IJIREEICE

ISSN (Online) 2321-2004
ISSN (Print) 2321-5526

International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering

Vol. 7, Issue 2, February 2019

Survey on Air Price Prediction using Machine Learning Algorithms

Abhilash[1], Ranjana Y[2], Shilpa S[3], Zubeda A Khan[4]

Assistant Professor, Department of Information Science & Engineering, Atria Institute of Technology, Bangalore, India[1]

Student, Department of Information Science & Engineering, Atria Institute of Technology, Bangalore, India[2,3,4]

Abstract: The existing airfare prediction method uses very complicated methods and algorithms for the prediction. They consider several financial and commercial factors and the prices changes dynamically which makes it difficult for customers to purchase the air ticket. Airlines implement dynamic pricing for their tickets, and base their pricing decisions on demand estimation models. The reason for such a complicated system is that each flight only has a set number of seats to sell, so airlines have to regulate demand. In the case where demand is expected to exceed capacity, the airline may increase prices, to decrease the rate at which seats fill. On the other hand, a seat that goes unsold represents a loss of revenue, and selling that seat for any price above the service cost for a single passenger would have been a more preferable scenario.

Keywords: Air fare, Random forest, Linear regression

I. INTRODUCTION

Machine Learning is an idea to learn from examples and experience, without being explicitly programmed. Instead of writing code, you feed data to the generic algorithm, and it builds logic based on the data given. The process of training an ML model involves providing an ML algorithm (that is, Machine Learning Algorithm) with training data to learn from. The data classification can be performed on structured or unstructured data. The main goal of classification is to identify the category/class to which a new data will fall under. We can use these ML models to get predictions on new data for which target is unknown.

II. GENRAL STRUCTURE

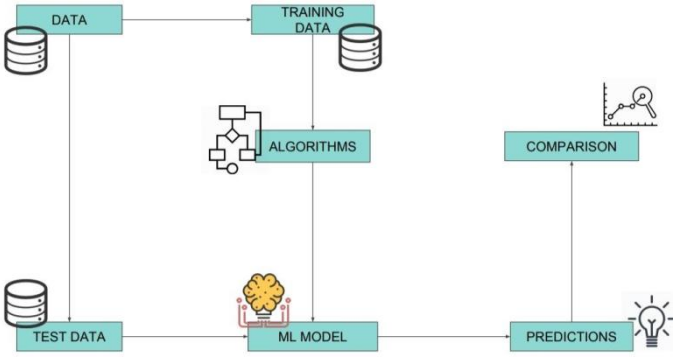

Fig : General Structure of ML Model

The above figure represents the general structure of an ML model. The data are split into Training and Test Data respectively. The Training Data is passed to trough the ML algorithms for enabling the machine to learn and apply it on the test data to predict the solutions. The such predicted solutions can be used for comparisons, calculate the accuracy of the prediction.

III. RELEATED SURVEY

The analysis of the pricing policy is adopted by Ryanair, which is the main low-cost carrier in Europe. Based on a year's fare data for all of Ryanair's European flights, using a family of hyperbolic price functions, the optimal pricing

IJIREEICE

ISSN (Online) 2321-2004
ISSN (Print) 2321-5526

International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering

Vol. 7, Issue 2, February 2019

curve for each distinct route is estimated. This analysis shows a positive correlation between average fare for each distinct route and its length, the frequency of flights operating on that route, and the percentage of fully booked flights. As the share of seats offered by the carrier at the departure and destination airports increases, fares tends to decrease. The correlation of dynamic pricing to the distinct route length and the frequency of flights is negative. Conversely, as competition increases discounts on advance fares rise.

Positive correlation is found between fares and distinct route length, distinct route frequency and percentage of fully booked flights. Length and distinct route frequency are also significant variables with negative correlation to dynamic pricing intensity. Ryanair grants fewer discounts on long distance and high distinct frequency routes, despite of advance purchases. We find a negative correlation between the Ryanair's importance in departure, arrival and offered fares. The offer of discounted fares appears as an incentive to use secondary airports.

For instance, an improved measure of the competitive pressure can be made through the analysis of fares applied by Ryanair's competitors [1].
Optimal timing for airline ticket purchasing from the customers perspective is challenging principally because customers have insufficient information for questioning about future price movements. This paper presents a model for computing expected future prices and reasoning about the risk of price changes. The model proposed is used to predict the future minimum expected price of all available flights on specific routes and dates based on historical price quotes. Also, this model to predict prices of flights with specific desirable properties such as flights from a non-stop only flights, specific airline, or multi-segment flights. By comparing models with different target properties, customers can determine the cost of preferences.

The primary data for analysis was collected using daily price quotes from a major travel search web site over the period Feb 22, 2011 to Jun 23, 2011. A Query was written for each unique route and departure date pair in our study, so the results should be representative of what an individual user could observe in the market. The query returned approximately 1,200 unique round-trip journeys from all airlines most queries returned results from 10 or more airlines. All journey was stored in a database, and feature values were computed as aggregates from the set of returned journeys on each day.

The algorithm used in this paper is Partial Least Square (PLS). It is related to Principle components regression (PCR) and Multivariate calibration (MLR). PCR captures maximum variance in X. MLR achieves maximum correlation between X and Y. PLS tries to do both by maximizing covariance between X and Y. PLS regression method considers the latent structures in both datasets and makes a regression model from X to Y so that in future we only require X and can predict Y from X.

This algorithm has multiple advantages. Firstly, PLS regression model can handle very high-dimensional inputs because it implicitly performs dimensionality reduction from the number of inputs to the number of PLS factors. Secondly, the model complexity can be adjusted by changing number of PLS factors to be used in computing the regression results. This value is adjusted in our experiments to determine optimal model complexity in each of the prediction class.

This investigation shows that, given sufficient publicly-observable information, it is possible to predict airline ticket prices to reduce costs for customers. It is observed that there is a significant market for this these kinds of models in the hands of customers. In addition to the results of this paper, there are additional cost reductions that can be found to obtain results closer to the optimal values [2].

Predicting airfare price dynamically based on their pricing decisions on demand estimation models. The reason for such a complicated system is that each flight only has limited number of seats. So, airlines must regulate demand. In the case where demand is expected to exceed capacity, the airline may increase prices, to decrease the rate at which seats fill. On the other hand, a seat that goes unsold represents a loss of revenue and selling that seat for any price is preferred. The purpose of this project was to study how airline ticket prices change over time. Then extract the factors that influence these fluctuations and describe how they're correlated. Using that information, a system is built to help customers in purchasing decision. There are multiple features that can be used to train a predictor, and for some it makes sense to collect them over a period.

Multiple websites expose their routes, time and airlines along with prices which can be scrapped. Some research groups have made their datasets available to other researchers which can be used too. Data collected from Bing Travel, which conveniently included historical information on lowest daily fares.

IJIREEICE

ISSN (Online) 2321-2004
ISSN (Print) 2321-5526

International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering

Vol. 7, Issue 2, February 2019

Consider the formula C = B + F
Where C = cost of ticket
B = Base Price
F = Fluctuation

Now we assume fluctuation is always relative to the base price, which makes the formula C = BF
Now let's assume feature that affect B is disjoint from features that affect F
As we focus on F, we need to substitute value for B approximately.
Approx. price of B can be

- price of the same flight a year ago
- the historical average for that route
- average price over the whole period from the moment ticketing opens up to 2 months before the date of departure.

Due to lack of these required data during the time when experiment was performed, this approach was discarded. Instead we chose to record previous days' prices in relation to the price at each specific day.

Fluctuations depends on following features

- No. of unsold seats
- Prices of other airlines flying the same route
- Date or time of booking
- Recent fluctuation of the price.

Weka Machine Learning Suite was used to experiment with different algorithms. The algorithms are

Ripple Down Rule Learner which was 74.5% accurate.
Logistic Regression which was 69.9% accurate
Linear SVM which was 69.4% accurate.

The main shortcoming of this paper is shortage of data. Which is available in our proposed project. Many features that can vary the base price of a ticket was not considered. It also gives us idea on the problems that will be encountered during manual data collection and previous works done in the field of airfare price prediction [3].

This paper finds the model that fits the behavior of the data well many days before departure. Different passenger on the same flight in the same flight class pay very different prices for their tickets while getting the exact same service. This research proposes statistical regression model for airline ticket prices and compare the goodness of fit. With this prediction model passengers can make a more informed decision whether to buy the ticket or wait a little longer.
We use a data set containing 126,412 observation of ticket prices of 2,271 different flights from San Fransisco Airport to John F. Kennedy Airport, these observations have been made daily by Infare. We find a model that fits the behaviour of the data well many days before departure. Therefore, this approach could help future air travellers to decide whether to buy a ticket or not.

Based on the data gathered a linear model is fit, to fit such a model a simple case such as linear line through the points fitted by ordinary least squares regression model.

FORMULA: $y_i = \alpha + \beta x_i + i$

The mathematical description of this linear line is given in equation where i indicates observation i. xi is the number of days left till departure. yi is the predicted price in USD for this observation. α, β, are respectively the estimated intercept, slope parameter and residuals. Assumed in this model is that each xi is independent. The slope parameter β indicates how the price behaves as the days left decreases. This simple model can therefore give us information about the price of this flight based on the number of days left before departure. This model tells us about how the mean of the price behaves as number of days till departure decreases.

This research takes the first step into applying linear quantile mixed model regression on airline tickets, a possible model is shown for predicting prices of airline tickets based on the number of days left till departure and if the flight leaves in the weekend or on a weekday. Results show that this model follows cheap tickets prices in many days before departure fairly well but tends not to be very effective several days before departure as it just not quite captures the

IJIREEICE

ISSN (Online) 2321-2004
ISSN (Print) 2321-5526

International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering

Vol. 7, Issue 2, February 2019

behaviour. Several steps can be taken to improve performance, therefore further research is needed. Such as for example the inclusion of more predictors for the price, such as the fuel prices, the distance between airports, holidays and restrictions on a specific flight tickets [4].

This paper shows the pilot study of airline ticket prices where data was recorded over 12,000 price observations over 41-days. When trained on this data, Hamlet the multi-strategy data mining algorithm generates a predictive model that saves 341 simulated passengers $198,074 by suggesting them when to buy and when to postpone ticket purchases. Remarkably, a clairvoyant algorithm with complete knowledge of future airfare prices will save at most $320,572 in simulation, thus Hamlet's savings is 61.8% optimal. The airfare data was directly collected from a major travel website. In order to extract large amount of data required for machine learning algorithms, flight data collection agent was built that runs at a scheduled interval, extracts the pricing data, and stores the result in a database.

In this paper generated training data is explained, and the various data mining methods were investigated: Ripper, Q-learning, and time series. Hamlet data mining algorithm combines the results of these methods using a variant of stacked generalization. The data consists of price observations recorded every 3 hours over 41 days. Goal here is to learn whether to buy a airline ticket or wait at a particular time point, for a particular flight, given the price history that has been recorded.

The models used are the following:
- Optimal: This model represents the maximal possible savings, which are computed by a "clairvoyant" algorithm with perfect information about future prices, and which obtained the best possible purchase price for each passenger.
- By hand: This model was hand-crafted by one of the authors after consulting with travel agents and thoroughly analysing our training data.
- Time series: This model was generated by the moving average method described earlier.
- Ripper: This model was generated by Ripper.
- Q-learning: This model was generated by our Qlearning method.
- Hamlet: This model was generated by our stacking generalizer which combined the results of Ripper, Qlearning, and

Time series.
Airfare data is gathered from the web and showed that it is feasible to predict airfare price changes for flights based on historical airfare data. Despite the complex algorithms used by the airlines, and absence of information on key variables such as the number of seats available on a flight, the data mining algorithms performed well. Most notably, Hamlet data mining algorithm achieved 61.8% accuracy in timing ticket purchases.
There are several promising directions for future work on price mining. Airline pricing with data collected over a longer period and over more routes must be included. multi-leg flights should be included in new data set. The pricing behaviour of multi-leg flights is different than that of non-stop flights because each leg in the flight can cause a change in price, and because pricing through airline hubs appears to behave differently as well [5].
Random Forests was introduced by Leo Breiman. It can be used for either a categorical response variable, referred to as "classification", or a continuous response, referred to as "regression". Similarly, predictor variables can be either categorical or continuous.

From a computational standpoint, Random Forests are appealing because they
- naturally handle both regression and classification.
- are relatively fast to train and to predict.
- depend only on one or two tuning parameters.
- have a built-in estimate of generalization error.
- can be used directly for high-dimensional problems.
- can easily be implemented in parallel.

Statistically, Random Forests are appealing because of the additional features they provide, such as,
- measures of variable importance.
- differential class weighting.
- missing value imputation.
- visualization.
- outlier detection.
- unsupervised learning.

IJIREEICE

ISSN (Online) 2321-2004
ISSN (Print) 2321-5526

International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering

Vol. 7, Issue 2, February 2019

Random Forests are a multipurpose tool, applicable to both regression and classification problems, including multiclass classification. They give an internal estimate of generalization error, so cross validation is unnecessary. They can be tuned, but often work quite well with default tuning parameters. Random Forests have been successfully used for a wide variety of applications and enjoy considerable popularity in several disciplines [6].

IV. CONCLUSION

We hereby after surveying the papers published before conclude that the majority of the methods used in papers above made use of traditional prediction models from the computational intelligence research field known as Machine Learning. It was difficult for the customer to purchase air ticket due to the high complexity of the pricing models applied by the airlines because the prices change dynamically. Many features that can vary the base price of a ticket was not considered. There were problems that was encountered during manual data collection and previous works done in the field of airfare price prediction. From the experiments we concluded which features influence the airfare prediction at most. Future this project could be extended to predict the airfare prices with higher performance considering few other features.

REFERENCES

[1] 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.
[2] A regression model for predicting optimal purchase timing for airline tickets William Groves and Maria Gini October 18, 2011
[3] Predicting Airfare Prices Manolis Papadakis, 2014
[4] Linear Quantile Mixed Regression Model for Prediction of Airline Ticket Prices, 2014
[5] To Buy or Not to Buy: Mining Airfare Data to Minimize, 2003
[6] Random Forests, January 2011