be 30, this may impact how quickly the tree is built and queried as well as how much memory is needed to store the tree.

5.4 Extra Trees Regressor

For the implementation of the extra tree regressor model the following model parameters were taken:

- n_estimators : The number of trees in the forest. n_estimators=100 was taken to build the model which is the default value.
- criterion **:** The feature that assesses a split's quality. "squared error" is a supported criterion for the mean squared error. Criterion=”squared_error” is taken to build the model, which is the default value
- max_depth: The maximum depth of the tree. Max_depth=None is considered for this model which is the default value.
- min_samples_split: The minimum number of samples required to split an internal node. min_samples_split=2 was taken which is the default value.

5.5 XGBoost and Gradient Boosting Regressor

This estimator allows for the optimization of any differentiable loss function and constructs an additive model in a forward stage-wise manner. A regression tree is fitted on the negative gradient of the provided loss function at each level.The following parameters have been tuned for the implementation of XG boost and gradient boosting, these values are also the default values for the implementation of XGBoost model:

- max_depth=4, this restricts the number of tree nodes.
- Loss= ‘squared_error’
- Learning_rate=0.1, by how much the contribution of each tree will decrease
- N_estimators=100, The number of steps of boosting that will be executed.
- Sub_sample=1.0

5.6 Bagging regressor

This algorithm incorporates various literary works. Pasting is the name of the algorithm used when random subsets of the dataset are used as random subsets of the samples. Samples drawn using replacement are known as "bagging" samples. The technique is referred to as "Random Subspaces" when random subsets of the dataset are created as random subsets of the features. Last but not least, the technique is referred to as Random Patches when base estimators are constructed on subsets of both samples and features.

The following parameters have been tuned for the implementation of Bagging regressor, these values are the default parameter values for Bagging regressor:

- Estimator: The base estimator to fit on random subsets of the dataset, the estimator value chosen is None.
- n_estimators: The number of base estimators in the ensemble. The value of n_estimators is taken to be 10.
- Max_samples= The number of samples to draw from X to train each base estimator. The value of max_sample is taken to be 1.0
- Max_features= The number of features to draw from X to train each base estimator. The value of max_features is taken to be 1.0

5.7 Lasso regression

A tuning parameter, λ controls the strength of the L1 penalty. λ is basically the amount of shrinkage:

- When λ = 0, no parameters are eliminated. The estimate is equal to the one found with linear regression.
- As λ increases, more and more coefficients are set to zero and eliminated (theoretically, when λ = ∞, *all* coefficients are eliminated).
- As λ increases, bias increases.
- As λ decreases, variance increases.

The following parameters that are the default parameter values have been tuned for the implementation of Lasso regressor:

- Alpha=1.0, A constant multiplied by the L1 term that controls the regularization strength.
- Fit_intercept=True, Whether the model's intercept should be calculated.
- Precompute=False, Whether or not to utilize a precomputed Gram matrix to accelerate computations
- Max_iter=1000

5.8 Ridge regression

This model resolves a regression model where the regularization is provided by the l2-norm and the loss function is the linear least squares function. sometimes referred to as Tikhonov regularization or Ridge Regression. Multi-variate regression is built-in support for this estimator.
The following parameters have been tuned for the implementation of Ridge regressor which are the default parameter values for the model:

- Alpha=1.0, A constant multiplied by the L2 term that controls the regularization strength
- Fit_intercept=True, Whether the model's intercept should be calculated.
- Copy_x=True
- Max_iter=None

6. Results

The Results and evaluation metrics for the models that have been generated for predicting the price of the flight on the training data are shown in the table figure below (shown in fig 11):
From the below table figure (fig 11) of results the following can be inferred:

- **XG boost** regressor produced the best set of results in training data with Adjusted R-Square value of 0.9845, MAE as 1111.31, RMSE as 2818.11 and MSE as 7.941 x 10^6 which means that the predicted prices vary by the factor of 0.9845 from the original prices and the average deviation for the predicted prices are deviated by the value of 1111.31 INR(Indian National Rupees).

	Model_Name	Adj_R_Square	Mean_Absolute_Error_MAE	Root_Mean_Squared_Error_RMSE	Mean_Squared_Error_MSE
0	XGBRegressor	0.984567	1111.313984	2818.111048	7.941750e+06
1	KNeighborsRegressor	0.982006	1209.984116	3043.043307	9.260113e+06
2	ExtraTreesRegressor	0.978741	1819.477614	3307.636674	1.094046e+07
3	DecisionTreeRegressor	0.973809	1247.902298	3671.225568	1.347790e+07
4	GradientBoostingRegressor	0.970023	1881.169731	3927.696573	1.542680e+07
5	Lasso Regression	0.957233	2785.167610	4691.360346	2.200886e+07
6	BaggingRegressor	0.906950	4572.481049	6919.907473	4.788512e+07
7	Ridge Regression	0.906950	4572.293761	6919.913344	4.788520e+07
8	LinearRegression	0.906949	4571.189881	6919.960632	4.788586e+07

Fig 11: Model results and metrics for computing flight prices from training data

- the **K-neighbours** model was the second best performing model that also had almost similar metrics to XGBoost but were slightly low than XG Boost having the Adjusted **R-Square** as 0.9820, **MAE** as 1209.98, **RMSE** as 3043.04 and **MSE** as 9.2 x10^6, which means that the predicted prices vary by the factor of 0.9820 from the original prices and with R-square value as 0.9820, so it can be inferred that the predicted and the actual values are very close, and the average deviation for the predicted prices are deviated by the value of 1208.98 INR.

- The **Extra Trees regressor** model was the third best performing model with the adjusted R-Square value of 0.9787, MAE as 1819.47, RMSE as 3671.22 and MSE 1.09 x 10^7, which means that the predicted prices vary by the factor of 0.9787 from the original prices, so it can also be inferred that the predicted and the actual values are very close and the average deviation for the predicted prices are deviated by the value of 1819.98 INR, which is significantly higher than the K-neighbour regressor.

- The **Decision Trees regressor** model was the fourth best performing model with the adjusted R-Square value of 0.9738, MAE as 1247.90, RMSE as 3307.63 and MSE 1.34 x 10^7, this suggests that the decision tree regressor have better MAE metrics than the Extra tree regressor, which means that the average price predicted price variation for decision tree regressor is lower than the Extra tree regressor. This also means that the predicted prices vary by the factor of 0.9738 from the original prices, so it can also be inferred here that the predicted and the actual values are very close and the average deviation for the predicted prices are deviated by the value of 1247.90 INR, which is significantly lower than the Extra Tree regressor.
- The **Gradient Boosting regressor** model was the fifth best performing model with the adjusted R-Square value of 0.9572, MAE as 1881.16, RMSE as 3927.69 and MSE 1.54 x 10^7, which means that the predicted prices vary by the factor of 0.9700 from the original

prices, the prediction metrics were slightly lower than the Decision tree regressor and the Extra trees regressor model, it can also be inferred that the predicted and the actual values are very close and the average deviation for the predicted prices and the prediction metrics have a are deviated by the value of 1881.98 INR, which is significantly higher than the Decision tree regressor and slightly higher than the Extra Tree Regressor.

- The **Lasso regressor** model was the sixth best performing model with the adjusted R-Square value of 0.9572, MAE as 2785.16, RMSE as 4691.36 and MSE 2.20 x 10^7, which means that the predicted prices vary by the factor of 0.9572 from the original prices, this means that the R-squared is noticeably lower than the top 5 models that have been explained above so it can also be inferred that the predicted and the actual values are relatively less close than the top 5 models and the average deviation for the predicted prices are deviated by the value of 2785.98 INR, which is significantly higher than the top 5 models presented above.

- The **Bagging and the Ridge regressor** model was the seventh best performing model, these 2 models have almost similar metrics with the adjusted R-Square value of 0.9069, MAE as 4572.48, RMSE as 6919.91 and MSE 4.78 x 10^7, which means that the predicted prices vary by the factor of 0.9069 from the original prices, this means that the R-squared is significantly lower than the top 6 models that have been presented above so it can also be inferred that the predicted and the actual values of prices are significantly less close than the top 5 models and the average deviation for the predicted prices are deviated by the value of 4572.48 INR, which is significantly higher than the top 6 models presented above.

- The **Linear regression** model had a lowest accuracy in predicting the flight prices as the adjusted R-Square was 0.9069, MAE was 4571.18, RMSE as 6919.96 and the MSE as 4.78 x 10^7 , which means that the predicted prices vary by the factor of 0.9069 from the original prices, this means that the R-squared is significantly lower than the top 6 models that have been presented above so it can also be inferred that the predicted and the actual values of prices are significantly less close than the top 6 models and the average deviation for the predicted prices are deviated by the value of 4571.18 INR, which is significantly higher than the top 6 models presented above and slightly lower than the Bagging and Ridge regressor.

Since the XG boost model had the best evaluation metrics in predicting the flight price, so the model was then fitted on the test data to get the flight price predictions on the test data. The below figure (as shown in fig 12) showcases the predictions of flight prices on the test data and plotted against the actual values of the flight prices. The test data was seen to fit well on the XG Boost model and the predicted prices were almost matching the actual price values of the flights.
Furthermore, the actual and predicted prices found on the test data of the flights were plotted with respect to the number of days left for the departure and the price of the flight (as shown in fig 13). The predicted price before the number of days left for departure almost matched the predicted values, which were found by fitting XGBoost model on the test data.

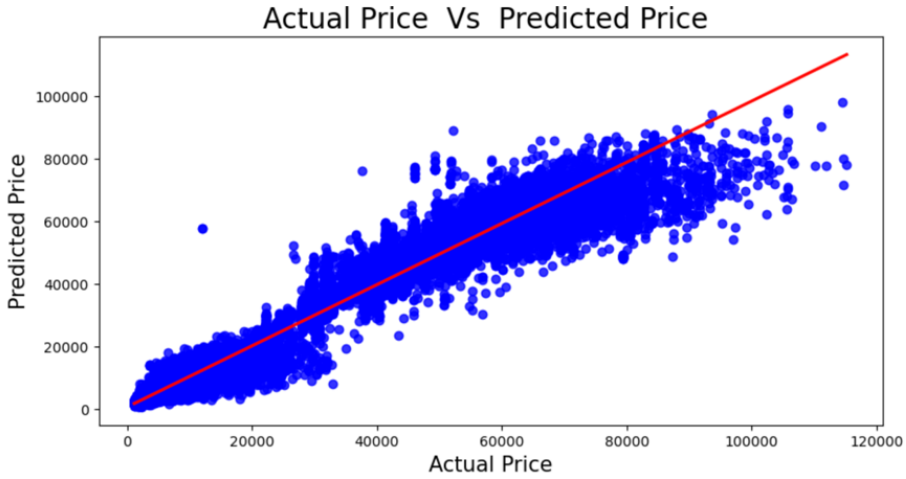

Fig 12: Actual vs the predicted prices of the flights

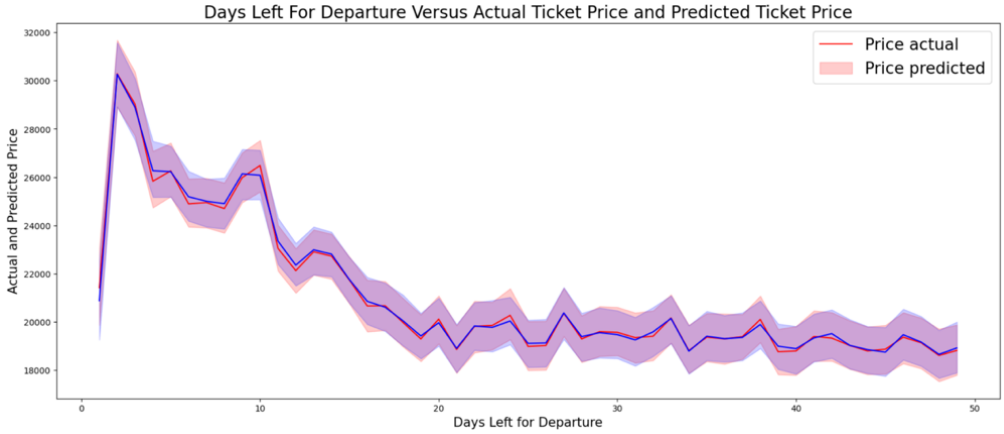

Fig 13: Days Left For Departure Versus Actual Ticket Price and Predicted Ticket Price

7. Discussions

This project was carried out step by stage, and it was difficult to identify a good dataset and to extract it, clean it, and convert it. After pre-processing the dataset, one hot encoding for the categorical variable, and scaling the data using the min max scaling transformation, the project's goal was achieved. Understanding the purpose of the variables in the dataset and identifying the relevant factors were the main challenges in the pre-processing of the data.

The Exploratory data analysis was performed after generating the pre-processed dataset which resulted in generating better insights than the previous work that has been done in predicting the airfares by (Liu et al., 2017) and gathering more information about the features that affect the variability of the airfares such as time of flight, the weekday of the departure, duration of the flight.

The number of days left before departure, timings of the day at which the flights are departed and arrived, the weekday on the flight is scheduled and the duration of the flight plays a vital role for determining the price of the flight price.

The case studies using a variety of sophisticated regression machine learning models have outperformed the earlier models in the literature study. With an adjusted R-squared value of 0.9845, this research produced the best prediction metrics to predict the price of the flight using the XG Boost model. This is significantly better than the research conducted by (Wang et al.,

2019), which produced the best adjusted R-squared value of 0.858 using the Random forest model. By accurately forecasting a customer's ideal flight cost, this study will have a significant positive impact on the airline industry's passengers and improve the possibility that they will make more purchases of airline ticket at the most optimal cost.

Conclusion and Future work

The flight price prediction is an important contribution various passenger who plan to book the flights at the most optimal cost. The literature work for the various studies relating to the flight price prediction was thoroughly studied and applied in this research work. All the steps of KDD(knowledge discovery of data) were studied and applied as part of this research work to predict the flight prices in India.
The application and background was initially understood followed by the understanding of the data and then data pre-processing steps were performed on the data. The exploratory data analysis(EDA) was then implemented to gain various insights from the flight pricing patterns based on various features. After performing the EDA the data transformation and scaling was done followed by understanding and implementing various regression models to predict the prices of the flight. The results were then evaluated using important metrics and then followed by testing of the best model on the test data to predict the price of flight was done.

The flight price prediction is an important contribution various passenger who plan to book the flights at the most optimal cost. This research work can have multiple outcomes, the following are listed below:

- This can fit in well with the customers who cannot foresee the flight prices pattern and are always looking to purchase the airline ticket at the lowest cost possible.
- This research work can contribute in the field of airline industry to help the customer to get the right date before the departure to make the flight booking at the lowest cost possible.
- The first 2 points can be used effectively by the major airline ticket agents to leverage this research to acquire more customers and offer them to book the airline ticket at the lowest prices, eg: Skyscanner, Expedia.

There were few limitations that can be taken care as part of the future work for predicting the prices of the flights. The limitations are as follows:

- The dataset contains record values for 50 days in a year, which limits the research to generate pattern of the price movement of airfares in a limited amount of time over the year.
- The seasonality in airfares can’t be interpreted well for the entire year.
- The model computation might take significantly more time for processing more than 10 million airline records to predict the airfares and it will require distributed computing resources to overcome this limitation.
- The dataset consists the airline data of 6 major cities of India, having more cities of India in the dataset can generate the prediction of airfares for other Indian cities.

REFERENCES

Abdella, J. A., Zaki, N., & Shuaib, K. (2019). Automatic Detection of Airline Ticket Price and Demand: A review. *Proceedings of the 2018 13th International Conference on Innovations in Information Technology, IIT 2018*. https://doi.org/10.1109/INNOVATIONS.2018.8606022

Boruah, A., Baruah, K., Das, B., Das, M. J., & Gohain, N. B. (2019). A Bayesian approach for flight fare prediction based on kalman filter. *Advances in Intelligent Systems and Computing*, *714*. https://doi.org/10.1007/978-981-13-0224-4_18

Dutta, G., & Santra, S. (2017). An empirical study of price movements in the airline industry in the Indian market with power divergence statistics. *Journal of Revenue and Pricing Management*, *16*(2). https://doi.org/10.1057/rpm.2016.12

Groves, W., & Gini, M. (2013). An agent for optimizing airline ticket purchasing. *12th International Conference on Autonomous Agents and Multiagent Systems 2013, AAMAS 2013*, *2*.

Janssen, T. (2014). A Linear Quantile Mixed Regression Model for Prediction of Airline Ticket Prices. *Cs.Ru.Nl*.

Liu, T., Cao, J., Tan, Y., & Xiao, Q. (2017). ACER: An adaptive context-aware ensemble regression model for airfare price prediction. *Proceedings of 2017 International Conference on Progress in Informatics and Computing, PIC 2017*. https://doi.org/10.1109/PIC.2017.8359563

Merkert, R., & Webber, T. (2018). How to manage seasonality in service industries – The case of price and seat factor management in airlines. *Journal of Air Transport Management*, *72*. https://doi.org/10.1016/j.jairtraman.2018.07.005

Parbat, T., Benhal, R. S., & Jain, H. (2021). Understanding the Customer Perception Using Machine Learning while Booking Flight Tickets. *Proceedings of the 3rd International Conference on Inventive Research in Computing Applications, ICIRCA 2021*. https://doi.org/10.1109/ICIRCA51532.2021.9544550

Puller, S. L., & Taylor, L. M. (2012). Price discrimination by day-of-week of purchase: Evidence from the U.S. airline industry. *Journal of Economic Behavior and Organization*, *84*(3). https://doi.org/10.1016/j.jebo.2012.09.022

Wang, T., Pouyanfar, S., Tian, H., Tao, Y., Alonso, M., Luis, S., & Chen, S. C. (2019). A framework for airfare price prediction: A machine learning approach. *Proceedings - 2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science, IRI 2019*. https://doi.org/10.1109/IRI.2019.00041

Williams, K. (2020). Dynamic Airline Pricing and Seat Availability. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3611696

Wozny, F. (2022). The Impact of COVID-19 on Airfares—A Machine Learning Counterfactual Analysis. *Econometrics*, *10*(1). https://doi.org/10.3390/econometrics10010008

Xu, Y., & Cao, J. (2017). OTPS: A decision support service for optimal airfare Ticket Purchase. *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017*, *2018-January*. https://doi.org/10.1109/BigData.2017.8258068

Zhai, Y. (2015). Selling Tickets as Options: Airline Price Dispersion with Limited Seat Capacity. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2663693

Ren, R., Yuan, S. and Yang, Y. (2015) *Prediction of airline ticket price - cs229.stanford.edu*. Available at: http://cs229.stanford.edu/proj2015/211_report.pdf

Papadakis, M. (2014) *Predicting airfare prices - Stanford University*. Available at: https://cs229.stanford.edu/proj2012/Papadakis-PredictingAirfarePrices.pdf