International Journal of Engineering Science Invention
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org ||Volume 6 Issue 11|| November 2017 || PP. 10-17

Airfare Analysis And Prediction Using Data Mining And Machine Learning

Bhavuk Chawla[1],Ms. Chandandeep Kaur[2]

[1](Bachelor of Technology, Department of Computer Science and Engineering,Dr. B.R. Ambedkar National Institute of Technology Jalandhar, India)

[2](Assistant Professor, Department of Computer Science and Engineering,Dr. B.R. Ambedkar National Institute of Technology Jalandhar, India)

Abstract: *As domestic air travel is getting more and more popular these days in India with various air ticket booking channels coming up online, travelers are trying to understand how these airline companies make decisions regarding ticket prices over time. Unfortunately this dynamic pricing strategy is usually carried out programmatically and is based on certain hidden parameters (e.g. number of days left till flight departure, or number of seats left). The paper works on mining the previous airfare data and developing data modeling technique to predict the price variation over time so that the consumer could benefit from it. This paper documents a study conducted to understand the airfare dependency over many hidden variables of which oil price, week day of departure, number of stops still have not received much attention from the research community, it also describes the two different methodologies adopted to model this price change, comparative analysis of algorithms under these two methodologies, applied on real world data has also been performed. The comparative analysis thus helped us to find out the most effective algorithm for the prediction of the airfare variations. The study suggests that mining historical airfare data and modeling using machine learning algorithms can help predict price trend and save consumers substantial sum of money.*

Keywords: *airfareprediction; classification;comparative analysis; data mining; machine learning; regression*

Date of Submission: 07-11-2017 Date of acceptance: 17-11-2017

I. Introduction

Almost all airline companies base their ticket price on demand estimation models and implement various dynamic pricing strategies in order to regulate seats demand and maximize their revenue. These corporations are said to use some proprietary software to evaluate ticket price per seat on a given day for a particular flight but the algorithms used are guarded with commercial secrets. These companies usually tie up with various online ticket sale channels (yatra.com, makemytrip.com, paytm.com) which maintains real time data on ticket price and constantly updates this price per seat over time. These channels are usually available over the internet where the traveler can book the ticket conveniently paying some convenience charges. This constant updating of prices results in high fluctuation which often confuses consumers as to when book their flight tickets to get best of the deals.

Previous research work carried on the same problem domain have not yet considered certain factors like oil price on a particular day and day of travel. Moreover not much study has been done on the Indian domestic air market. The study documented here {1} explains dependency of airfare over certain important factors {2} proposes two different methods adopted to model this price trend and {3} describes our implementation, simulation and performance of various algorithms on test data. The study was done following the standard CRISP-DM model with real airfare data extracted and collected over a span of 2 months from an Indian online air ticket sale channel.

The remainder of paper is organized as follows. Section 2 describes our data collection mechanism, data preparation and discusses dependency of airfare on various factors. Section 3 introduces two methods we designed to model the price changes, various algorithms used under these two methods and their performance on test data while section 4 describes a prototype application designed to help consumers in their buying decisions. Finally we conclude with a little discussion of possible future work.

II. Data Preparation

In this section we explain various factors considered in this study, our data collection mechanism, data preparation and finally the analysis of factors' impact on airfare changes.

2.1 Factors Considered

There are a number of factors upon which airfare may depend of which we took into account five parameters to conduct our study namely {1} number of days left till flight departure {2} oil price on a particular day {3} week day of departure (e.g. Monday, Tuesday) {4} number of intermediate stops {5} number of competitors on the route.

2.2 Data Collection

The data was collected from a number of sources. For the airfare data we scrapped it from a major online flight ticket booking website (Yatra.com) [17]. In order to collect a large amount of data, a PHP [22] script was built, which ran at scheduled intervals, data from the website was scrapped and stored on an online database. Seven continuous dates were selected (corresponding to seven days of week i.e. Monday, Tuesday etc.). These were the dates when the flights would depart and data for 2 months continuous was collected. This helped us to consider the effect of week day on the price fluctuations. Six different routes were considered each having different number of competitors. The considered routes and the number of competitors on the corresponding routes are given in table I. Then we built a script that ran once each day for continuous two months extracting the minimum ticket price for each route, for each of seven days of travel and store it on an online database. We collected approximately 20000 tuples of airfare data. However, for this study we restricted ourselves to only economy class ticket prices.

Table I. Routes and number of competitors

Route	Number of competitors
New Delhi To Aurangabad	2
Amritsar to New Delhi	3
New Delhi To Ranchi	4
New Delhi to Ahmedabad	6
New Delhi to Kolkata	7
New Delhi to Mumbai	8

We collected historical oil price data from Macrotrends [18] and stored it in another database. While for the number of intermediate stops of flights we collected data manually from various airline companies' websites.

2.3 Data Preparation

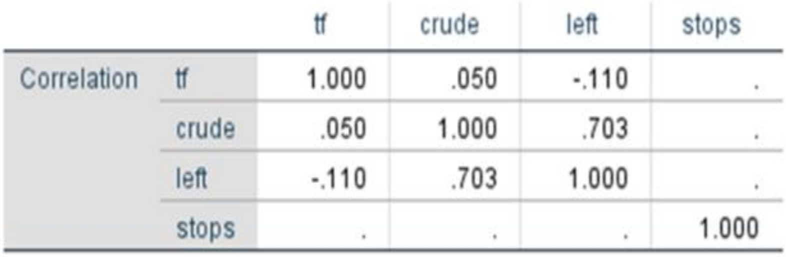
As the data was collected from a number of sources and it contained redundant data as well as missing data the data was integrated, transformed and cleaned to construct a final dataset from the initial raw data. One of the important steps during data preparation was normalizing the airfare data in the range 0-100 as the ticket prices are different on different routes. We used the equation (1) for normalizing the ticket prices.

$$Xnew = ((Xold - Xmin)/(Xmax - Xmin))*100 \qquad (1)$$

Xnew is normalized price, Xold is normal price, Xmin is the minimum price observed in dataset while Xmax is maximum price observed in dataset.

2.4 Factor Analysis

Factor analysis is a statistical method used to describe variability among observed, correlated variables to check which factors are more important than other. Factor Analysis was done in order to check dependency of airfare on factors considered. We used SPSS [20] software to build a correlation matrix showing correlation between various factors. The correlation matrix between factors for the Delhi-Kolkata route is shown in Fig 1.



		tf	crude	left	stops
Correlation	tf	1.000	.050	-.110	.
	crude	.050	1.000	.703	.
	left	-.110	.703	1.000	.
	stops	.	.	.	1.000

Figure 1. Correlation matrix

Here “tf” is the airfare, “crude” refers to crude oil, and “left” corresponds to number of days left till departure. A positive value of correlation between two factors shows that they are directly correlated or in other words if one increase other increases and vice versa. Negative values show the inverse correlation.. The matrix provides a rough idea of how the ticket price is dependent on crude oil price and number of days left. For example the matrix shows negative correlation between the number of days left and ticket price so as the number of days left decreases the ticket price increases. The same is evident in the graph shown in Fig 2 between number of days left and ticket price. The graph suggests that ticket price is at peak, 2 weeks before departure and the magnitude of fluctuations are at max within this period.

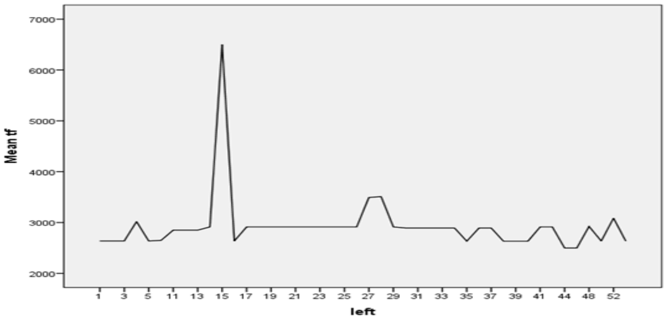


Figure 2. Price change over time for flight on route Delhi-Kolkata departing on 13-feb-2017. The figure shows how price fluctuates before 55 days of departure.

Similarly a positive correlation between crude oil price and ticket price suggests that as crude oil price increases the ticket price increases. The same is depicted in the graph in Fig 3.

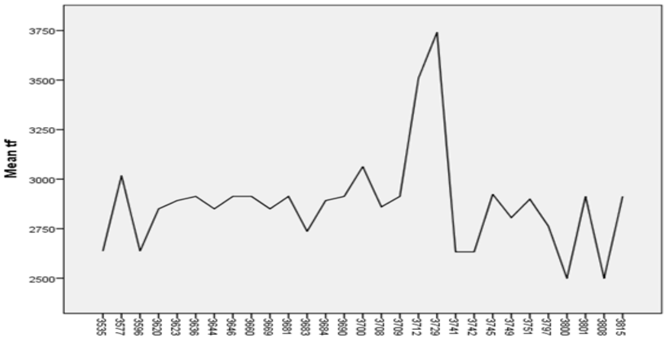


Figure 3. Price change over time for flight on route Delhi-Kolkata departing on 13-feb-2017. The figure shows how ticket price varies as the crude oil price changes.

Further analysis of factors brought out some interesting relations, one of which is depicted in Fig 4 where one can see that ticket price appears to fluctuate often when the number of competitors are few as compared to when the number of competitors are more. This suggests that more competition on the same route leads to less ticket price variations. In Fig 4 ticket price remains almost constant on the route having 7 competitors.

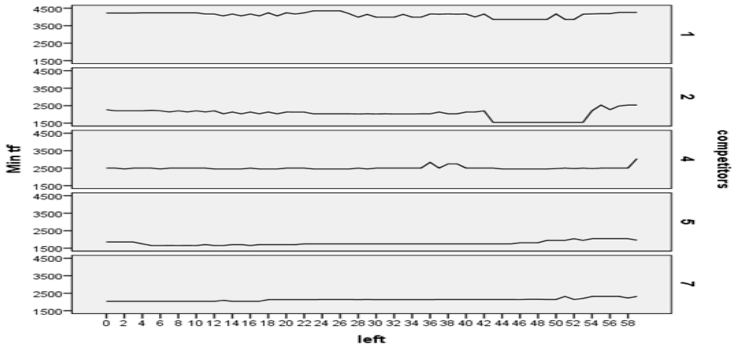


Figure 4. Price change over time for flight departing on 13-feb-2017 vs number of days left on different routes with different number of competitors.

Data collected over span of 60 days had mainly flights with 0 stops in dataset, flights with 1 stop were few and flights with 2 or more stops were almost negligible. So the effect of number of stops on price fluctuations could not be studied much. This could be because almost all the flights available on the travel booking site were nonstop.

III. Data Modelling

In this section we explain the two approaches we used to train and test various machine learning algorithms on the dataset. We then report the performance of the algorithms and make a comparative analysis. Our data consisted of price observations recorded every day for all the six routes and for each day of travel collected over a span of 60 days. We had approximately 6x7x60 i.e. 2520 tuples of data. Each tuple is a vector of following features:

- Number of days left till departure (numeric value 1-60)
- Day of travel (numeric value 0-6)
- Crude Oil price
- Number of competitors on route
- Number of stops

These tuples of features constituted the X set or the independent set while the dependent set or Y set consisted of price observations normalized to range of 0-100.

Motive was to build a model which could help consumer make a decision on whether to buy ticket or should wait for particular time point, for a particular route and flight given the historical fare data. Two different supervised learning approaches have been proposed to build these models. [24].

3.1 Regression based modelling

Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable (s) (predictor). Regression tries to predict a real valued output (numerical value) of some variable based on the relationship discovered. Here the dependent variable is the X set while the independent variable is the normalized ticket price. So under this approach we used a number of regression based algorithms, built and tested models using K-fold cross- validation [23]. We used 10-fold cross-validation here.

Procedure

Input:D: dataset; Tnd: training data set; Tsd: testing data set; M: model; Xnd: vector of features (tuple) in training data; Xsd: vector of features (tuple) in testing data set; Ynd: independent variable (airfare in training data set); Ysd: independent variable (airfare in testing data set); t: tuple; Acc: accuracy; temp: temporary variable; lensd = number of tuples in testing data set; R: regression algorithm

Output:

```
Split D into 10 equally sized sets.
Loop1 : For i in 1-10:
    Tsd = D[i]; Tnd = D – D[i];
    M = build model using some R
    Train M on Tnd
    Loop 2: For t in Xsd:
        Predict Ysd; Store Ysd;
    End Loop2
    Transform Ynd into class labels “yes” or “no”
    Transform Ysd into class labels “yes” or “no”
    Count number of matching labels i.e. when Ynd[j]=Ysd[j] and store in temp
    Acc = Acc + (temp/lensd)*100
End Loop1
Acc = Acc/10
```

Acc is the final accuracy achieved with model M built using R In other words Acc gives the percentage of test results predicted correctly by the model M.

Regression algorithms used:

1. Support Vector Regression
2. Random Forest Regression
3. Gradient Boosting Regression

4. AdaBoost Regression
5. K-nearest neighbors Regression

3.2 Classification based modelling

Classification is a form of modelling technique in which we take an existing dataset and using it generate a predictive model which learns to classify data points into certain classes, later on this model is used to classify future data points. Here data points are tuples of features and class labels “yes” and “no” are classes corresponding to whether user should buy ticket or not. Similarly under this approach we used a number of classification based algorithms, built and tested models using 10-fold cross-validation.

Procedure

Input:D: dataset; Tnd: training data set; Tsd: testing data set; M: model; Xnd: vector of features (tuple) in training data; Xsd: vector of features (tuple) in testing data set; Ynd: class label in training data set; Ysd: class label in testing data set; t: tuple; Acc: accuracy; temp: temporary variable; lensd = number of tuples in testing data set; C: classification algorithm

Output:

```
1.  Split D into 10 equally sized sets.
2.  Loop1 : For i in 1-10:
3.  Tsd = D[i]; Tnd = D – D[i];
4.  Transform Ynd into class labels “yes” or “no”
5.  M = build model using some C
6.  Train M on Tnd
7.  Loop 2: For t in Xsd:
8.  Predict Ysd; Store Ysd;
9.  End Loop2
10. Count number of matching labels i.e. when Ynd[j]=Ysd[j] and store in temp
11. Acc = Acc + (temp/lensd)*100
12. End Loop1
13. Acc = Acc/10
```

Acc is the final accuracy achieved with model M built using C and gives the percentage of test results predicted correctly by the model M.

Classification algorithms used were:

1. AdaBoost Classification
2. Random Forest Classification
3. Support Vector Machine
4. Naïve Bayes Classification
5. K-nearest neighbors Classification
6. Logistic Regression

The scripts were coded in python language while using sckit-learn [21] library for implementation of various machine learning algorithms.

3.3 Comparitive Analysis

The percentage accuracies obtained for the algorithms investigated above are given in table II. Some of the top performing algorithms were naïve-bayes and SVM with accuracies of 84.01% and 83.97% respectively. An accuracy of 84% is much better than 74.5% obtained in [6].

Table i. Comparitive analysis

Algorithm	Category	Accuracy
Support Vector Regression	Regression	83.27%
Random Forest Regression	Regression	82.11%
Gradient Boosting Regression	Regression	80.66%
AdaBoost Regression	Regression	56.57%
K-nearest neighbors Regression	Regression	82.72%
AdaBoost Classification	Classification	82.66%

Random Forest Classification	Classification	73.4%
Support Vector Machine	Classification	83.97%
Naïve Bayes Classification	**Classification**	**84.01%**
K-nearest Classification	Classification	82.27%
Logistic Regression	Classification	81.81%

IV. Deployment

The goal of this study was to analyze the dependency of airfare on certain factors and build some data modelling technique (which is accurate up to an appreciable degree) that could help consumers on their buying strategy. After the study was conducted successfully, a prototype application was built using python Qt [25] which ran the naïve-bayes model in the backend for predicting whether the user should buy the ticket or wait for a certain amount of days after the user gave input as route and date of departure. The application was in the form of a desktop app that could run on any platform having python environment and required libraries installed.

Oil price prediction

It is interesting to note that while evaluating various algorithms we used the collected oil price data in the dataset for training and testing which was right but if we build an application like discussed before we must generate or predict the oil price on a future date before predicting what would be the airfare on that date and making a buying decision. Hence again we have to apply some modelling technique that could accurately predict the oil price on a future date. Number of algorithms on historical oil price data of 1 year and finalized Support Vector Regression (SVR) for predicting future oil price as it gave the maximum accuracy. An image of how SVR would fit curve in the historical oil price data is shown in Fig 5. So before predicting the buying decision in above application, the script first predicted the future oil price and then using other features together ran the naïve-bayes model to generate a price trend for the flight. The price trend was analyzed programmatically in the background and finally suggested the user whether he/she should buy or wait.

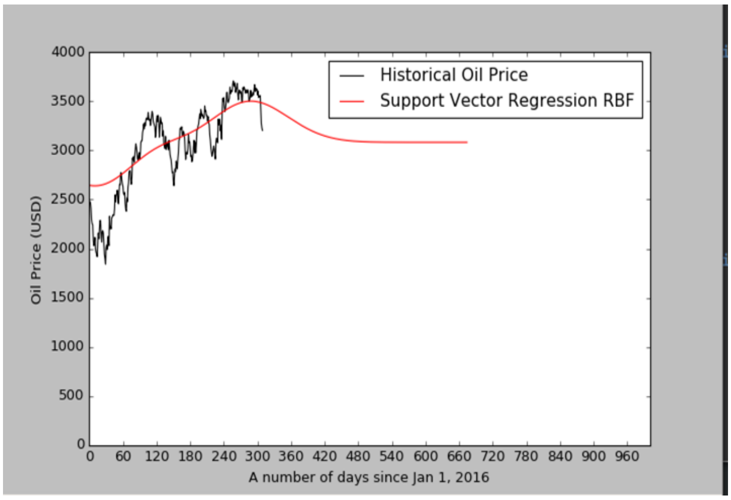


Figure 5. SVR modelling on crude oil price data of year 2016

V. Conclusion And Future Work

The paper reported on a study conducted to understand the airfare changes using historical data extracted from web. We collected data from various sources, cleaned, integrated and transformed data to generate a final dataset from raw unorganized data. Later on we investigated a number of algorithms under the classification and regression approaches which could be used to model this problem. Despite of the complex commercial algorithms used by airline corporations and the absence of data on key potential parameters like number of seats left on flight our data modelling technique performed surprisingly well. We were able to achieve an accuracy of 84% using naïve-bayes algorithm. We owe this high accuracy to the fact that an appreciable amount of previous airfare data was collected and some key factors like oil price data, day of travel were taken in to account which were not much considered in previous research works done in this problem domain. Secondly the study here is one of the few done on Indian domestic air travel market.

However it is interesting to note that getting good performance on a metric like above might not actually translate to good performance on commercial application as the oil price prediction too have some limitations. One can further expand on this research work by considering the factors on which oil price depends which might further lead to better accuracy in predicting future oil prices. Another thing which could receive attention is taking into account festive seasons while predicting airfare. For the study documented here, impact of festive seasons on airfare was not taken into account. However during the period study was conducted there were no major festivals.

Similar studies in other domains like hotel, auctions could be done but this initial study shows the potential of scrapping previous price data and applying data mining and machine learning algorithms to model price change and helping make smarter buying decisions. We believe price mining of this sort is a fertile area of future research.

References

Journal Papers:

[1]. Yuwen Chen, Jian Cao, Shanshan Feng and Yudong Tan, "An ensemble learning based approach for building airfare forecast service" Big Data (Big Data), 2015 IEEE International Conference, 29 Oct.-1 Nov. 2015.

[2]. AnastasiiaGordiievych and Igor Shubin, "Forecasting of Airfare Prices Using Time Series" in Information Technologies in Innovation Business (ITIB), 7-9 October, 2015, Kharkiv, Ukraine, pp. 68-71.

[3]. Till Wohlfarth, St´ephanCl´emenc¸on, Franc¸oisRoueff and Xavier Casellato, "A Data-Mining Approach to Travel Price Forecasting" in 10th International Conference on Machine Learning and Applications, 2011, pp. 84-89.

[4]. O. Etzioni, R. Tuchinda, C. A. Knoblock, and A. Yates, "Mining airfare data to minimize ticket purchase price," in ACM SIGKDD, ser. KDD'03. New York, NY, USA: ACM, 2003, pp. 119–128.

[5]. Currie, Cheng, HKSmith, "Dynamic Pricing of airline tickets with competition", Journal of the Operational Research Society, 2007, 5, pp.1-12.

[6]. ManolisPapadakis, "Predicting Airfare Prices" in Stanford, 2013

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

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

Chapters in Books:

[9]. Jiawei Han, MichelineKamber, Jian Pei, "Getting to know your data", in Data Mining Concepts and Techniques, 3rd ed. (Upper Saddle River New Jersey: Pearson), ch 2, pp. 39-78.

[10]. Nong Ye, "Naïve Bayes classifier", in Data Mining Theories Algorithms And Examples, (New York: CRC Press), ch 3, pp. 31-36.

[11]. Efraim Turban, Ramesh Sharda, Dursun Delen, "Data Mining for Business Intelligence ", in Decision Support and Business Intelligence Systems, 9th ed. (Upper Saddle River New Jersey: Pearson), ch 5, pp. 190-230

[12]. Jiawei Han, Micheline Kamber, Jian Pei, "Data preprocessing", in Data Mining Concepts And Techniques, 3rd ed. (Upper Saddle River New Jersey: Pearson), ch 3, pp. 83-117

[13]. Nong Ye, "k-Nearest Neighbour Classifier and Supervided Clustering", in Data Mining Theories Algorithms and Examples, (New York: CRC Press), ch 7, pp. 117-136

[14]. Jiawei Han, Micheline Kamber, Jian Pei, "Classification: Advances Methods", in Data Mining Concepts and Techniques, 3rd ed. (Upper Saddle River New Jersey: Pearson), ch 9, pp. 393-426

[15]. Raúl Garreta , Guillermo Moncecchi, "Machine Learning – A Gentle Introduction", in Learning scikit-learn: Machine Learning in Python, (Birmingham, UK: PACKT), ch 1, pp. 5-21

[16]. Raúl Garreta , Guillermo Moncecchi, "Supervised Learning", in Learning scikit-learn: Machine Learning in Python, (Birmingham, UK: PACKT), ch 2, pp. 25-58

Online Webpages:

[17]. Fox Travels, https://www.foxworldtravel.com

[18]. Yatra travel booking website, https://www.yatra.com

[19]. Macrotrends , http://www.macrotrends.net

[20]. Weka Tool http://www.cs.waikato.ac.nz/ml/weka

[21]. SPSS Predictive Analytics Software https://www.spss.co.in.

[22]. scikit-learn Machine Learning in Python https://www.scikit-learn.org

[23]. PHP Hypertext Preprocessor http://www.php.net

[24]. K-fold cross-validation https://goo.gl/iF5t58

[25]. Supervised Learning https://goo.gl/hPkcfR

[26]. Python Qt https://wiki.python.org/moin/PyQt

International Journal of Engineering Science Invention(IJESI) is UGC approved Journal with Sl. No. 3822, Journal no. 43302.

Bhavuk Chawla Airfare Analysis And Prediction Using Data Mining And Machine Learning." International Journal of Engineering Science Invention(IJESI), vol. 6, no. 11, 2017, pp. 10-17.