INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 5, Issue 3)
Available online at: www.ijariit.com

Analysis of dynamic pricing in airlines and predicting least fare

Purbid Bambroo
bambroopurbid@gmail.com
BMS College of Engineering, Bengaluru, Karnataka

Sheetal
1bm15cs095@bmsce.ac.in
BMS College of Engineering, Bengaluru, Karnataka

Nitin Agrawal
1bm15cs066@bmsce.ac.in
BMS College of Engineering, Bengaluru, Karnataka

Dr. Kavitha Sooda
kavithas.cse@bmsce.ac.in
BMS College of Engineering, Bengaluru, Karnataka

ABSTRACT

Airline companies have been using dynamic pricing to vary the ticket prices to maximize the profit for a limited number of seats. Though the algorithm is different for all the airlines and never disclosed, it is possible to predict the variation in ticket prices. There have been studies in the past for the same, none explicitly for the Indian market; considering the major holidays. Applying techniques from Machine learning model of neural networks and back-propagation, we could predict the upcoming surge or dip in the ticket prices. We aim at predicting if the price of the ticket will go down in the future or the current price is the lowest.

Keywords— *Neural networks, Machine learning, Airlines, Dynamic pricing*

1. INTRODUCTION

The ticket prices for a journey are a culmination of many underlying factors. Some of these (Base fare, oil fares, taxes, etc) are fairly similar for all airlines. But there are other deciding factors like route (source/destination), number of stops, aircraft model, time of the year (holiday season or not) which decide the fluctuation in prices of these tickets. While the airline companies have made up to believe there is no complex algorithm in deciding price fares and as the date of journey approaches, the fare increases, on analysis of historical data there can be interesting trends that are seen in fluctuation of prices, that show that price is not always inversely proportional to the number of days left. Most of the researches carried out in the field take into analysis domestic flights in the USA.

1.1 Data analytics

Data analytics is a process of finding information from data to decide and subsequently act on it. Data is information in raw format. With increasing data size, it has become a need for inspecting, cleaning, transforming, and modelling data with the goal of finding useful information, making conclusions, and supporting decision making. Like in this project, data of various parameters (Base fare, the current price of oil, taxes, etc) are gathered to make a particular decision on it.

1.2 Machine learning

Machine learning is an application of Artificial Intelligence (AI) that provides systems with the ability to automatically learn and improve from experience without being explicitly programmed. Machine Learning algorithm is trained using a training data set to create a model. When new input data is introduced to the ML algorithm, it makes a prediction on the basis of the model. In order to find the least price of a flight for a particular destination, we are using various ML models to reach a perfect accuracy

1.3 Regression analysis

Regression analysis is a powerful statistical method that allows you to examine the relationship between two or more variables of interest. Regression is used to predict continuous values. Regression analysis is most commonly used in forecasting or predicting how a set of conditions will impact an outcome. For example, you had a set of data on flight parameters and variables; you could build a simple linear regression model where oil price, tax, base fare and date are our variables. Then, once you have the model, and provided it is a good fit, we can use the model to predict minimum fare based on those parameters. Other sectors where machine learning can help is health care and network-enabled manufacturing (NEM) development for Boeing aircraft [1].

1.4 AdaBoost

AdaBoost is a type of "Ensemble Learning" where multiple learners are employed to build a stronger learning algorithm. AdaBoost works by choosing a base algorithm (e.g. decision trees) and iteratively improving it by accounting for the incorrectly classified examples in the training set.

Equal weights are assigned to all the training examples and a base algorithm is chosen. At each step of the iteration, the base algorithm is applied to the training set and weights of the incorrectly classified examples are increased. It is iterated n times, each time applying base learner on the training set with updated weights. The final model is obtained as the weighted sum of the n learners

2. FACTORS THAT INFLUENCE THE PRICE OF AIRLINE TICKET

For most of the researches carried out, the following criteria were considered while collecting data and training. It is to be noted that in no paper were all of them considered together, that too for the Indian subcontinent hence we would like to include a few more factors specific to India, which are mentioned later in the paper.

- Price of Oil (Historic as well as current)
- Flight Distance (the distance between source and destination)
- Competition (Number of flights flying on the same route)
- Timing of Purchase and flight (Number of days from date of purchase)
- Empty Middle Seats
- Fare Rules (are the flights refundable or non- refundable)

3. PRESENT WORK

3.1 Data collection

Most of the previous researchers collected data over a period of 2-3 months. Most of the previous works have considered scraping data of travel websites or airline's website to get the cheapest ticket between source and destination. [2] Collected the data from the database of liligo.com which stored the trends and patterns of what users searched over time. Another study used historical data from Bing travel, a portal that had the ticket prices for the last 42 days [3]. In order to get results for a large database and audience, the study took 6 source-destination pairs and then collected the data for them daily for over 2 months. Now, there are two schools of thoughts, one that the data should be collected manually and the other that it should be collected automatically from the internet with some script/program. While extracting data manually can be slow, it for sure will give the desired data format and fields, unlike an online dataset [4]. This was used for fight analysis in the United States. The program gave the data as a fare chart, which was then turned to a graph digitizer and later exported as a CSV file. In most of the surveys/data collection, the major focus remained on fetching one-way trip data and any layovers were ignored. In fact, only 5 of them took a round trip into account.

One of the research in 2017 uses a set of 8 parameters on which the prices of the airline tickets are assumed to be dependent [5]. Data were continuously collected in the period between July and December for one-way flights between source and destination, based on which the results were calculated.

In another research, the entire analysis part was done based on daily price quotes from a major aeroplane search web over a period of three hundred-days, over a set of 8 different routes. Using the technique of web crawling, each route information was procured (the departure date in specific). This crawler was run every day at a specific time to get more consistent results. In these two different approaches were used for the analysis:

- Specific problem approach: in this approach, all the data collected over time (historical) was considered for analytics.
- Generalized problem approach: This approach considered the case where the results from historic data weren't present and hence the results/analytics made was only on the basis of the currently available data.

3.2 Data preparation

After the data was collected, the most important fields were retained and used for analytics (using R language mostly).

Since the routes were different, one particular study used normalized values on a scale of 0-100. [7]. Following was the formula used for normalization.

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

Another paper derived a relation between the Cost Predicted(C) and the actual current price as C=BF [3], where B is the base fare and hence usually remains constant and F is the fluctuation on that base fare. There were various metrics considered to estimate the performance of the model. Namely:

- Random Purchase Price
- Optimal Price
- Predicted Price

Finally, the following formulae were derived for optimal and normalized Performance:

Performance = ((Random Purchase Price-Predicted Price)/Random Purchase price) %

Optimal Performance = ((Random Purchase Price- Optimal Price)/Random Purchase price) %

Normalize Performance = (Performance/Optimal Performance) %

3.3 Results from data analysis

To determine the various factors for determining the price, a correlation matrix was built using SPSS, a software package for interactive or batched statistical analysis. An interesting and worth noting trend in one of the other studies was that ticket prices see maximum fluctuation around 2 weeks before departure date and consequently they are even more expensive around this time. Another interesting trend was that more competitors on a route meant lesser fluctuations in prices [8]. We plan to derive more of such observations and trends from our analysis.

3.4 Data modelling

In most of the researches carried out, regression modelling[9][10][11] was used to establish a concrete relation between the independent variable (the normalized price) and the dependent variable (the price to be estimated).[12] showed the ticket fare to be estimated as a function of days until departure as peace wise continuous function of time $t \in [T(i)0 - 28, T(i)0]$, where T (i) 0 shows the departure time (number of days left). In regression-based modelling, the algorithms used mainly support vector regression, random forest regression, gradient boosting regression. Also, it was noted that regression-based models yielded better results than classification based models [2]. One research conducted also gave a 60% accurate results when considering only two factors: days left for departure and previous day's price. Another research gave 69.9% accuracy when using logistic regression, while 69.4% while using SVM (Support Vector Machine). Another research gave a much higher (83.97%) [7] Accuracy on using SVM. This can be because the latter research considered factors like fuel prices and the dataset differed for both the researches. (Former being domestic flights in the US and later Indian domestic flights).

Other researches carried out modelling using classification based algorithms, where a data point is classified using labels ('yes' and 'no') to indicate if the ticket should be brought at the current price or not. Algorithms used were Adaboost, SVM, Naive Bayes, K-nearest neighbour's classification. An impressive accuracy of 84.01% was recorded for the Naive Bayes algorithm [7].Another study used multilayer perceptron (MLP) and achieved an accuracy of 80.28%.

In [5] after set of rigorous experiments in which one or more parameters were considered at a time or eliminated, it was observed that the models of bagging regressive tree, and random forest regression tree had the highest efficiency or accuracy

levels in terms of predicting the actual prices which were as close to 88%.

3.5 Results

It was seen in general that accuracies were higher with regression models, even up to 84.5%. In particular, SVM stood as the second highest to the Naïve Bayes algorithm. To no surprise, linear regression could give only 57% accuracy while one bagging regression tree gave a high accuracy of 87.42%.

In [6], the authors suggest that EPFL Classification is the best model, which performed much better (61.35%) than booking at no fixed pattern. This algorithm also had no variation with respect to the routes chosen. (8 to be precise).

[13] shows that competition among airlines in more pronounce on longer routes than on shorter routes, while [14] shows that the fares are expected to increase by 2% every time a seat is sold, in bi-directional routes in the UK. If we were to look at the analysis with respect to different routes, Uniform Blending Classification is chosen as the best model with relatively high performance. On the other hand, the Q- Learning method got a relatively high performance as well. The HMM Sequence Classification based Adaboost-Decision Tree Classification model got a much better result over 12 new routes, which has 31.71% better performance than the random purchases [6].

4. IMPLEMENTATION

4.1 Shortcomings of current models

While a lot of quality work has been done in the given domain, there are still many areas which remain unexplored. Firstly most of the researches considered are pre-2013, which gives a 5-year gap from 2018. Given the rate at which travel industry has boomed in the last five years (part of whose credit goes to social media boom), it would be wrong to be still using these algorithms and techniques and expect equally accurate results.

Moreover, there are very few researches conducted for Indian airlines, which leaves us with a very huge market to tap into and get interesting trends from.

No research has been carried out for the Indian market that considers holidays and festivals while deciding prices. We plan to get during and analyze prices from 3 major holidays (Dussehra, Diwali and Christmas/New year). Also, we plan to use a vast majority of factors which have not been used together for research till now.

Lastly, almost all studies take into account only data where the minimum flight price is included. While this is convenient for the training and would give accurate results in the most number of cases, the problem here is on considering only the lowest fare, we are letting go of other airlines' prices for the day/route. This might give only a small set of airlines over time which means our results would not hold good for all the airlines in consideration, which lets variance creep in. We would consider not only the lowest but all the fares between the source and destination and see if we can get better results.

From the analytics point of view, we are particularly interested in finding how the prices vary near festivals like Diwali and New Year, to get a better insight into the dynamic pricing algorithm being used.

4.2 Our approach

We intend to first collect the data from a popular travel booking website for Indian flight market. The data collection involves getting flight prices for 6 cheapest flights for all the days considered over 3 separate routes. This is done for a period of minimum of 2 months to get a good amount of data. We explicitly consider holidays and festivals in the Indian market to see how it affects the ticket pricing.

Following this, the data is analyzed and using plots and graphs to draw conclusions on fluctuations of the prices.

There is a dependent variable (the price of the ticket) and independent variables (listed in section II). The dependent variable is plotted against the independent variables. We plan to use a multilayer back propagation neural network. The dataset would be divided into a number of epochs and fed into the input layer of this neural network. After several iterations and least deviation from the actual value of the fare, we stop and consider our model ready for deployment.

The following flow chart shows the workflow.

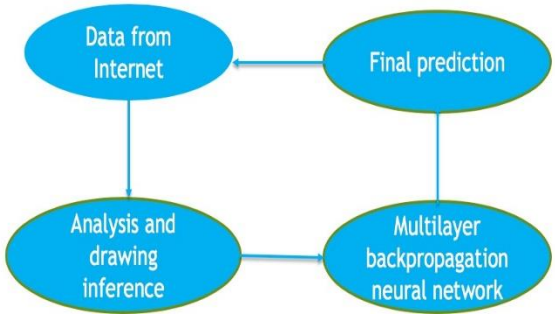

Fig. 1: Workflow

5. REFERENCES

[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] Till Wohlfarth, Stéphan Clémençon, François Roueff, Xavier Casellato A Data-Mining Approach to Travel Price Forecasting. HAL Id: Hal-00665041

[3] William Groves and Maria Gini. A regression model for predicting optimal purchase timing for airline tickets.

[4] Predicting airfare prices: Manolis Papadaki

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

[6] Jun Lu, Computer Science, EPFL: Machine learning modelling for time series problem; predicting flight ticket prices.

[7] Airfare Bhavuk Chawla1, Ms Chandandeep Kaur2: Analysis and Prediction Using Data Mining and Machine Learning. ISSN (Print): 2319 – 6726

[8] P. Malighetti, S. Paleari and R. Redondi, "Pricing strategies of low-cost airlines: The Ryanair case study," Journal of Air Transport Management, vol. 15, no. 4, pp. 195- 203, 2009.

[9] Kevin R. Williams DYNAMIC AIRLINE PRICING AND SEAT AVAILABILITY

[10] R. Preston McAfee, Vera te VeldeDynamic Pricing in the airline Industry. California Institute of Technology.

[11] Anastasia Lantseva, Ksenia Mukhina, Anna Nikishova, Sergey Ivanov and Konstantin Knyazkov Data-driven Modeling of Airlines Pricing. Procedia Computer Science Volume 66, 2015, Pages 267–276. YSC 2015.

[12] 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.

[13] Brander, J. A., & Zhang, A. (1990). Market conduct in the airline industry: an empirical investigation. The RAND Journal of Economics, 567-583.

[14] Marco Alderighi, Alberto A. Gaggero, Claudio A. Piga.The hidden side of dynamic pricing in airline markets. MPRA Paper No. 76977.

