Dissertation Submitted for the partial fulfillment of the **B.Sc. as a part of M.Sc. (Integrated) Five Years Program AIML** degree to the Department of AIML & Data Science.

**Project Dissertation**

**Flight Ticket Price Prediction and Deployment**

submitted to

****

By

**Lathiya Jay Sureshbhai**

**Semester-VI**

**M.Sc. (Integrated) Five Years Program AIML**

Department of AIML & Data Science.

School of Emerging Science and Technology

Gujarat University

**June - 2022**

**DECLARATION**

This is to certify that the research work reported in this dissertation entitled “**Flight Ticket Price Prediction and Deployment**” for the partial fulfilment of B.Sc. as a part of M.Sc. (Integrated) in Artificial Intelligence and Machine Learning degree is the result of investigation done by myself.

|  |  |
| --- | --- |
| **Place:** Ahmedabad | **Name of Student** |
| **Date:** 6th June 2022 | Jay S. Lathiya |

**ACKNOWLEGEMENT**

On this occasion of submitting my research work, I would like to thank all the people who have made this possible. The role of Department of Artificial intelligent, Data Science and Actuarial Science, Gujarat University in shaping my professional capability is very important.

I thank **Dr. Ravi Gor**, co-coordinator of the Department of Artificial intelligent and Data science, for all his support. He has always supported me and encouraged me to overcome many difficulties. I am extremely grateful to him for his valuable

suggestions and guidance he has provided me during my research work.

**Ms. Rashmi Kumari Panday**, **Mr. Eric Shah** and **Mr. Dilip Dubey** my guide has always guided me on the right path and has been a motivating source for innovative research work. This would have not been possible without his guidance, support, and encouragement. With the help of his guidance, I magnificently overcame the difficulties during my work and leamed at each step.

lam heartily and loving thankful to my **parents** for their sacrifice, support and endless love which encouraged me to efficiently overcome the difficulties in my pursuit of the research and who always helped me in my work.

Last but not the least, I would not be able to complete the work without the blessings of almighty and would like to thank God for giving me strength and patience to carry out my work with full of dedication and ethical values.

Lathiya Jay S.

**Index**

|  |  |  |
| --- | --- | --- |
| **Sr. No** | **Content** | **Page No.** |
| 01 | Abstract & Key Words |  |
| 02 | Introduction |  |
| 03 | Literature review |  |
| 04 | Methodology |  |
| 05 | Data Analysis |  |
| 06 | Implementation |  |
| 07 | Result & Discussion |  |
| 08 | UI Development |  |
| 09 | Limitation of system |  |
| 10 | Conclusion |  |
| 11 | Bibliography |  |
|  |  |  |

**Chapter 1**

**Abstract & Key Words**

**Abstract**

As domestic air travel in India is becoming increasingly popular with different air ticket booking channels coming online these days, passengers are trying to understand how these airline companies make decisions over time about ticket prices. Therefore, many methods are ready to provide the proper time to do so. The customer who buys an air ticket by estimating the price of the airfare is recently proposed. The majority of these strategies make use of sophisticated Computational Intelligence Prediction Models an area of science known as Machine Learning (ML). This paper highlights the parameters and also includes the guidelines that are important for project work to be developed that is indicated above.

**Keywords:** Computational intelligence, Machine learning

**Chapter 2**

**Introduction**

These days, domestic air travel is becoming more and more common in India. Everybody knows that holidays always call for a much-needed vacation and finalizing the travel itinerary becomes a tedious task. With the worldwide growth of internet and E-commerce, commercial aviation industry has witnessed a tremendous growth and has become a regulated marketplace. Hence, for Airline revenue management, different strategies like customer profiling, financial marketing, social factors are used for setting ticket fairs. It is often seen that airfares are low when tickets are booked months in advanced and then they rise when booked in urgency. But, number of days/hours until departure isn’t the only factor which decides flight fare, there are numerous other factors as well.

Because of this complex pricing model of aviation industry, customers find it very difficult to find a perfect and cheapest ticket deal. To solve this problem, Machine Learning and Deep Learning based several technologies and modals are developed and extensive research is also underway. This paper throws light on Machine Learning based Flight fare Prediction System which uses Random Forest Regression to estimate prices of airline tickets. Various features that influence prices are also studied along with system’s experimental analysis. In Section III, literature survey was carried out wherein, technical papers and some existing models and systems were studied. Differences in the features considered are also mapped down, In Section IV, Methodology is discussed. In Section V, Data Analysis is Discussed with maps and graphs. In Section VI, Implementation of different machine learning algorithms. In Section VII, results are presented along with various comparisons between findings. In Section VIII, UI Development is discussed. In section IX, Limitation of system discussed. In Section X, conclusions are stated and possible advances for future research are mentioned.

**Chapter 3**

**Literature Review**

K. Tziridis, Th. Kalampokas, et.al in have developed an airfare price prediction system. The paper begins with a piece of general information about Machine learning and then the authors further proceed to the methodology comprising of four distinct phases of Feature Selection that influence airfare prices, collection of data from Greek Aegean Airlines, Selection of accurate ML Regression model, and its evaluation. The airline dataset had the following eight features- departure and arrival time, number of free luggage, days before departure, number of intermediate stops, holiday, time of day, any day of the week. The authors performed prediction using eight state-of-art regression Machine Learning models including, MLP, GRNN, ELM, Random Forest Regression Tree, Regression Tree, Bagging Tree, Bagging Regression Tree, Regression SVM, and Linear Regression. Performances of these ML models were also compared and evaluated. The Bagging Regression Tree model outperforms other models with its accuracy of 87.42%.

Tianyi Wang, Samira Pouyanfar, et. al in states the problem of market segment level airfare price prediction and propose a novel application for the same using a Machine learning approach. For training and evaluation of the proposed model, two public datasets, DBIB and T-100 were collected with minimal features. The methodology includes data cleaning, data transformation, data pre-processing, selection of extracted features, and applying ML model. The extracted features include distance, seat class, passenger volume, load factor, competition factor, LCC presence, Crude oil price, CPI, and Quarter. Random Forest Model is used for development because of its best performance on the data in comparison to other models including LR< SVM and Neural Networks. This prediction framework achieves high accuracy with an R squared score of 0.869.

Tao Liu, Jian Cao, et. al in address the problem of airfare forecasting and introduce an ACER framework for airfare price prediction which predicts the lowest ticket price available before departure day. The model is deployed using three steps, namely Feature Selection and Extraction, Selecting a Forecast algorithm, and Multistep Forecasting. The dataset is collected from leading OTAs in China. For feature extraction, a matrix-like a schema is used with matrix rows comprising consecutive departure dates and columns with the number of days before departure. Model’s input features include prices of the same itinerary, prices of itineraries departing in the last few days, statistical values, route features, and airfare searching times. Bayesian Regression is used as the base model and result analysis is based on the metrics of RMSE. Results from the experimental analysis showed that ACER performed better with an error of just between 3.7% and 6%.

Supriya Rajankar, Neha Sakharkar, et, al. in put forward Machine Learning Regression methods to predict the price of a flight ticket at a given time. The paper describes its methodology which starts with the data collection process and the dataset is procured from makemytrip.com. This dataset has seven components namely, Date of journey, time of departure, place of departure, time of arrival, place of destination/arrival, airway company, and total fare. Next, the data is cleaned, pre-processed and analysis is performed using different AI models. Authors perform a comparative study of results based on the performance of various Machine Learning models like LR, Decision tree, SVM, KNN, Random Forest, and Bagging Regression Tree. It was observed that KNN gives R-squared value nearing 1 indicating high accuracy.

Juhar Ahmed Abdella, Nazar Zaki, et, al. in present a review of deep learning and social media data-based Airline ticket price prediction model. The authors introduce the current airline ticket pricing situation with the factors that affect ticket prices. They also touch upon the strategies which airlines induce to increase their revenue and maximize profits. This model helps its users by advising them whether to buy tickets or wait for a suitable time to get the optimal deal. It uses data mining techniques like Rule Learning, Reinforcement Learning, time-series methods, and their combinations to achieve greater accuracy in predicting the fare of flights. Features considered for the study include flight number, hours till departure, the current price of a ticket, airline, and its route. The model attained maximum accuracy of 61.9% when a combination of the above-mentioned techniques was used.

**Chapter 4**

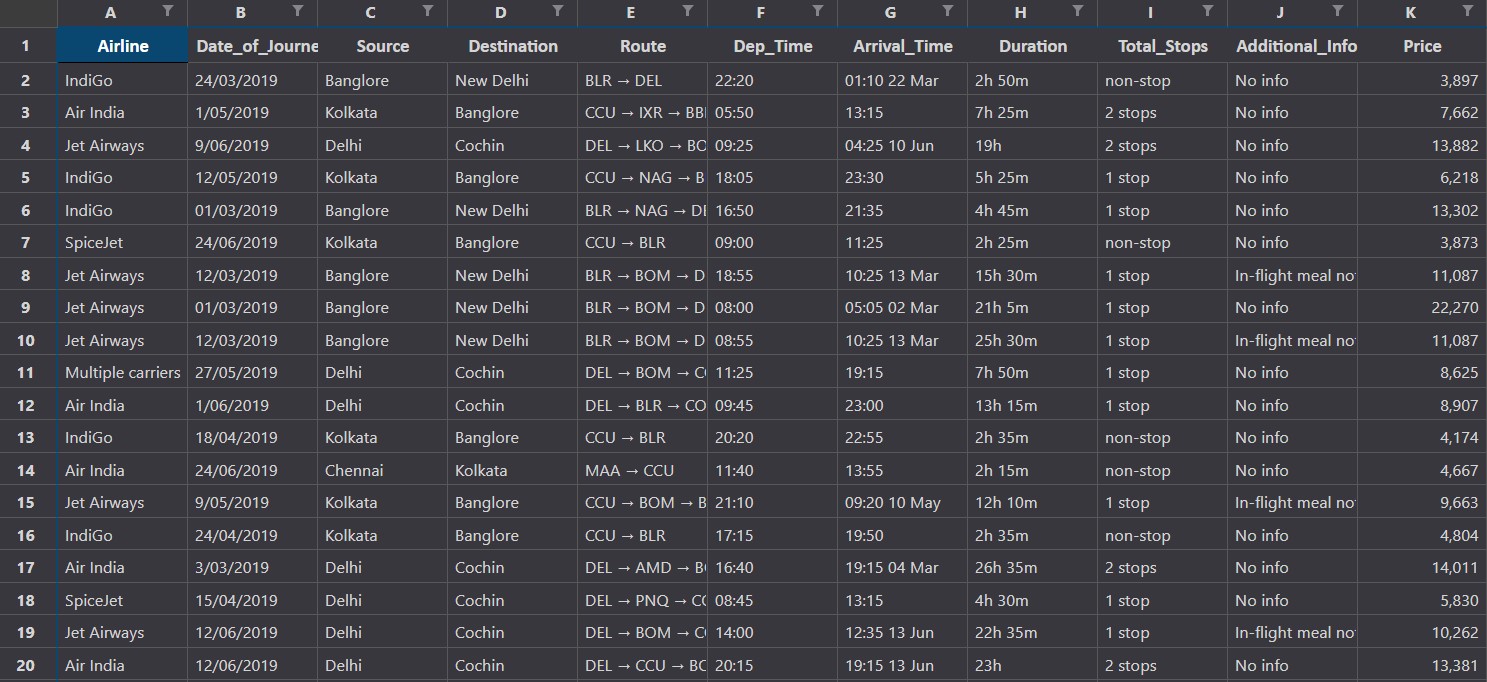
**Methodology**

**A*.*Data Collection**

Both the training and testing datasets have been extracted from Kaggle data repository. They contain categorical as well as nominal data related to the Indian Airlines from the year 2019. The dataset provides vital information about some impacting features to predict the fare of a flight - such as the places of departures and arrivals, time of departure and arrivals, the route of the flight, the number of halts during the journey and the price of the ticket depending on those features. It’s an enormous dataset of 10683 rows and 11 columns (each representing one attribute).

The document contains the data with features and its details. Choosing the features needed for the estimation of the predicted flight price is an important prospect. The site's output contains the number of parameters for each flight: but not all are needed, so only the accompanying components are required.

* Airway company
* Date of Journey
* Data of Arrival
* Data of Departure
* Time of Arrival
* Route
* Source
* Destination
* Total number of Stops
* Price



**Table 1**: Row Data

The original dataset obtained from kaggle.com is shown in Table I. It is essentially raw information containing all the characteristics. For many routes, this information has been obtained.

**B. Data Pre-processing**

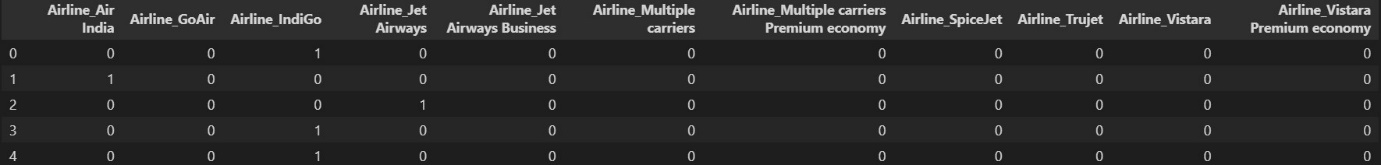
While pre-processing the data, we converted the date of journey, departure time and the arrival time from string datatype to date-time object and extracted the numeric values from them; the month-date numeric value from the date of journey attribute and hour-minute numeric value from the departure time and arrival time attributes respectively. Later, we have implemented the ‘One hot encoding’ method for the nominal categorical data and the label encoding method for ordinal categorical data present in both the training as well as the testing dataset. ‘One hot encoding’ is a process of converting the categorical data variables into numerical values thus making it suitable to use while implementing machine learning algorithms. One hot encoding method was applied to nominal categorical data attributes such as the ‘source’, the ‘destination’ and the ‘airline company’ chosen by the user.

‘Label encoding’ helps us convert the labels into numeric values in order to make the dataset suitable for use. Label encoding method was applied to the nominal categorical data attributes such as the ‘total number of halts in the journey’. The columns were re - arranged at the last step.

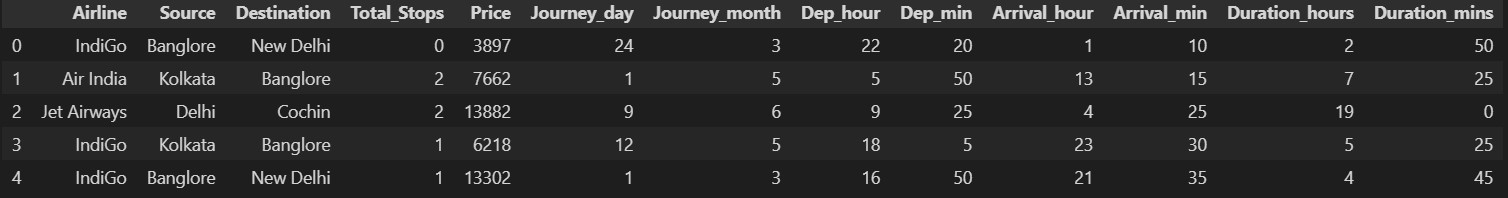
One can find many ways to handle categorical data. Some of them categorical data are,

**Nominal Data** -> data are not in any order -> **OneHotEncoder** is used in this case

**Ordinal Data** -> data are in order -> **Label Encoder** is used in this case



**Table 2**: OneHotEncoder



**Table 3**: Label Encoder

**C. Data Cleaning**

The null values present in the training dataset where removed. A few columns which were of no use for the feature selection process were deleted from the dataset. The columns of attributes having the categorical data were dropped from the dataset after the new columns containing the numerical values extracted from the pre-processed data were stored for the prediction. Thus, the training dataset suitable for use was obtained and it had the following attribute columns.

**Description of the Attributes**

|  |  |
| --- | --- |
| Data Attribute | Description |
| Total Stops | The number of halts in the journey |
| Journey Day | The numerical value of ‘day’ selected from the calendar |
| Journey Month | The numerical value of ‘month’ selected from calendar |
| Dep\_hour | The numerical value of ‘hour’ in departure time |
| Dep\_Min | The numerical value of ‘minutes’ in departure time |
| Arrival\_hour | The numerical value of ‘hour’ in arrival time |
| Arrival\_Min | The numerical value of ‘minutes’ in arrival time |
| Duration\_Hour | The numerical value of ‘hours’ in duration time |
| Duration\_Min | The numerical value of the minutes in duration time |
| Airline Company (One hot encoding applied) | Display ‘1’ for the chosen Airline company and display ‘0’ for the rest |
| Source (One hot encoding applied) | Display ‘1’ for the chosen Source and display ‘0’ for the rest |
| Destination (one hot encoding applied) | Display ‘1’ for the chosen Destination and display ‘0’ for the rest |

**Table 4:** Description of the Attributes

**D.** **Generating the Model**

The model has been generated using the Random Forest Regression.

**E.** **Presenting the Final Prediction**

The user input fields will be provided on a webpage developed using the flask framework. The webpage body was built using HTML5 and the same was styled using CSS3. After the user fills all the required input fields and submits the form, the data will be sent to the generate random forest regression model and the predicted value of the ticket price will be displayed.

**Data Collection**

**Data Pre-processing**

**Data Cleaning**

Rearranging the columns

Check the dataset for null values

Performing One Hot Encoding on Nominal categorical data

Filling NaN values with mean, median, and mode

Performing Label encoding on Ordinal Categorical data

Dropping the columns having non-impacting information

Convert string data type to datetime object

**Model Generation**

**(Random Forest Regression)**

**User Input**

**(Departure Data and Time, Arrival Time, Source, Destination, Number of Stops, Airline)**

**Prediction**

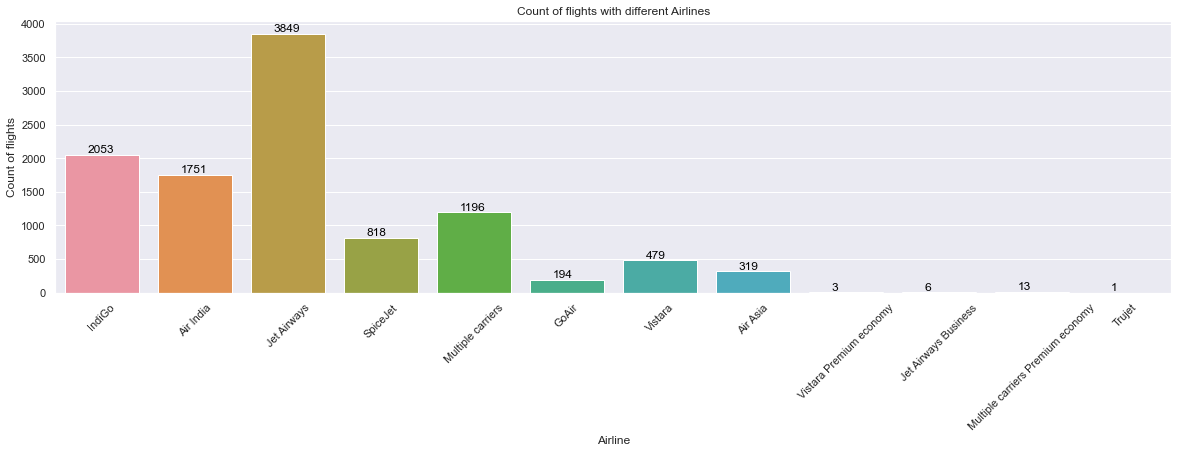
**Flight Ticket Price**

**Figure 4**: Methodology block diagram

**Chapter 5**

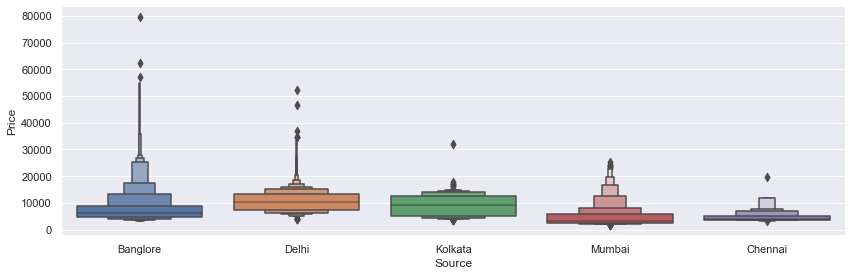
**Data Analysis**

Data preparation is monitored by breaking down the data, exposing the hidden patterns and applying various regression models afterwards. Similarly, from the existing features, a few features can be calculated. Flight days can be given by measuring the difference between the date of the flight and the date of collection of the details. In addition, the flight date, whether on a festive day or a weekday or a weekend, is significant. The flights scheduled during the weekends instinctively cost more than the flights on weekdays. In addition, time plays a major role.

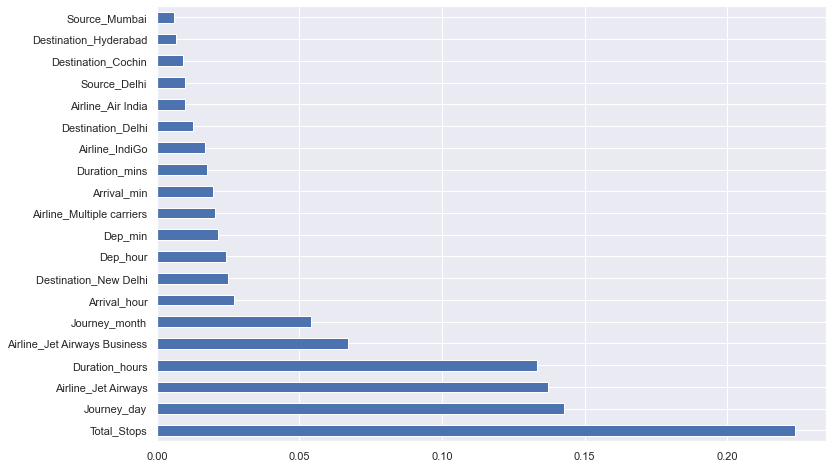


**Figure 1**: Count of flights with different Airlines

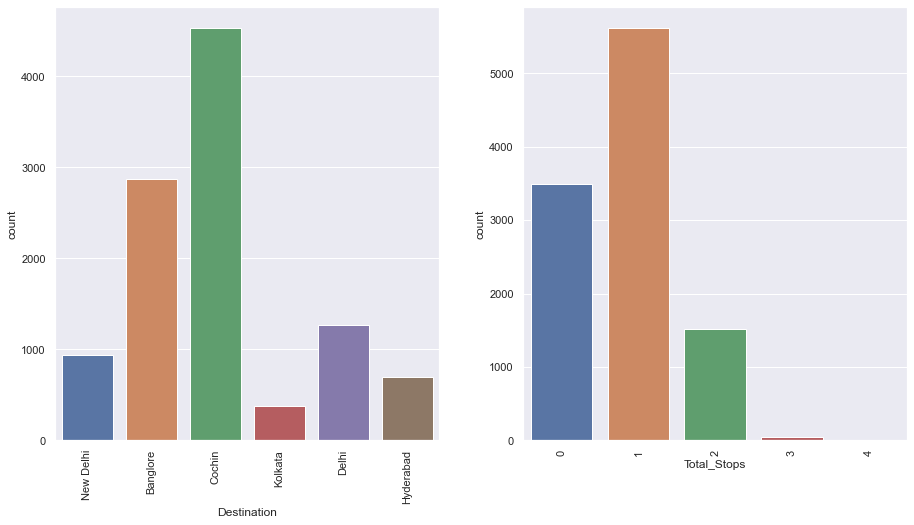
From graph we can see that Jet Airways Business have the highest price. Apart from the first Airline almost all are having similar median.



**Figure 2:** Source Vs Price



**Figure 3:** Important parameters



**Figure 4:** Count plot

**Chapter 6**

**Machine Learning Modals Performance**

In machine learning, several algorithms are applied to forecast the prices of flight tickets. The algorithms are: Linear regression, Decision tree, K-Nearest neighbors, and Random Forest Algorithm. These models have been implemented using the python library Sklearn. The parameters like MAE and MSE, RMSE are considered to check the efficiency of these models.

**Linear Regression**

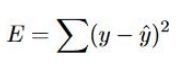
Simple linear regression analysis is used to determine the association between two continuous variables. The indicator variable of what importance is to be found is one of the two variables. It is not the deterministic relationship between two variables that gives the statistical relationship. The linear regression algorithm gives the given data the best fit line for which the prediction error is limited. The two main factors for understanding linear regression are gradient descent and cost function. The equation for linear regression is:

**Y(pred) = b0 + b1 \* x**

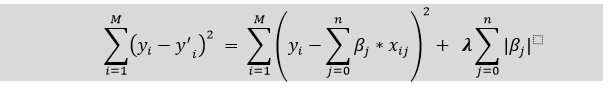
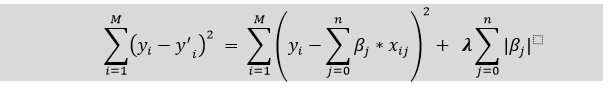
The value of coefficients b1 and b0 is selected such that the value of the error is as minimal as possible. The error is given by the square of the expected and actual value difference. The mean square error is taken to deal with the negative values (MSE). The positive or negative relationship between x and y is given by b0 here, while b1 is called bias. Regression problem accuracy is calculated in terms of R-squared, MAE, and MSE.

**Decision Tree**

This tree count isolates the data obtained into small subsets, rendering it permanent at a comparable time. The new findings show the tree with the decision centers, and the leaf centers as well. At any rate, this decision-center point will contain two branches. Think of the entire knowledge index as a root at first. Function aspects are kicked out of the opportunity. If the characteristics are relentless, then before structuring the model, they have to be discretized. In view of estimation property records are corrected recursively. In the decision of tree computation, Knowledge Gain and Gini index are two basic properties. Information Gain is characterized as the change in entropy in quantity. Higher entropy suggests the substance's greater efficacy. Therefore, entropy is a proportion of an arbitrary variable's susceptibility. The Gini Index tests how to falsely identify an arbitrarily chosen component on a regular basis. This suggests that a feature with a lower Gini index should be liked. For Regression tree, cost capacity can be a basic squared condition:

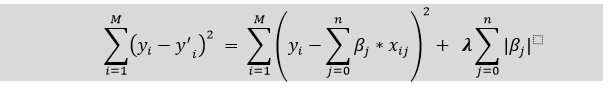


Where y is the actual value from the dataset and y cap is predicted value. Have a class with the maximum sum of the expected value obtained by a split function called the gain of knowledge. If the class is kept dividing and dividing at the leaf node without any condition, the algorithm will be truly massive, slow and over-fitted. To stop this, a minimum count on the training example on the leaf node is assigned.



**Lasso**

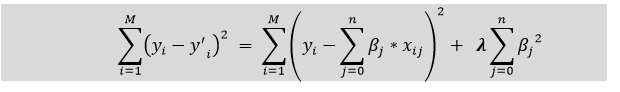
Lasso regression is a regularization technique. It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.



**Ridge**

Ridge regression is a regularization technique, which is used to reduce the complexity of the model. It is also called as L2 regularization.

In this technique, the cost function is altered by adding the penalty term to it. The amount of bias added to the model is called Ridge Regression penalty.



**GradientBoostingRegressor**

Gradient Boosting is a popular boosting algorithm. In gradient boosting, each predictor corrects its predecessor’s error. In contrast to Adaboost, the weights of the training instances are not tweaked, instead, each predictor is trained using the residual errors of predecessor as labels. There is a technique called the Gradient Boosted Trees whose base learner is CART (Classification and Regression Trees).

**Y(pred) = y1 + (eta \* r1) + (eta \* r2) + …… + (eta \* rN)**

The class of the gradient boosting regression in scikit-learn is GradientBoostingRegressor. A similar algorithm is used for classification known as GradientBoostingClassifier.

**Random Forest Regression**

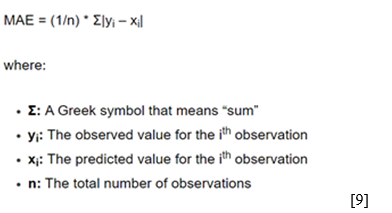
**Random Forest Regression** is a supervised learning algorithm that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model. It operates by building decision trees during training time and outputting the mean of the classes as the prediction of all the trees.

**Chapter 7**

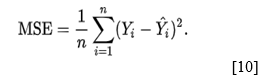
**Result & Discussion**

We are using evaluation metrics such as MAE (Mean Absolute Error), MSE (Mean Squared Error), RMSE (Root Mean Squared Error) and R squared Value for evaluating all the 3 models.

1. **Mean Absolute Error (MAE)** is the average of difference between the actual data value and the predicted data value. It is calculated as shown below:



1. **Mean Squared Error** is the average squared difference between the estimated values and the actual value.



Where, n = Data set observations

Yi = Observation values

Y^i = Predicted Values

**3.** **Root Mean Squared Error is the root of MSE**

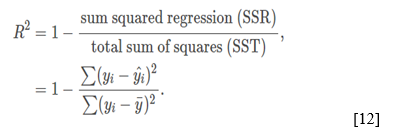


Where, n = Data set observations

Si = Predicted values

Oi = Observations

**4.** **R squared** value is used for measuring the accuracy of the model.



Where,

R^2 = coefficient of determination

Following are the values of evaluation metrics for the Random Forest Regression Model.

**Values of Evaluation Metrics**

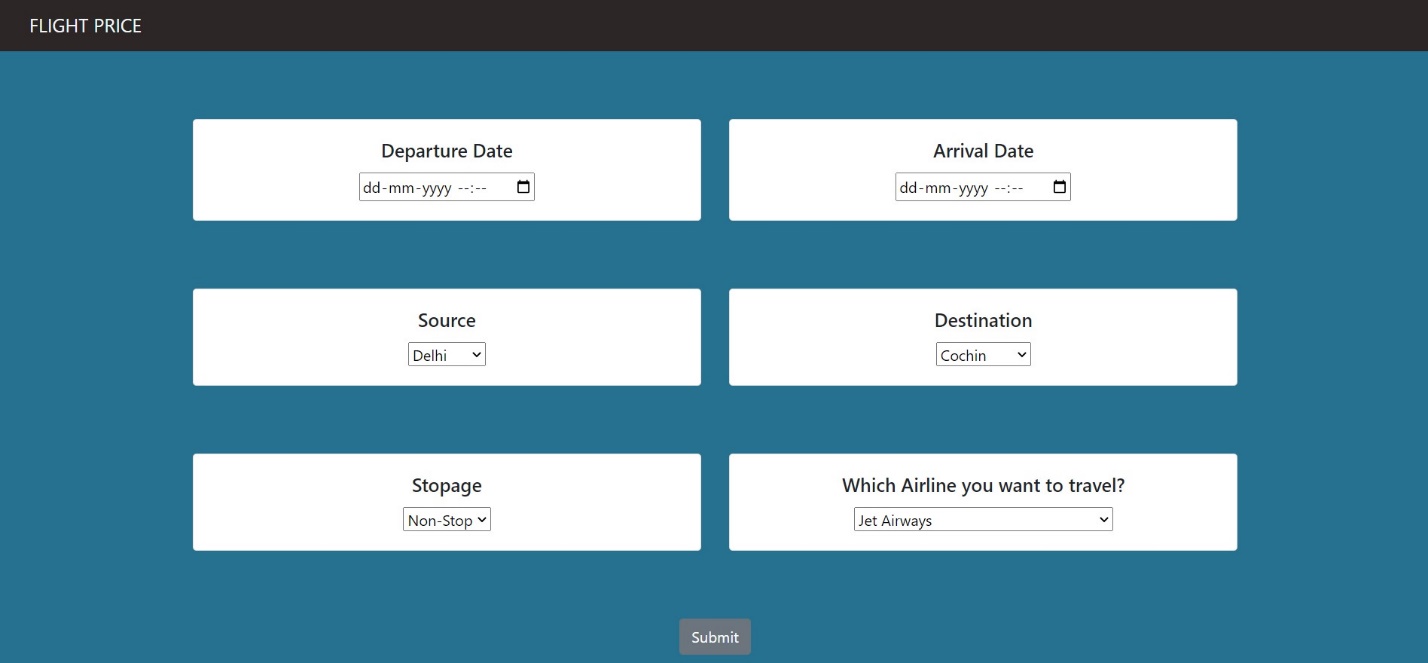
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Algorithms** | **Accuracy**  **(%)** | **MAE** | **MSE** | **RMSE** |
| SVR | 0.040 | 3535.5200 | 20300555.7626 | 4505.6138 |
| AdaBoostRegressor | 42.90 | 2829.4132 | 11647247.3243 | 3412.8063 |
| GradientBoostingRegressor | 79.08 | 1489.6617 | 4265985.9109 | 2065.4232 |
| LinearRegression | 62.63 | 1937.8989 | 7621946.5167 | 2760.7873 |
| Lasso | 62.62 | 1937.2911 | 7625316.7654 | 2761.3976 |
| Ridge | 62.57 | 1940.4783 | 7635170.4637 | 2763.1812 |
| RandomForestRegressor | 79.81 | 1176.2320 | 4352308.4389 | 2086.2186 |

**Table 5:** Regressions with its accuracy and errors

**Chapter 8**

**UI Development**

In this project, Flask framework has been used for the UI development. The main web page of the project takes the required inputs from the user in order to predict the price for the flight. The user inputs required are Departure date and Departure Time, Arrival time of the flight, Source and Destination of the journey, the number of halts during the whole journey and most importantly the airline company which we choose to travel with. After inputting all the fields, the user will click the Submit” button and then the form is submitted. Model enters the scenario at the backend after the submission of the form. The inputs take the help of the historical data and are analysed through supervised machine learning techniques resulting in the prediction of the ticket price. The routing of the pages is done based on the URLs. When the browser finds the ‘/’ in the URL it redirects the user to the home page. After the submission of the form, the user is redirected to the '/result' URL i.e., to the result page where we can see the final result i.e., the prediction of the ticket price. The webpage body was built using HTML5 and the same was styled using CSS3.



**Figure 5: UI page**

**Chapter 9**

**Limitation of system**

Flight ticket prices can be something hard to guess, today we might see a price, check out the price of the same flight tomorrow, it will be a different story. We might have often heard travellers saying that flight ticket prices are so unpredictable. As data scientists, we are going to prove that given the right data anything can be predicted. So, the collected train data should be accurate if not it may result in wrong prediction. And also, it is necessary to update the train data time to time for best results.

**Chapter 10**

**Conclusion**

For this paper, an extensive study was carried out with dataset collection from Kaggle and Random Forest Machine Learning model was used for deployment. Using visualization, we were able to determine the features which influence airfare prices the most. With experimental analysis, it can be concluded that Random Forest Regression model achieves good accuracy. The future aim is to work more on the feature selection and model accuracy. We also plan to extend the study by working with larger datasets and greater number of experimentations on the same to procure more accurate airfares which will in turn help users to get an estimated cost of their next airplane travel and can benefit them to make the best deal. We also plan to level up web applications’ user interface to provide a premium user experience. We can also consider various other crucial features that affect airplane ticket prices like public holidays, number of luggage, number of hours till departure, crude oil price, etc. in order to get best results. In the near future, there is also a plan to host the web application.

**Chapter 11**

**Bibliography**

Tom Chitty, CMBC Business News, “This is how airplanes price tickets”, August 3, 2018. Available: <https://www.cnbc.com/2018/08/03/how-do-airlines-price-seat-tickets.html>

Moira McCormick, BlackCurce, “Behind the Scenes of Airline Pricing Strategies”, September 19, 2017. Available: <https://blog.blackcurve.com/behind-the-scenes-of-airline-pricing-strategies>.

 K. Tziridis, Th. Kalampokas, G. A. Papakostas, “Airfare Prices Prediction Using Machine Learning Techniques”, 25th European Signal Processing Conference (EUSIPCO), IEEE, October 26, 2017.

Tianyi Wang, Samira Pouyanfar, Haiman Tian, Yudong Tao, Miguel Alonso Jr., Steven Luis and Shu-Ching Chen, “A Framework for Airfare Price Prediction: A Machine Learning Approach”, 2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI), September 9, 2019.

Tao Liu, Jian Cao, Yudong Tan, Quanwu Xiao, “ACER: An Adaptive Context- Aware Ensemble Regression Model for Airfare Price Prediction”, 2017 International Conference on Progress in Informatics and Computing (PIC), December, 2017.

Supriya Rajankar, Neha Sakharkar, Omprakash Rajankar, “Predicting the price of a flight ticket with the use of Machine Learning algorithms”, international journal of scientific & technology research volume 8, December, 2019.

Juhar Ahmed Abdella, Nazar Zaki and Khaled Shuaib, “Automatic Detection of Airline Ticket Price and Demand: A Review”, 13th International Conference on Innovations in Information technology (IIT), January 10, 2019.

Chaya Bakshi, Medium, “Random Forest Regression”, June 9, 2020, Available: <https://levelup.gitconnected.com/random-forest-regression-209c0f354c84>

Zach, Statology, “How to calculate mean Absolute Error in Python”, January 8. 2021, Available: <https://www.statology.org/mean-absolute-error-python/>

NCL.AC.UK, “Coefficient of Determination, R-squared”, Available: <https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/coefficient-of-determination-r-squared.html>

[Google Scholar](https://scholar.google.com/)

[Kaggle, a subsidiary of Google LLC, is an online community of data scientists and machine learning practitioner.Kaggle offers a no-setup, customizable, Jupyter Notebooks environment. Access free GPUs and a huge repository of community published data & code.](https://www.kaggle.com/)

[W. Groves and M. Gini, “An agent for optimizing airline ticket purchasing,” 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2013), St. Paul, MN, May 06 - 10, 2013, pp. 1341-1342.](https://www.ifaamas.org/Proceedings/aamas2013/docs/p1341.pdf)

[T. Janssen, “A linear quantile mixed regression model for prediction of airline ticket prices,” Bachelor Thesis, Radboud University, 2014.](https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.566.5436&rep=rep1&type=pdf)

[Wohlfarth, T. Clemencon, S.Roueff, “A Dat mining approach to travel price forecasting”, 10th international conference on machine learning Honolulu 2011.](https://www.sciencedirect.com/science/article/pii/S131915781830884X)

[Dominguez Menchero, J. Santo, Reviera, ”optimal purchase timing in airline markets” ,2014.](https://www.ijert.org/a-survey-on-flight-pricing-prediction-using-machine-learning)