Prediction of Airline Ticket Price

Ruixuan Ren, Yunzhe Yang, Shenli Yuan

Introduction

Airline industry is one of the most sophisticated in its use of dynamic pricing strategies to maximize revenue, based on proprietary algorithms and hidden variables. Therefore, it is challenging for consumers to predict the price change in the future [i]. With the information of the airfare available online, buyers are trying to track the prices of the flight over a certain period of time, and anticipate the price change in the future. However, it turns out to be rather difficult to predict the price of the flight precisely only by observation.

Our project is aimed at building up models to predict the airline ticket price. The input of our models are the factors that may influence the price, such as the weekday of departure and the number of stops in the itinerary. We applied linear regression, Naïve Bayes, Softmax regression, and Support Vector Machine (SVM) to predict the corresponding price.

Related Work

Previous works have been done regarding airfare prediction using machine learning. Etzioni et al.[i] have performed a pilot study on 12000 price observations over a 41-day period. Their multi-strategy data mining algorithm – Hamlet generated a predictive model that could potentially save substantial amount of money which consumers pay on airline tickets. Groves et al.[iv] proposed a model to predict expected minimum price of all flights on a particular route. The model was also used to predict price with different target properties such as prediction from a specific flight, non-stop only flight and etc. Rama-Murthy[ii] built a model to predict the airfare price with specific focus on how different factors influence the price of airline tickets. Papadakis[iii] studied how prices of airline tickets change overtime by extracting several factors that potentially affect the price fluctuation and finding out their correlation.

Dataset and Features

The dataset used in our project is provided by Professor Gini [iv] from University of Minnesota. It was originally collected using daily price quotes from a major travel search web site over the period February 22, 2011 to June 23, 2011. The data were used to build a regression model for computing expected future prices and reasoning about the risk of price changes. The data source contains information of seven different routes operated by several flight companies. The features selected to use in our model include: the departure week begin, weekday of the departure, price quote date, weekday of the price quote, number of days between fetch days and the departure, and the number of stops in the itinerary.

To shed light on how the dataset looks like, **Error! Reference source not found.** shows a small portion of this dataset. It presents the trend of mean lowest price offered by all the airlines versus the number of days between fetchdays (purchase date) and the departure. To simplify the figure, depart week begin (number of days between 1 Jan 2011 and the departure week's Monday) is set to be 128, 135, 142, and 352, respectively. The weekday of departure date is Monday.

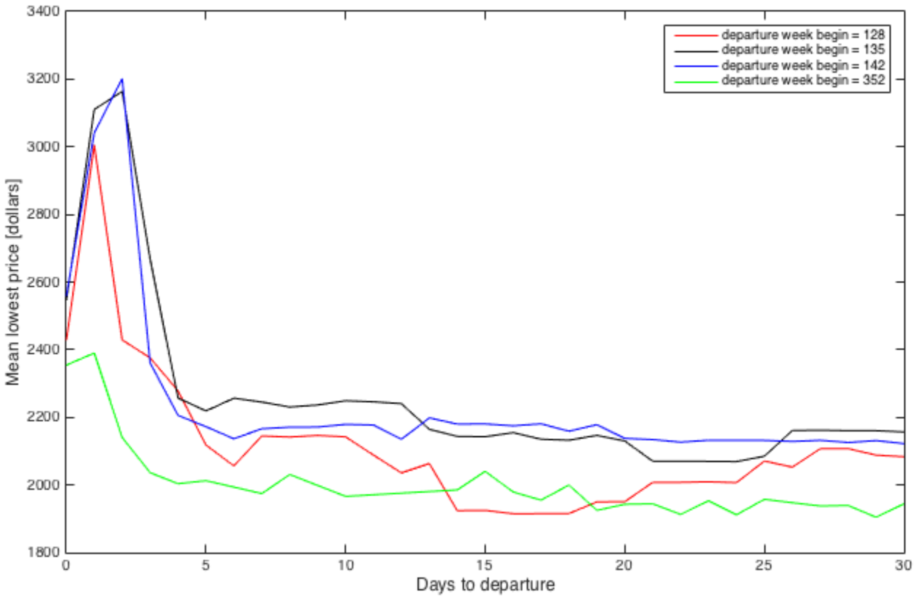

Figure 1 Days to departure vs. Mean lowest price

Methods

Since the price of the flight varies due to the difference in distance, popularity of airport, and other factors, it is hard to build up a model which performs well for all the flights. We decided to train different models for each airline route and the model trained is only applicable to its corresponding route. For the continuous model, the target variable is simply the one-day-average price. For Naïve Bayes and Softmax regression, we classified the prices into five bins using three different classification methods, and the target variable would be 1 to 5, representing each bin. For SVM, the prices are classified using equal interval.

1. Linear Regression Model

Linear regression was performed as the first attempt due to its simplicity. The four features selected in the model include the weekday of the departure date, denoted as x_1, where $x_1 = k$ $(k = 1, 2, \dots, 7)$, the weekday of the quote fetch date denoted as x_2, the number of days between quote fetch date and the departure date denoted as x_3, and the number of stops during the itinerary denoted as x_4. Normal equation was used in linear regression models. Both weighted and unweighted linear regression were performed for comparison. The band width values used in weighted linear regression are 0.8, 2 and 10, respectively. The normalized error is defined as

$$\varepsilon = \sqrt{\frac{1}{n}\sum\left(\frac{h_\theta(x^{(i)})}{y^{(i)}} - 1\right)^2}$$

2. Naïve Bayes Model

To convert the problem to a classification problem, we applied three different discretization methods: equal probability discretization, equal interval discretization, and K-means cluster. We separated the relative price into five bins and selected the same features as the ones used in linear regression. Multinomial event model of Naïve Bayes with Laplace smoothing is applied to model the trend of the price. Several tests indicate that equal interval discretization method always gives the highest accuracy. Therefore, we decided to apply this method when developing Naïve Bayes model.

To parameterize the distribution of relative price, y, over 5 possible outcomes, we use 5 parameters specifying the probability of each of the outcomes: $\phi_1, \phi_2, \ldots, \phi_5$, where $\sum_{i=1}^{5} \phi_i = 1$. Maximizing the log joint likelihood of the training set with respect to $\phi_y, \phi_{x|y=j}$ gives the maximum likelihood estimates:

$$\phi_{y=j} = \frac{\sum_{i=0}^{m} 1\{y^{(i)} = j\} + 1}{m + 5}$$

$$\phi_{x_k=t|y=j} = \frac{\sum_{i=0}^{m} 1\{x_k^{(i)} = t \wedge y^{(i)} = j\} + 1}{\sum_{i=0}^{m} 1\{y^{(i)} = j\} + V}$$

3. Softmax Regression

Using the same discretization method of relative price described in Naïve Bayes model, we can also develop softmax regression model to predict the price. We have

$$p\left(y^{(i)} = j \middle| x^{(i)}; \theta\right) = \frac{e^{\theta_j^T x^{(i)}}}{\sum_{l=1}^{k} e^{\theta_l^T x^{(i)}}}$$

The parameters of softmax regression are updated based on the following equation:

$$\theta_j := \theta_j + \alpha \frac{1}{m} \sum_{i=1}^{m} [x^{(i)}(1\{y^{(i)} = j\} - p(y^{(i)} = j|x^{(i)}; \theta))]$$

4. SVM

We divided prices into several bins according to their relative values compared to the overall average air ticket price. An example set of bins would be 60% to 80%, 80% to 100%, 100% to 120% and 120% to 140% (of average price) etc. Using L2-regularization and L2 loss function, we are able to achieve an accuracy of 60.54%, which is not very ideal. However, when using only two bins (higher and lower than average), we are able to achieve an accuracy of 80.6%.

SVM regression has also been used as a continuous model, which does not generate satisfying results, thus discarded.

Result and Discussion

We selected the price of the round trip between New York Airport and Charles de Gaulle Airport to test the performance of the model developed, applying leave out one cross validation. There are 9390 data points included in the testing dataset.

The training error produced by each model are listed in the following table:

Table 1 Training Error

Model		Error
Linear Regression	Unweighted	0.2304
	$\tau = 0.8$	0.2294
	$\tau = 2$	0.2297
	$\tau = 10$	0.2304
Naïve Bayes		0.2694
Softmax Regression		0.2316
SVM(two bins)		0.1939

For weighted linear regression, when the bandwidth value is large ($\tau = 10$), the hypothesis is closer to the one of unweighted linear regression; the errors of these two cases are almost the same. In addition, not surprisingly, the smaller the bandwidth value is, the better the prediction fits the training data, and the smaller the training error would be.

In order to reduce the training error, we utilized learning curve to investigate the price prediction problem. Since Naïve Bayes is computational efficient and gives reasonable results when classifying 5 bins, and SVM gives the lowest error when classifying only 2 bins, we picked them to investigate further.

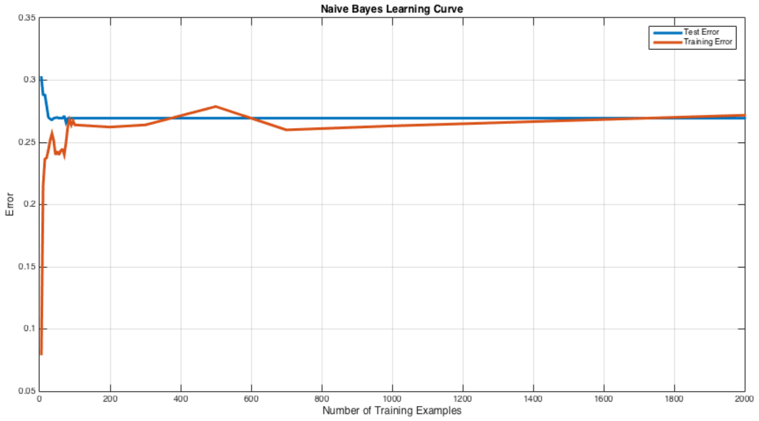

Figure 2 Naïve Bayes Learning Curve

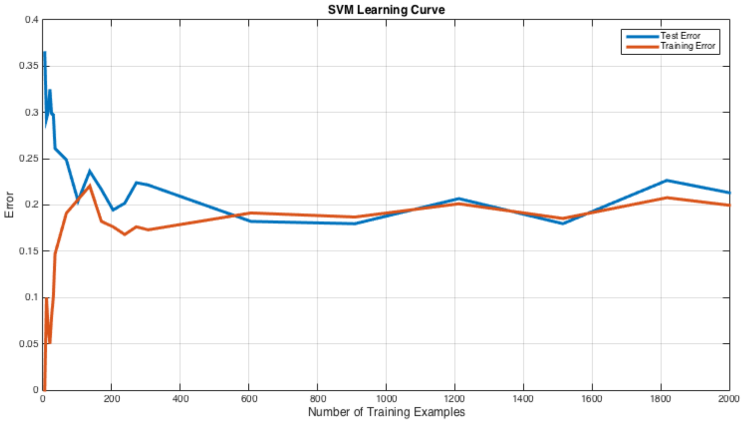

Figure 3 SVM Learning Curve

From the plots above, it is apparent that both models have high biases, indicating adding more features will help reduce the error. Due to the limited accessible information from price-quoting websites, we are only able to add the count of itineraries from all airlines to our features, and we found that the training error of Naïve Bayes and Softmax regression reduced to 24.88% and 20.22%, respectively. Similarly, the error of SVM model is also reduced by approximately 1%. Ideally, if we could have access to information such as number of seats on airplane or number of vacant seats etc., we would be able to predict the price much better. Still, we could have gathered information such as the departure time of the day to improve the accuracy of our model. But these information is not included in the dataset from Professor Gini and given the limited time of this project, we are unable to collect enough amount of data.

In addition to adding new features, we also study the performance of our model in each individual bin. The results showed that Naïve Bayes only misclassifies data in the first bin (lowest price interval), with a training error of 36.40%. This means Naive Bayes is quite reliable in all other price intervals except the lowest one.

Conclusion

This study shows that it is feasible to predict the airline ticket price based on historical data. One possible way to increase the accuracy can be combining different models after carefully studying their own performance on each individual bin. Additionally, as the learning curve indicates, adding more features will increase the accuracy of our models. However, limited by the current data source that we have, we are unable to extract more information of a particular flight. In the future, more features, such as the available seat, the departure time of a day, and whether the departure day is a holiday or not, can be added to the model to improve the performance of the predicting model.

Reference

[i] Etzioni, O., Tuchinda, R., Knoblock, C. A., & Yates, A. (2003, August). To buy or not to buy: mining airfare data to minimize ticket purchase price. In*Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 119-128). ACM.

[ii] Rama-Murthy, K. (2006). Modeling of United States Airline Fares--Using the Official Airline Guide (OAG) and Airline Origin and Destination Survey (DB1B).

[iii] Papadakis, M. (2014). Predicting Airfare Prices.

[iv] Groves, W., & Gini, M. (2011). *A regression model for predicting optimal purchase timing for airline tickets*. Technical Report 11-025, University of Minnesota, Minneapolis, MN.