COMP-SCI 5542 (SP17) - Big Data Analytics and Applications

Project Proposal - Due 01/30/17 by 11:59 PM

Yunlong Liu **(22)**
Chen Wang **(44)**
Dayu Wang **(45)**

1. Project Title, Team Number, and Members (Table 1)

Table 1. Basic project team information for the class COMP-SCI 5542 (SP17).

Project Title	**Buy Now or Wait? Data-Mining-Based Minimization of Airfare**
Team Number	**9**
Members	***Yunlong Liu* (22)** - yln69@mail.umkc.edu ***Chen Wang* (44)** - cwrp3@mail.umkc.edu ***Dayu Wang* (45)** - dayuwang@mail.umkc.edu

2. Project Goal and Objectives

2.1. Motivation

In today's America, most of people's traveling relies on aircraft. When we are planning a future excursion, booking flights is one of the most crucial parts for our entire trip. Therefore, how to minimize the airfare is a very valuable research topic for data scientists that can definitely facilitate people's everyday life. Someone might claim that the earlier you purchase the air ticket, the lower price of the ticket you pay. Such logic does sound plausible, but in reality, it is not quite true. For example, Figure 2 demonstrates the average air ticket booking prices from Kansas City (MO) to Houston (TX) for the actual traveling data of Jan 1st, 2017[1], which obviously proved that the logic of "the earlier, the cheaper" is not correct at all. The pattern of how air ticket prices are altering is hidden deeply inside the huge amount of historical data of air tickets. Therefore, we would like to develop a more complex system that can "study" the previous airfare data, and provides tips to the user of whether it is better to purchase the ticket now or to wait until the price decreases sooner or later.

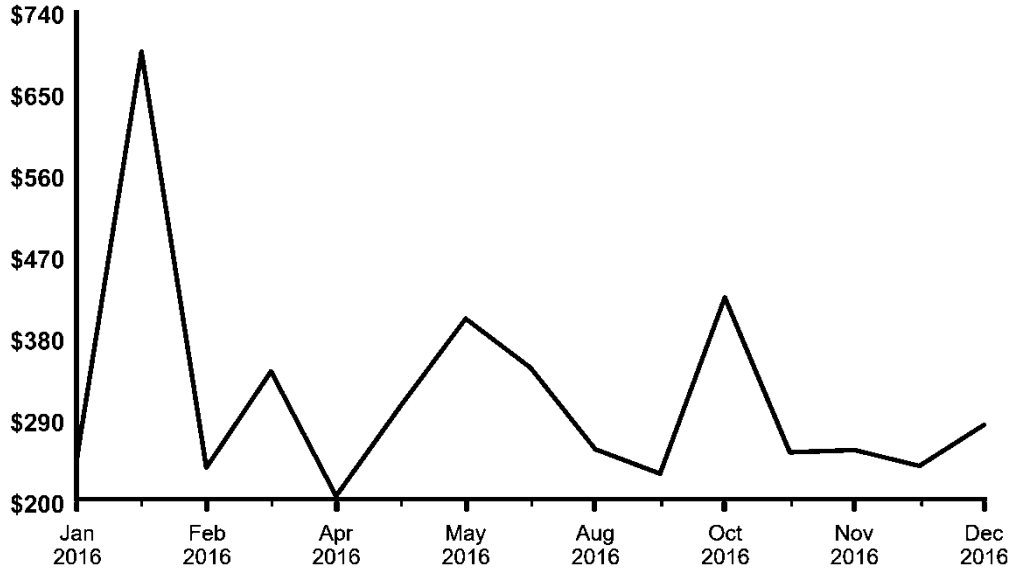

Figure 2. Historical booking prices in the entire year of 2016 for the flights from Kansas City (MO) to Houston (TX) on Jan 1st, 2017[1].

2.2. Significance/Uniqueness

From Figure 2, we have the a few observations to present. First of all, if your target date of travel is within 30 days, the price of air ticket will probably keep going up, unless special discounts/events occur during that time period. This can be logically explained by the fact that for the last 30 days, the number of remaining seats in the flight becomes the dominant factor to determine the cost of the ticket. This observation can help our system being developed delve out the **date** as a key factor to generate the tip of "Buy Now or Wait".

Second, there are several peaks of the curve of the booking prices. And what is more, though the prices for the peaks were dramatically different, those peaks seemed to appear **periodically**. Therefore, such a periodic characteristic can provide a nice opportunity for the system to construct a self-adjusting machine learning model which can accurately predict the peak booking prices in future, telling the user to avoid booking air tickets on those days.

Third, there exist several valleys (low prices) in the curve which also seemed like a **periodic** property. This can be modeled by our developed system to know the next cheap-booking date. **The uniqueness of our system is that not only "Buy Now or Wait" decision is generated by our machine learning models, but also "When to Check Again" and "How Long You Can Wait" are provided to the user, since no one has the time to check air ticket prices every day since you tell him/her to wait.**

2.3. Objectives

In this project, we would like to implement a self-adjusting machine learning model to fit the changing curves of air ticket prices from scratch. Based on the class schedule in this semester, we have 4 iterations in our project development. Depending on the 4 iterations of the Agile project development, our objectives in each iteration are listed in Table 3.

Table 3. Team objectives for the 4 iterations of the Agile project development.

Iteration	Objectives
1	• **Literature Searching** - Get to know what others did in this area. • **Data Collection** - Collect a satisfactory amount of raw data for future use. • **Real-Time Data** - Try to make our data be real-time. Set up queries for the data.
2	• **Theoretical Models** - Separate parameters to ease the machine learning process. • **Implementation** - Implement the self-adjusting model for the air ticket prices. • **Test** - Test the performance of our implemented model.
3	• **Refinement** - Elaborate the logic of our model and prove the reasonability. • **Application** - Build an web application for our developed system.
4	• **Application Test** - Test the application and fix the bugs. • **Publishing** - Publish our application to Amazon EWS/IBM Bluemix. • **Documentation** - Prepare final documentation and submission package.

2.4. System Features

We present the system features from both the developer's view (Table 4) and the user's view (Table 5), since some of the features are hidden from the presentation of the final application.

Table 4. System features based on the developer's view.

Number	Feature Description
D_1	**Real-Time** Database Management
D_2	**Machine Learning** Models
D_3	**Deep Learning** of the Data
D_4	**Multi-Device Compatible** Web Application

Table 5. System features based on the user's view.

Number	Feature Description
U_1	Simple **Air Ticket Searching Engine**
U_2	**Price Tip** - "Buy Now or Wait"
U_3	Waiting Suggestion - "**When to Check Again**"

3. Related Work

We would like to present 2 projects here, a research paper (algorithm development) and an actual application (Bing Travel).

3.1. To Buy or Not to Buy: Mining Airline Fare Data to Minimize Ticket Purchase Price[2]

Although huge amount of data of prices of air tickets is available on the internet, the philosophy of how airline companies vary their ticket prices is still mysterious to the customer. This paper developed an algorithm, which came from the data mining of existing data, to reveal the underlying pattern of air ticket prices. In this paper, many data mining methods are applied, which gave satisfactory results to save customer's money.

Full Paper: https://docs.google.com/viewer?url=https://github.com/dwk894/COMP-SCI_5542_SP17_Team_9_Project/raw/master/References/Etzioni_-_2003_-_PDF.pdf

3.2. Bing Travel

This is a very interesting case, because the airfare predictor feature in Bing Travel has been removed now. Nevertheless, the idea of airfare predictor is still valuable to us, since Bing provided another key factor, named **confidence** (Figure 6), to tell the user how risky to follow their tips of "Buy Now or Wait". The confidence value kept changing when more and more data was passed by its engine, which represented a very classic model of machine learning.

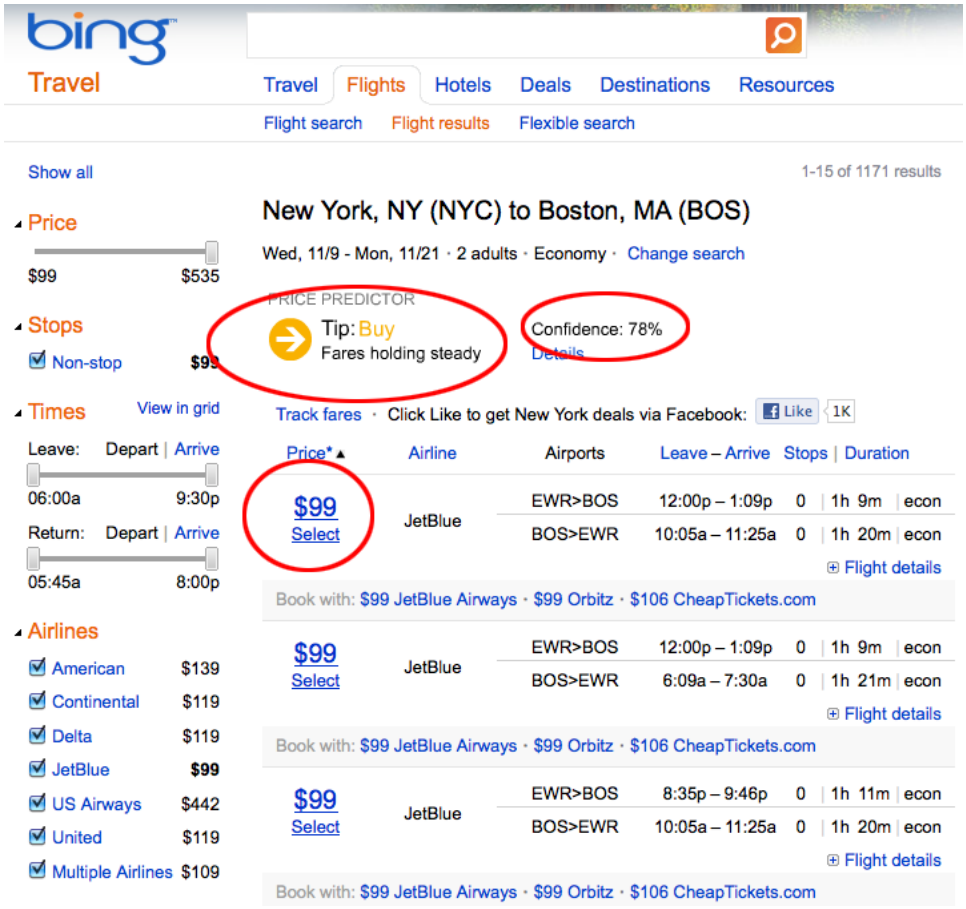

Figure 6. The **confidence** value presented in the airfare predictor of Bing Travel.

4. Backup Project (Table 7)

Table 7. **Backup** project proposed by Team 9 for the class COMP-SCI 5542 (SP17).

Project Title	**Friendly or Unfriendly? Study of MEAN COMMENTS in Social Media**
Project Description	People who use social media (Facebook, Twitter, etc.) sometimes receive **mean comments**, even from his/her best friends or family members, after he/she just posted what he/she did or the achievement just accomplished. Traditional sentiment analysis can only give "positive/negative" results for those text-based comments. However, some negative comments were friendly, since they wanted to share their ideas and had a discussion with you, or just a joke; some comments were unfriendly, in which the only purpose of the comment was to hurt you. For the latter case, those "friends" should no longer appear in your social media account. Admittedly, people often have the feeling of whether a mean comment is friendly or not. But, can a machine still read that? This project will challenge us further than the regular sentiment analysis. We would like to implement an even deeper machine learning model, in order to read people's hearts amongst those mean comments in social media.

5. Bibliography

[1] fare | detective
http://www.faredetective.com

[2] Etzioni, Oren, *et al*. "To Buy or Not to Buy: Mining Airfare Data to Minimize Ticket Purchase Price." Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2003.
https://docs.google.com/viewer?url=https://github.com/dwk894/COMP-SCI_5542_SP17_Team_9_Project/raw/master/References/Etzioni_-_2003_-_PDF.pdf