A NOVEL DATA SCIENCE AND ML APPROACH TO PREDICT AIRFARE

[1]Tadvi Shabana, [2]Khan Mariya, [3]Shaikh Afifa, [4]Sayyed Naziya Begum

[1]Professor, [2]Student, [3]Student, [4]Student
[1,2,3,4] Department of Computer Science ,
[1,2,3,4]M.H.Saboo Siddik College of Engineering, Mumbai, India

Abstract : Ticket purchasing in advance is a well- known traditional approach but it entirely depends on the Airline industry to change the fare according to factors whether the travel is during the holidays, the number of free seats in the plane etc. Some of the features are seen, but some of them remained hidden. We are using Indian Domestic Airline Dataset which contains multiple columns so over a period as the data increases (approx. 1 year) we will be able to extract few more hidden features to increase the efficiency and accuracy of the system. The goal is to use machine learning techniques to model the behavior of flight ticket prices over the time. In other words system will be able to provide a general idea to the clients when to increase or decrease the fares i.e. prediction of Airfare. For that after collecting the dataset the proposed system will extract important features from dataset, cleaning of data and using Regression Machine Learning Algorithms multiple models will be trained and the accuracy of those models will be compared and prediction report will be given to client.

***IndexTerms* – Airfare, Regression, Data Analytics, Airline, Machine Learning , Prediction.**

I. INTRODUCTION

Business Services provides us intangible products such as accounting, banking, consulting and finance etc. Airline industry is one of the sectors which provide us with the above given services. It also provides air transport services for travelling passenger. These airline services can be categorized as being intercontinental, domestic, regional or international. The goal of improving business sales (revenue) in airline industry depends on the flight ticket selling.

Now-a-days the airline industries are using complex algorithms and strategies for the airfare prices in a dynamic way to regulate seats demand and maximize their revenue. Though these strategies developed that works properly but aren't that accurate. The airfare price depends on many factors like base fare (airfare), booking date, departure date, meal and route. There are so many airline agencies through which we can also book the tickets and sometimes we find their prices lower than direct booking.

If the journey is not too long then passenger can probably skip meal, Wi-Fi or other facilities charges in order to decrease the fare price [1]. But the **Base Fare** will be common for all the passengers containing **Air Charges**. So airfare prices can be dynamically changed by considering many factors. For generating more revenue it is necessary for an airline industry to create some good strategy to predict airfare prices so that is increases industry's profit as well as customer's welfare.

The contribution of the proposed system includes the following activities :

1) **Airfare prices prediction in India for domestic airline**

The dataset contains 45 different columns from which we are extracting those columns (features) which will be used to train the model and predict the given goal.

2) **Investigation and analysis of the features that affects the airfare.**

The proposed system will use Data analytics for completely exploring the data, identifying relationship among those columns, finding some patterns of work, cleaning and refining the dataset to reduce complexity of data. Then the system will extract useful features from the bulky dataset and also will derive new features for making system processing easier.

3) **Performance analysis of the ML models.**

Random forest regression, Support vector machine, K-nearest neighbors Regression etc. Machine Learning algorithms will be used to train the model. Since the airfare is continuous value therefore Regression techniques will be used. At the end the system will generate report for "AIRFARE PREDICTION".

The rest of the paper is organized as follow: **Section II** Study and reporting of the previous work done for predicting air prices. **Section III** Experimental approach of the proposed system and the models used. **Section IV** Discusses the result by comparing the accuracy of different prediction models. **Section V** Conclusion and future work.

II. RELATED WORK

In the existing Airfare prediction systems, for predicting dynamically increasing or decreasing prices of the air different behavioral characteristics of the timing factors are considered and are given much priority. Although, different features are also taken into consideration but the increasing price of crude oil is not taken into consideration [4] which indeed have a high tendency of affecting the airfare and that too at a higher level. The systems taken into considerations are all mentioned in the literature survey.

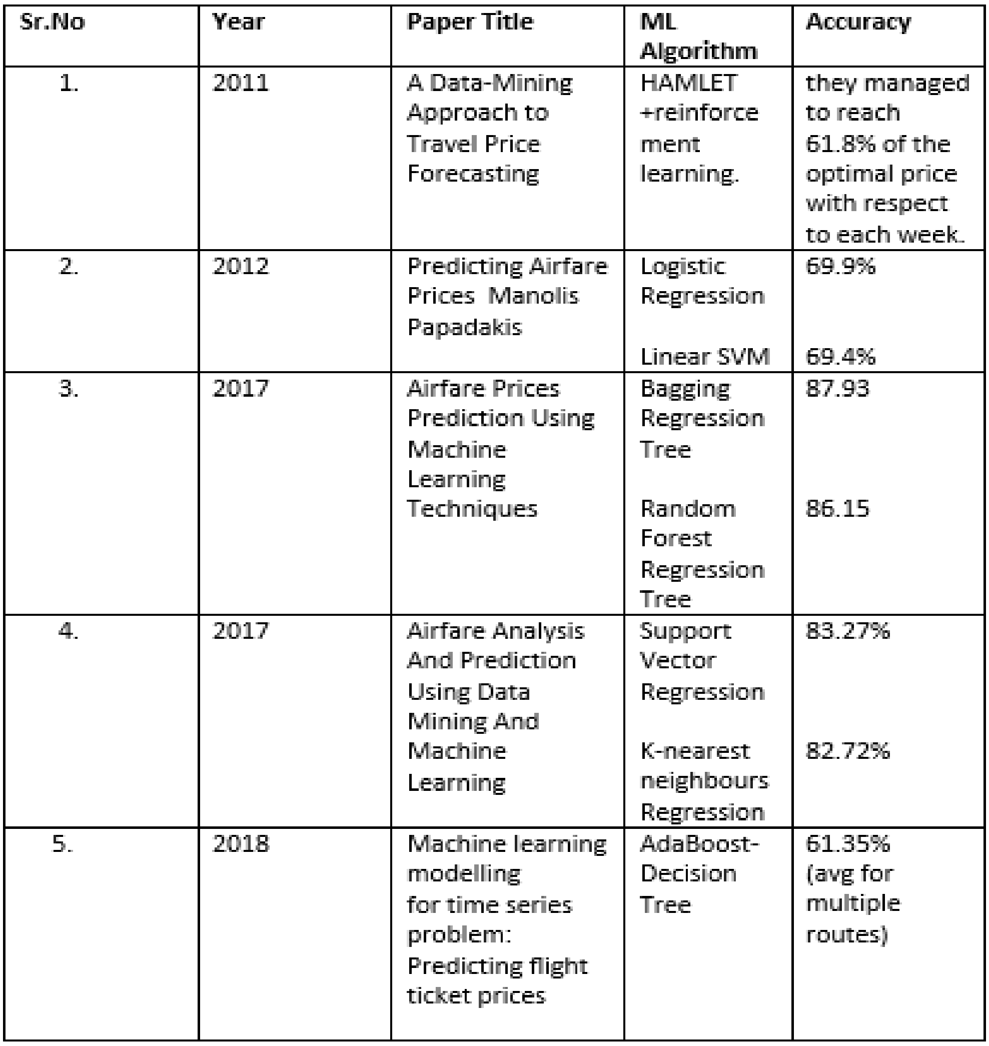
Limitations:

The existing systems doesn't provide any severe drawbacks but it did have certain limitations

a. The system doesn't have sufficient data for better prediction.

b. The system changes accuracy with changing algorithm and so it becomes a bit confusing though the accuracy only changes much when features important features are removed.

Table 1 Comparing ML algorithms based on accuracy



Sr.No	Year	Paper Title	ML Algorithm	Accuracy
1.	2011	A Data-Mining Approach to Travel Price Forecasting	HAMLET +reinforce ment learning.	they managed to reach 61.8% of the optimal price with respect to each week.
2.	2012	Predicting Airfare Prices Manolis Papadakis	Logistic Regression Linear SVM	69.9% 69.4%
3.	2017	Airfare Prices Prediction Using Machine Learning Techniques	Bagging Regression Tree Random Forest Regression Tree	87.93 86.15
4.	2017	Airfare Analysis And Prediction Using Data Mining And Machine Learning	Support Vector Regression K-nearest neighbours Regression	83.27% 82.72%
5.	2018	Machine learning modelling for time series problem: Predicting flight ticket prices	AdaBoost-Decision Tree	61.35% (avg for multiple routes)

III. METHODOLOGY

To develop a web application for "Airfare Prediction" based on previous airline ticket sales dataset for improving sales in Indian Domestic Airline. Our main motive is to provide the client with a prediction system from which it can take a right decision of increasing or decreasing the Airfare so that the flight doesn't go empty or no money is lost due to sudden increase in crude oil.

a. To perform data analytics on customer's ticket booking data for a brief amount of time.

b. To refine the data i.e. Removing duplicate records, ambiguity etc.

c. To perform Feature engineering in order to extract important feature from dataset for prediction.

d. To Brainstorm the Features i.e. to decide how to use those features

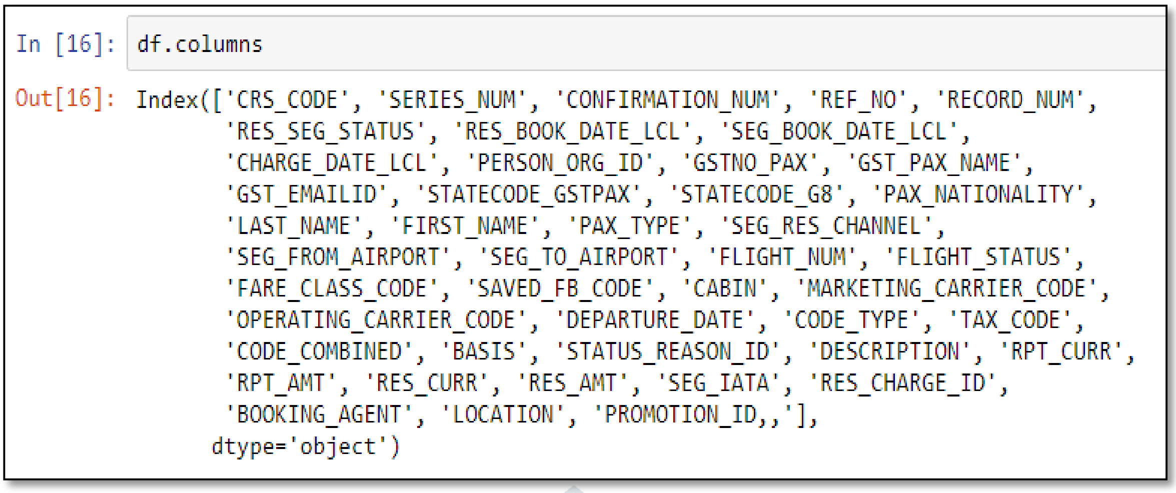
e. To create features i.e. to derive new features from those useful features.

The proposed system is composed of four phases:

1. Data input
2. Feature extraction
3. Machine learning model selection
4. Prediction

Phase 1: Data Input

The input data file is in .csv file will be provided to system and that input file contains all customer ticket booking information. The training data contains 45 columns from which important features are extracted. The information is limited to domestic airline.



```
In [16]: df.columns

Out[16]: Index(['CRS_CODE', 'SERIES_NUM', 'CONFIRMATION_NUM', 'REF_NO', 'RECORD_NUM',
                'RES_SEG_STATUS', 'RES_BOOK_DATE_LCL', 'SEG_BOOK_DATE_LCL',
                'CHARGE_DATE_LCL', 'PERSON_ORG_ID', 'GSTNO_PAX', 'GST_PAX_NAME',
                'GST_EMAILID', 'STATECODE_GSTPAX', 'STATECODE_G8', 'PAX_NATIONALITY',
                'LAST_NAME', 'FIRST_NAME', 'PAX_TYPE', 'SEG_RES_CHANNEL',
                'SEG_FROM_AIRPORT', 'SEG_TO_AIRPORT', 'FLIGHT_NUM', 'FLIGHT_STATUS',
                'FARE_CLASS_CODE', 'SAVED_FB_CODE', 'CABIN', 'MARKETING_CARRIER_CODE',
                'OPERATING_CARRIER_CODE', 'DEPARTURE_DATE', 'CODE_TYPE', 'TAX_CODE',
                'CODE_COMBINED', 'BASIS', 'STATUS_REASON_ID', 'DESCRIPTION', 'RPT_CURR',
                'RPT_AMT', 'RES_CURR', 'RES_AMT', 'SEG_IATA', 'RES_CHARGE_ID',
                'BOOKING_AGENT', 'LOCATION', 'PROMOTION_ID,,'],
               dtype='object')
```

Figure 1 List of columns present in the dataset

Phase 2: Data Cleaning

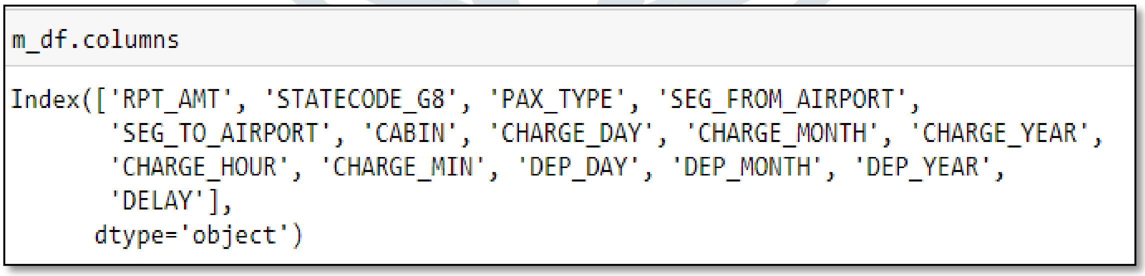
According to the research 80% of the work is done in cleaning data and retrieving useful information from it. As the data is collected from live public domain i.e. Airline industry it contains many Null values, redundant entries, merged values, referential features and many unnecessary columns. Data cleaning steps are as follows :

1. Removing null values
2. Formatting of date columns
3. Removing outliers
4. Conversion of object, string and other data types into numeric form (Encoding).

Phase 3: Feature Extraction

During this phase most of the informative features from the airline dataset that determines the prices of the air tickets are extracted. Features that can be considered are as follows:

Feature 1: Booking date and time
Feature 2: Departure date and time
Feature 3: Numbers of days till flight departure
Feature 4: Category of passenger (Adult/Child)
Feature 4: Cabin (Economy/Business)
Feature 5: Source Location
Feature 6: Destination Location



```
m_df.columns

Index(['RPT_AMT', 'STATECODE_G8', 'PAX_TYPE', 'SEG_FROM_AIRPORT',
       'SEG_TO_AIRPORT', 'CABIN', 'CHARGE_DAY', 'CHARGE_MONTH', 'CHARGE_YEAR',
       'CHARGE_HOUR', 'CHARGE_MIN', 'DEP_DAY', 'DEP_MONTH', 'DEP_YEAR',
       'DELAY'],
      dtype='object')
```

Figure 2 Extracted Features for training the model.

Phase 4: Machine Learning Model Selection

Machine learning is a science that uses statistical techniques to give computer system ability to learn from the given dataset without being explicitly programmed. The supervised learning algorithm deals with labeled data set training for predicting the results. Our system will be provided with label dataset and it is expected to predict the new input data. Therefore, we will use supervised machine learning algorithm. [8]. For continuous flight fare changing data we will use regression machine learning models which are as follows:

1. Random Forest regression tree.
2. Logistic regression.
3. Decision Tree.
4. Support Vector Machine.
5. Linear regression.

Phase 5: Prediction and Evaluation

When the input file or input data will be provided to trained ML model then it will predict some output which will be compared with the expected outcome. If outcome matches with the expected output then it will be accepted else it will again be given to the ML model [7].

IV. RESULT AND DISCUSSION

The main objective of this paper is to analyze the dependency of air prices on different features and build a prediction model that could help Airline industry to predict price of Air ticket and gain maximum profit. The following is the result of analysis done on the given dataset for domestic Airline using jupyter-notebook and it uses various python libraries like pandas, numpy and scikit-learn.

1. Correlation among the variables.

The term "correlation" refers to a mutual relationship or association between quantities. It decides prediction of one quantity from the other. Figure 2 shows dependencies among columns and this output gives a broader look in deciding the columns in model preparation.

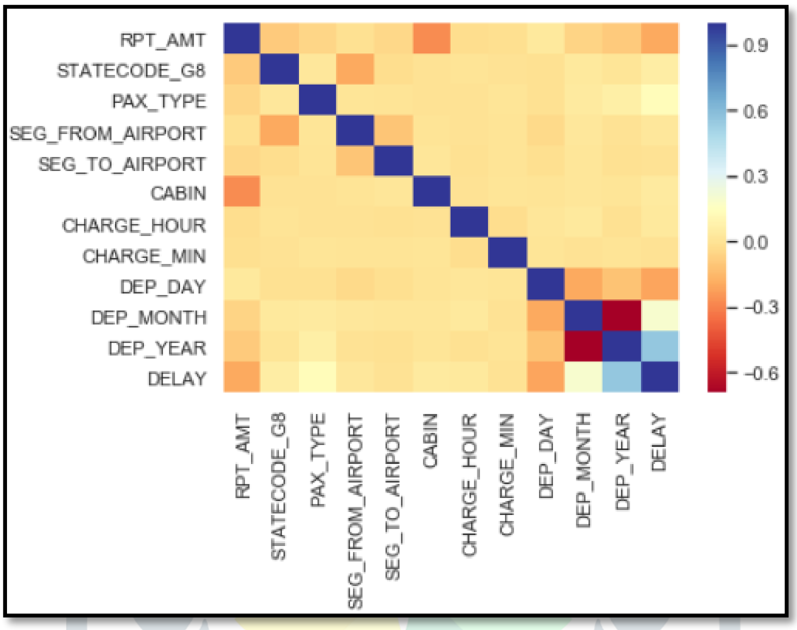


Figure 3 Correlation

2. Random Forest Regression Algorithm

In order to implement Random forest regression tree we used number of estimators as 1000 and number of random states were 42. This algorithm is well suited for unstructured data where dependencies among the features are quite difficult to identify.

Table 2 Results of Random Forest Regression Tree

Performance Measure	Values
Mean Absolute Error	1232.72721
Mean Squared Error	9751483.56970
Root Mean Squared Error	3122.7365

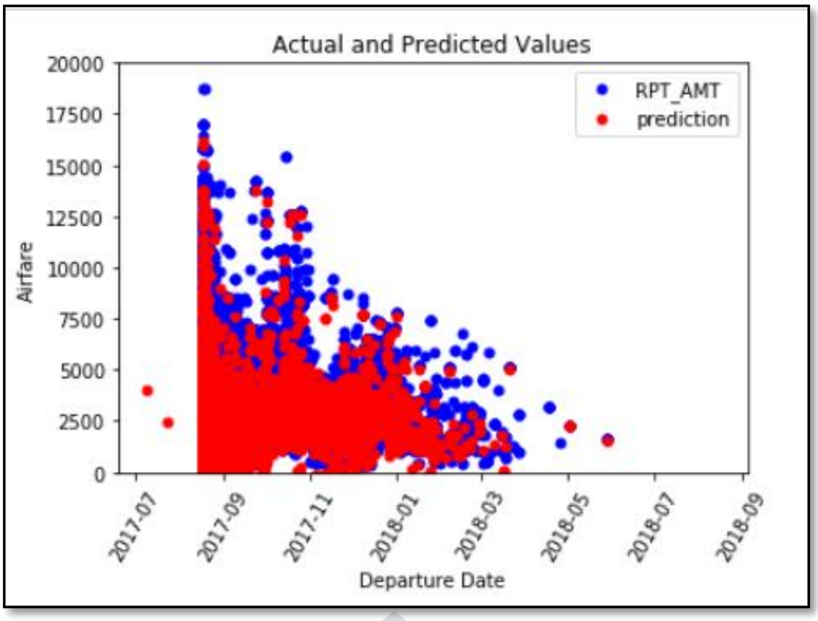


Figure 4 Graph of Train data Vs. Test data

3. Decision Tree

In decision tree there is a tree like model of all decisions and their possible consequences. For implementing we took number of states as 42 which means that there will be total 42 branches in the tree.

Table 3 Result of Decision Tree

Performance Measure	Values
Mean Absolute Error	768.7658
Mean Squared Error	2004395.2946
Root Mean Squared Error	1415.7666

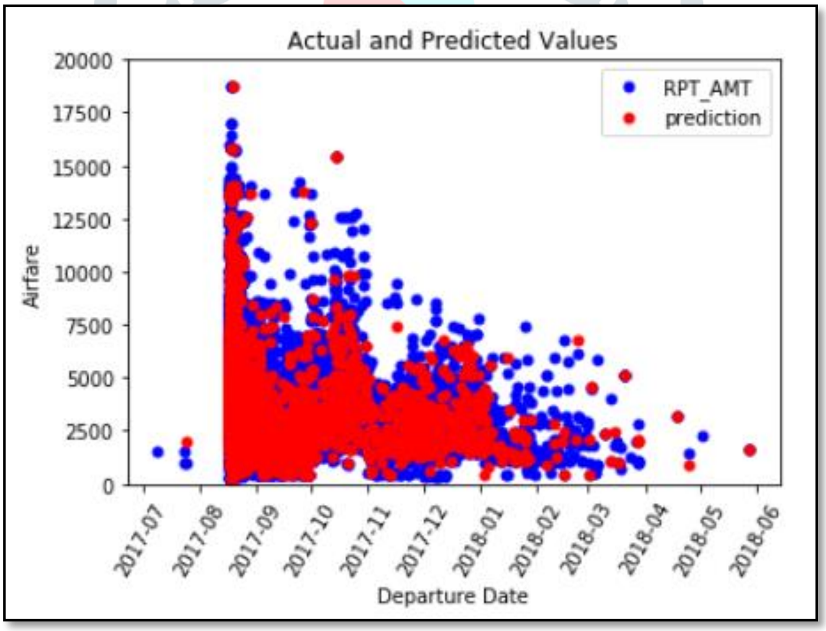


Figure 5 Graph of Train data Vs. Test data

4. Support Vector Machine (kernel = rbf)

Support vector machine is a supervised learning model that analyses the data for classification and regression. Here we used different parameters of SVR function like gamma='scale', kernel='rbf', degree=3.

Table 6 Result of SVR

Performance Measure	Values
Mean Absolute Error	657.57318
Mean Squared Error	14355128.8023
Root Mean Squared Error	3788.8162

V. CONCLUSION AND FUTURE SCOPE

This paper reported a study on airfare prediction using data analytics and machine learning. We gathered dataset from Domestic Indian Airlines of one month and applied data cleaning, features extraction and finding correlation [4].The experimental result shows what features are necessary for developing a prediction model and how they are interrelated with each other. To train the model different ML algorithms are used such as Random Forest, Decision Tree and Support Vector Regression.

Mean Error, Mean Squared Error and Root Mean Squared Error are least in Random forest and Decision Tree Algorithm from it we can get more accuracy.

Apart from selected features other factors can be also involved to improve the accuracy of the model. In future this proposed model can be trained for more than one year of data, if this is extended further then the processing speed as well as the power of the computer required will be more so this can be implemented using Big data analytics to improve sales in business services[3]

REFERENCES

[1] A regression model for predicting optimal purchase timing for airline tickets.

[2] Airfare Prices Prediction Using Machine Learning Techniques K. Tziridis, Th. Kalampokas, G.A. Papakostas HUMAIN-Lab

[3] International Journal of Computer Science and Mobile Computing “big data analysis of airline data set using hive” by p. swathi1, j. kumari2.

[4] International journal of engineering science invention “airfare analysis and prediction using data mining and machine learning” by bhavuk chawla1,ms.chandandeep kaur2.

[5] Machine learning modeling for time series problem: Predicting flight ticket prices Jun Lu,

[6] Airline Data Set,United States Department of Transportation, Office of the Assistant Secretary for Research and Technology,BureauofTransportationStatistics,http://www.tr anstats.bts.gov/DL_SelectFields.asp?Table_ID=236

[7] William Groves and Maria Gini, ”On Optimizing Airline Ticket Purchase Timing”, University of Minnesota, 2011

[8] ManolisPapadakis, “Predicting Airfare Prices” in Stanford,2013

[9] A Data-Mining Approach to Travel Price Forecasting

[10] WEKA Manual for Version 3-6-8, The University of Waikato, 2012