**Flight Prices Prediction**

**A Project Report**

***Submitted by***

**Aishwarya Dubey – 1700117C203**

**Nikita Agarwala – 1700247C203**

**Surya Kiran Kanduri-1700218C203**

**Nishant Pandey-1700250C203**

**Batch - 2017 CSE 1**

***Course***

**MACHINE LEARNING and DATA MINING - CSE2702**

****

**SCHOOL OF ENGINEERING & TECHNOLOGY**

**BML MUNJAL UNIVERSITY**

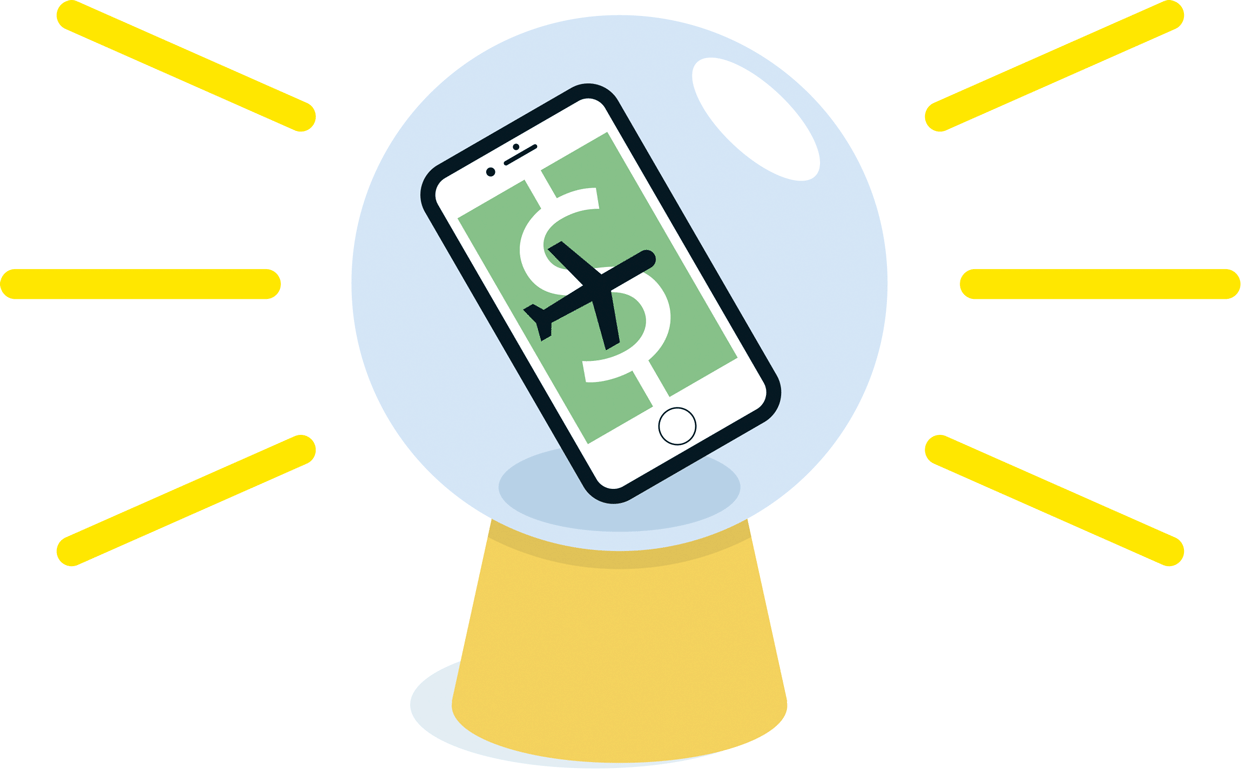
**April 2020**

***Abstract –*** These days, carrier ticket costs can shift powerfully and altogether for a similar flight, in any event, for close by seats inside a similar lodge. Clients are looking to get these seats at minimal cost while aircrafts are attempting to keep their general income as high to increase their profits. Aircrafts use different computational models to expand their income, for example, request forecast and value separation. From the client side, two sorts of models are proposed by various analysts to set aside cash for clients: models that foresee the ideal time to purchase a ticket and models that anticipate the base ticket cost. In our project we try anticipating the base price for the consumers for various routes in regards with the Indian market.We see how the base price depends upon the services being provided as well as the number of stops that are being taken.We model various ML algorithms to do the above mentioned work and then compare their performance.The derived experimental results show that by stacking of ML models are able to handle this regression problem with 92% accuracy after extensive data preprocessing.

***Keywords –*** Machine learning; Pricing models;Data cleaning;Prediction model;Ensemble Learning; Airfare price.

1. **INTRODUCTION**

Our project deals with airfare prices prediction. Airlines use various algorithms to fluctuate and control the airfares in order to increase their revenue. According to Google Trends, "Cheap Air Tickets" is the most searched query in the aviation sector in India. Because of the high unpredictability of the valuing models applied by the aircrafts, it is extremely troublesome for a client to buy an air ticket in the most minimal cost, since the value changes progressively.



*Figure 1 [1]*

We often hear travelers saying that flight ticket costs are so high.We are going to demonstrate that given the correct information anything can be anticipated.Since this is such a critical issue a lot of research is going on in this field but a little is open source and a lot of companies are cashing in using various algorithms. Most of the research that has been done is based on the databases of flight information from other countries.We got one database for Indian Aviation and we used Linear Regression, XGB Regression, Random Forest Regression and Extra Tree Regression to build our model .In the final iteration of the problem we stacked all these models and see how ensemble learning can be useful in applications like these.

**II. LITERATURE REVIEW**

It is very difficult for the customer to purchase a flight ticket at the minimum price. Air ticket price prediction is a challenging task since the factors involved in pricing change dynamically and make the price fluctuate. For this reason, several techniques are proposed to provide the right time for the buyer to purchase an air ticket by predicting the airfare price. The majority of these methods are making use of sophisticated prediction models from the computational intelligence research field known as Machine Learning (ML). In the last decade, researchers have incorporated machine learning algorithms and data mining strategies to better model observed prices. Among them, regression models, such as Linear Regression (LR), Support Vector Machines (SVMs), Random Forests (RF), are frequently used in predicting accurate airfare prices

K. Tziridis, through his research paper “Airfare Prices Prediction, Using Machine Learning Techniques” has shown that it is feasible to use ML models for predicting airfare prices.He has claimed that “Bagging Regression Tree”, “Random Forest Regression Tree”, “Regression Tree” are the most stable model with variation in their accuracy scores.He has build all these models again and again with considering different features like departure time and overnight journey (yes/no).If we consider execution time as a parameter then “Random Forest Regression Tree” and “Regression tree” are the best performing. But he has worked on some Greek airlines.

Also, Vinal Raja and his colleagues have worked on this problem statement in context with the Indian aviation sector and in their model they have tried predicting that whether it is the right time to buy a ticket or not.In their work Logistic Regression,Regression Tree and Bagging Regression Tree have given high accuracy.They also found out that for some routes like Mumbai-Delhi prices remain constant and they attributed this behaviour to high frequency of flights ,heavy competition and high demand.

Gini and Groves took the Partial Least Square Regression(PLSR) for developing a model of predicting the best purchase time for flight tickets. The data was collected from major travel journey booking websites from 22 February 2011 to 23 June 2011. Additional data were also collected and are used to check the comparisons of the performances of the final model.

Janssen built up an expectation model utilizing the Linear Quantile Blended Regression strategy for SanFrancisco to NewYork course with existing every day airfares given by www.infare.com. The model utilized two highlights including the number of days left until the takeoff date and whether the flight date is at the end of the week or weekday. The model predicts airfare well for the days that are a long way from the takeoff date, anyway for a considerable length of time close to the takeoff date, the expectation isn't compelling.

A study by Dominguez-Menchero recommends the ideal buying time dependent on nonparametric isotonic relapse method for a particular course, carriers, and timeframe. The model gives the most extreme number of days before buying a flight ticket.The passage taken and date of procurement are the two variables that are considered for expectation

In our paper, we address the problem of market segment level airfare price prediction by using publicly available datasets and stacking various ML models to predict market segment level airfare price. The goal of our proposed work is to correctly approach this problem with emphasis on ensemble learning as a way of solving regression problems like these.Also highlighting the benefit of using One hot encoder over label encoder so that ordering can be removed.

**III. METHODOLOGY**

**We tackled our problem in phases.The phases are discussed below in detail:**

1. **Data Collection-** Initially we thought we would do some data mining for our problem statement from Expedia using API but due to the ongoing pandemic we could not do so as all flights were cancelled .So we used a database that was available on MachineHack and proceeded with it.The original database was of the shape [10463,11].10 features are intrinsic in nature and ‘Price’ is the label for our project work.
2. **Cleaning Of Data/Data Preprocessing-** On analysing the data we found that there were many empty rows and columns with null value that were of no use.We did following things in this phase.
   1. Dropped rows having null values.
   2. Dropped rows having duplicate values.
   3. Replace repeating values in the data set :For example,in the database New Delhi and Delhi both were used.So we changed all values of Delhi into New Delhi.
   4. Converting ‘Duration of Flight’ and ‘Date’ into workable format.
3. **Feature Selection-** During this phase we have identified various features which are required for accurate prediction of flight prices. For every flight we have collected the following features.

* F1: Feature-1- Airline Name
* F2: Feature-2- Date of Journey
* F3: Feature-3- Source
* F4: Feature-4- Destination
* F5: Feature-5- Route
* F6: Feature-6- Departure Time
* F7: Feature-7- Arrival Time
* F8: Feature-8- Total Stops
* F9: Feature-9- Holiday (yes/no)
* F10: Feature-10- Price

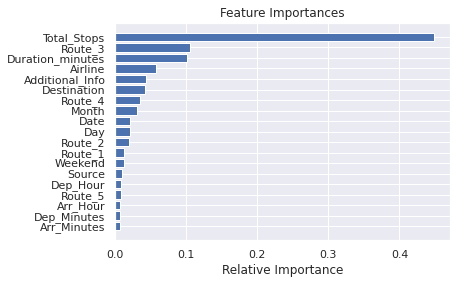
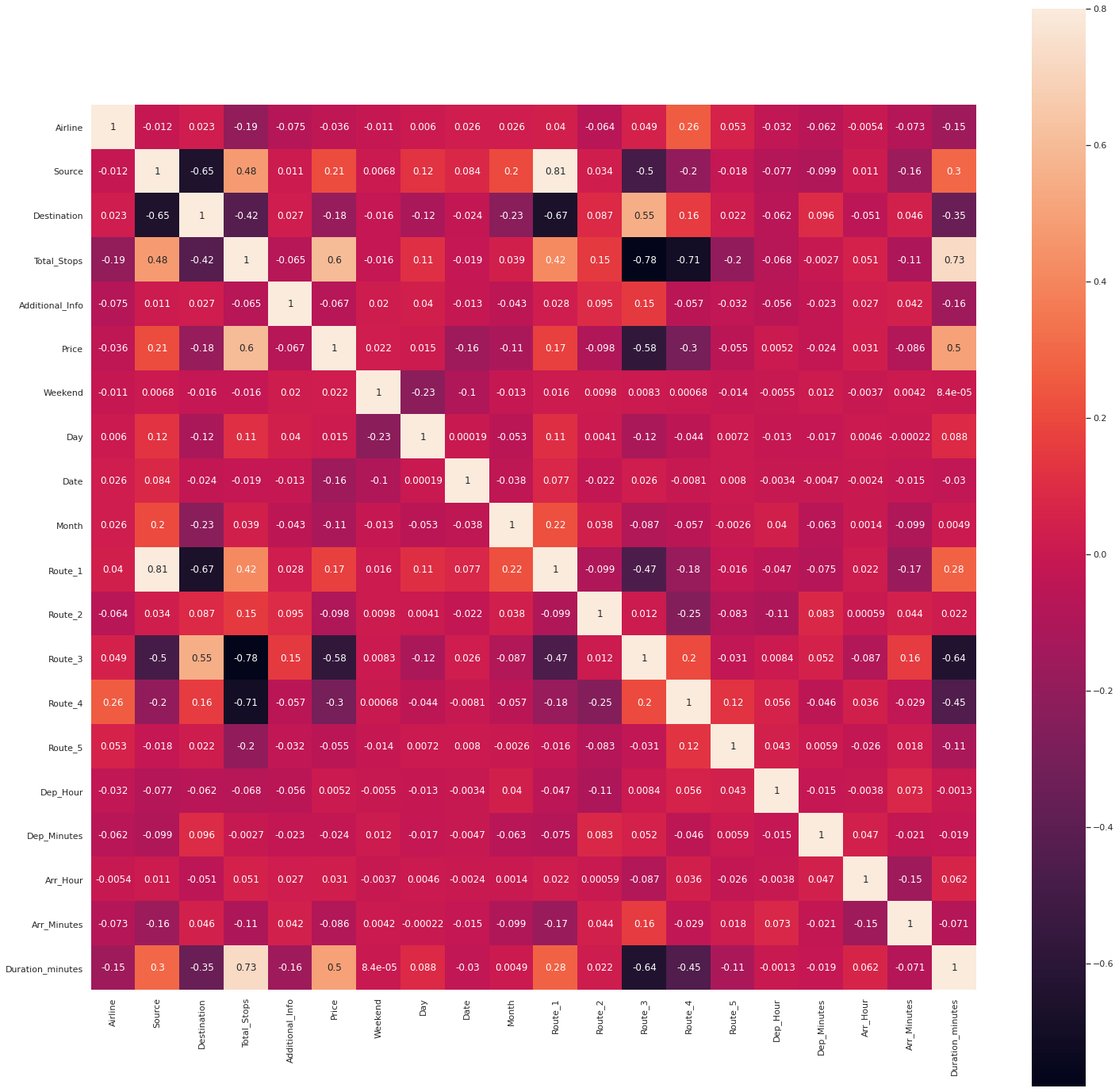
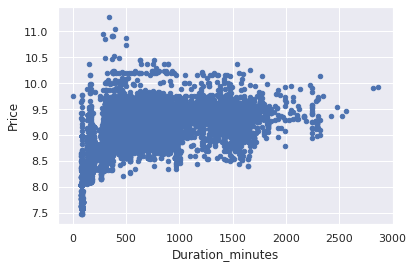
*Fig 2:Feature Importance according to XGBRegressor*

Fig 3.Correlation Matrix

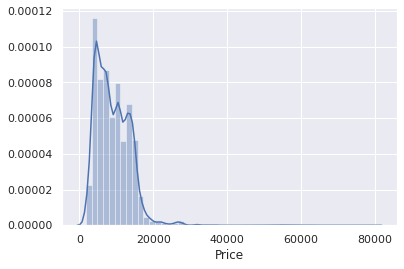
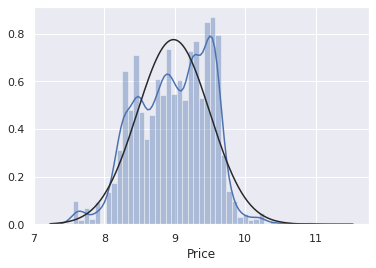
We see that Total Number of Stops,Duration of Flight ,Additional Info are some of the important features of the dataset

**4. Feature Engineering-** In this step we can add a new feature or replace the previous feature that is not required with a new feature that can be additionally used for analysing the data.

* For eg., adding the weekend price of the flight column which can provide travellers easy access to the cost of the flight on the weekend.
* **Attribute construction**:We constructed some new features out of the existing features to get more information.For example we further broke Route Attribute into 5 different features as it was important.A value like BLR->DEL was broken individually into BLR and DEL as intuitively path taken has some effect on airfare.
* One might think ‘Duration’ would be linearly related with ‘Price’ but that is not the case as seen in Figure 4.

*Fig 4.Plotting Duration\_minutes wrt Price*

**5. Preparing for model-**  During the initial iterations we realised that our label ‘Price’ was skewed.So we did log transformation on ‘Price’ to normalize our label.

*Fig 5:Normalization of ‘Price’*

**6. Encoding the data :**Since the regression models work better with numerical inputs we encoded features like ’Source’,’Destination’,’Additional\_Info’,’Day’ and different routes using Label Encoder and ‘Airlines’ using One-Hot Encoder.Label encoder converts categorical data into numeric values but this can misinterpreted by algorithms as ordinal data.So we use ‘One-Hot Encoder’ of scikit.It does binary encoding and adds a new column in the data frame for each unique category value.So for “Airlines’ we got 12 new columns and that is the disadvantage.It swells the dataset thus the execution time also increases.

**7. Selection Of The Model-** In our project we have used following 4 models:

A. XGB Regression

B. Random Forest Regression Tree

C. Extra Tree Regression

D. Linear Regression

We also tried little ensemble learning.It is ML technique in which several models are combined to make one final model.We did stacking of the above mentioned models.We used Extra Tree Regression ,Random Forest Regression Tree and XGB Regression as base models and Linear Regression as the meta model in stacking.

A picture containing drawing

Description automatically generated

**IV. PERFORMANCE**

After all the data preprocessing we modeled our data using various regression models.To gauge the performance of various models we took the help of performance metrics from scikit library of Python.We mainly used 5 components to measure the performance of every model.

1. R Squared
2. Mean Absolute Error
3. Mean Squared Error
4. Root Mean Squared Error
5. Root Mean Squared Error Log

R squared explains how much variance in our label (in our case,Airfare) can be explained by the various features of our model.So we will use mainly R-squared to say how good a model is explaining the data.

Also for all the models the training set and testing set has been divided in the ratio of 70:30.

**V. RESULT**

For all the models we initially trained our model using training data and based on testing data we measured the performance of each model with our main performance indicator being R squared.

In Table 1 we see R-squared for various models and see that by Extremely Randomized Trees and by stacking the models we get the highest R-squared,but this was our first iteration.We realised that our label was screwed and so log transformed Price to get higher R2.After normalization also Extremely Randomized Trees performed well but we got a higher R2  when we stacked the models.

Table I:RESULT FOR SKEWED AND NORMALIZED DATA

|  |  |  |
| --- | --- | --- |
| *ML Model* | *R-squared for skewed data* | *After normalization* |
| Linear Regression | 0.49 | 0.60 |
| Random Forest Regression Tree | 0.88 | 0.93 |
| Extremely Randomized Trees | **0.90** | **0.93\*** |
| XGB Regressor | 0.87 | 0.92 |
| Ensemble Learning-Stacking | **0.90** | **0.94** |

\*Read R2 as 93% of the variation in Price can be explained by the other features

When we doing literature review of this project we realised for categorical data in the dataset being used various people have used Label Encoder.It is better to use One Hot Encoder on features like ‘Airlines’,’Source’,’Destination’ and ‘Additional Info’ as these features are categorical in nature but not ordinal .So intuitively we should not use label encoder for encoding such features. and rather use One Hot Encoder. On doing this our R squared dropped to 0.91 when we stacked various models but this should work.

Table II:RESULT USING ONE HOT ENCODER AND AFTER NORMALIZATION

|  |  |
| --- | --- |
| *ML Model* | *One-Hot Encoder+After normalization* |
| Linear Regression | 0.60 |
| Random Forest Regression Tree | **0.90** |
| Extremely Randomized Trees | 0.90 |
| XGB Regressor | 0.89 |
| Ensemble Learning-Stacking | **0.92** |

One can claim after the analysis of the result that Extremely Randomized Trees and XGBRegressor work well but we should give special attention to ensemble learning as it helps us to overcome any unfair bias that is introduced because of one single model.

**VI. CONCLUSION AND FUTURE SCOPE**

We can conclude that for datasets like these we need to do a lot of data preprocessing and feature engineering but it does pay in the end because it improves our ML models.In predicting flight prices regression works and Random Forest Regression Tree and XGBRegressor perform well.Also ensemble learning is a great way of training a model as it makes a stronger ML learner which has reduced bias .

As of now we could only binary encode one feature that is ‘Airline’ and all the other features are label encoded.In the future,we can binary encode more features with One Hot Encoder but we will also need to do Principal Component Analysis with it to reduce the dimensions of our dataset as binary encoding itself makes our dataset sparse.

**REFERENCES**

# References

|  |  |
| --- | --- |
| [1] | J. V. S. R. Vinal Raja, "Prediction of Airfare Using Machine Learning," Shirpur, India , 2018. |
| [2] | N. S. Supriya Rajankar, "A Survey on Flight Pricing Prediction using Machine Learning," Pune, India, 2019. |
| [3] | T. K. G. P. K. Tziridis, "Airfare Prices Prediction Using Machine Learning Techniques," Kavala, Greece, 2017. |
| [4] | T. Janssen, "“A linear quantile mixed regression model for prediction of airline ticket prices," Radboud University, 2014. |
| [5] | N. Z. a. K. S. J. A. Abdella, ""Automatic Detection of Airline Ticket Price and Demand," United Arab Emirate, 2019. |
| [6] | W. G. a. M. Gini, "A regression model for predicting optimal purchase timing for airline tickets,," University of Minnesota, Minneapolis, 2011. |
| [7] | S. P. a. R. R. P. Malighetti, "Pricing strategies of low-cost airlines: The Ryanair case study," Journal of Air Transport Management, 2009. |