Fordham University

Gabelli School of Business

WEB CRAWLING AND AIRLINE TICKET FARE ANALYSIS IN EXPEDIA

Yuqi Chen
Nada Rachid
Betsy Mendieta Brito
Haoxiang Jia
Claudia Villanueva Robles
BYGB7978004 -Jenny Gong
December 2022

Web Crawling and Airline Ticket Fare Analysis in Expedia

Expedia has been a pioneering online travel service for many years and is one of the most popular places for travelers to search, contrast, and book flights. Even though the COVID-19 epidemic caused a disruption in travel, it wasn't until recently that things returned to normal on a worldwide scale prompting a surge in interest in low-cost flights when people realized how much post-COVID ticket costs had risen.

Expedia is now more necessary than ever for the sole purpose of seeking and comparing flights, but the high and continuously shifting ticket prices have slowed down the process of turning prospective clients into actual customers. The goal of this project is to test an appropriate model that will determine the optimal day and time to get the cheapest fare for the desired flight. Therefore, speeding up the process of searching, comparing, and buying airline tickets in Expedia and, ultimately, shorten and simplify the customer experience.

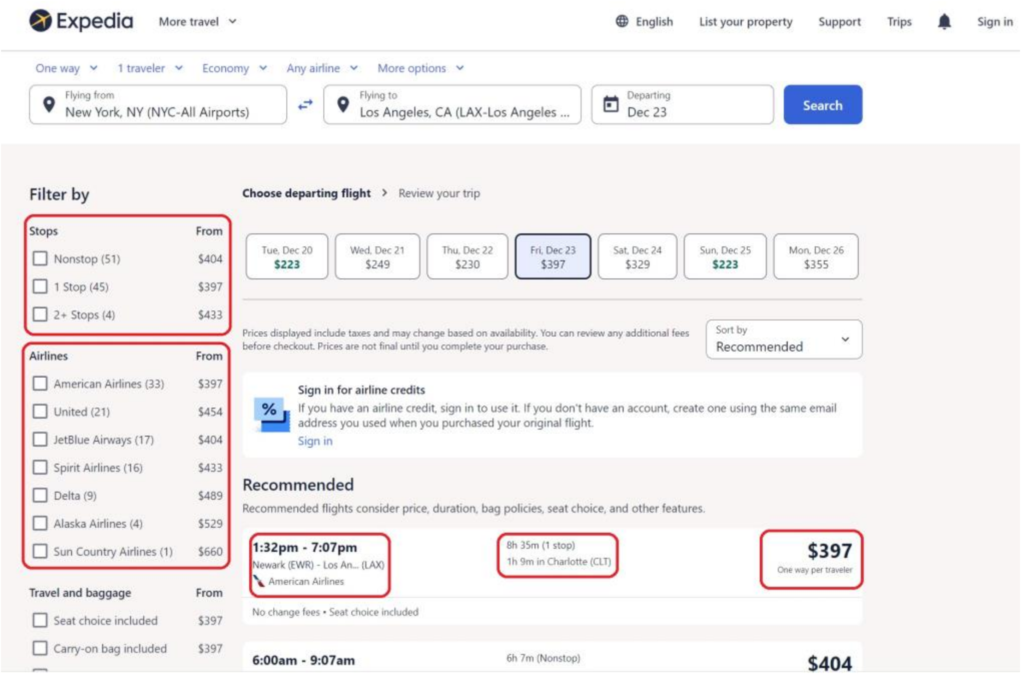
Business Goal Analysis

The purpose of this research is to predict the cheapest flight available for a certain route on a certain day. As a starting point, we chose a sample flight which is one way from New York to Los Angeles, leaving on the 23rd of December. We will measure the flight price change within an 8-hour window for one day. To make use of our acquired knowledge in the Web Analytics class, we will crawl the Expedia website utilizing Selenium WebDriver and Tweepy API.

We will use Selenium WebDriver to mimic user clicks on the Expedia website to collect data about a certain flight and create the dataset that can be analyzed by applying LASSO Regression to get insights about the relationship between the parameters and the dependent variable “price”. Furthermore, we will determine which is the most important predictor that determines the cheapest price. Additionally, we will use Tweepy API to crawl data on Twitter and perform sentiment analysis to get a better understanding of customer’s experiences in those particular airlines.

This model will certainly help the customers reduce the burden of continuously searching and comparing flights.

Dataset Description



Flying from: New York, NY (NYC-All Airports)

Flying to: Los Angeles, CA (LAX-Los Angeles ...

Departing: Dec 23

Stops	From
□ Nonstop (51)	$404
□ 1 Stop (45)	$397
□ 2+ Stops (4)	$433

Airlines	From
□ American Airlines (33)	$397
□ United (21)	$454
□ JetBlue Airways (17)	$404
□ Spirit Airlines (16)	$433
□ Delta (9)	$489
□ Alaska Airlines (4)	$529
□ Sun Country Airlines (1)	$660

Travel and baggage	From
□ Seat choice included	$397
□ Carry-on bag included	$397

Tue, Dec 20	Wed, Dec 21	Thu, Dec 22	Fri, Dec 23	Sat, Dec 24	Sun, Dec 25	Mon, Dec 26
$223	$249	$230	$397	$329	$223	$355

Prices displayed include taxes and may change based on availability. You can review any additional fees before checkout. Prices are not final until you complete your purchase.

Sort by: Recommended

Sign in for airline credits

If you have an airline credit, sign in to use it. If you don't have an account, create one using the same email address you used when you purchased your original flight.

Recommended

Recommended flights consider price, duration, bag policies, seat choice, and other features.

1:32pm - 7:07pm
Newark (EWR) - Los An... (LAX)
American Airlines

8h 35m (1 stop)
1h 9m in Charlotte (CLT)

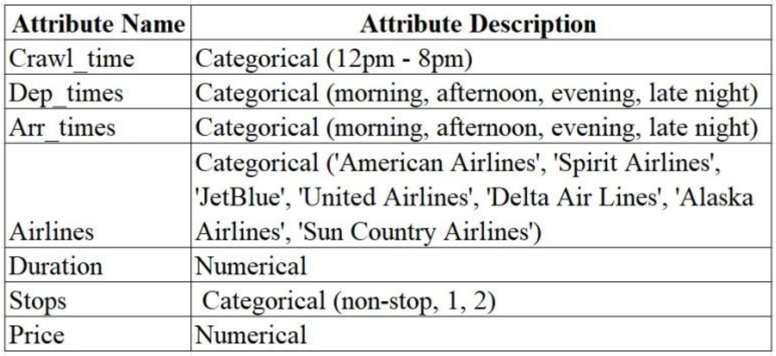
$397
One way per traveler

No change fees • Seat choice included

6:00am - 9:07am
6h 7m (Nonstop)
$404

Figure 1. Dataset Example

As shown in Figure 1, we scrapped from Expedia all the elements highlighted in the table, we have 6 attributes with 793 records in total. 793 records might seem like a small dataset; however, since the prediction insight is the expected lowest price for an airfare to satisfy the dataset portrait-shape requirements we need (number of observations = 10 * number of predictors) 10*6 = 60 observations and we have 793 in total, which is far beyond the minimum requirement. Hence, we have enough observations to conduct this analysis. Below are the attributes along with their description.



Attribute Name	Attribute Description
Crawl_time	Categorical (12pm - 8pm)
Dep_times	Categorical (morning, afternoon, evening, late night)
Arr_times	Categorical (morning, afternoon, evening, late night)
Airlines	Categorical ('American Airlines', 'Spirit Airlines', 'JetBlue', 'United Airlines', 'Delta Air Lines', 'Alaska Airlines', 'Sun Country Airlines')
Duration	Numerical
Stops	Categorical (non-stop, 1, 2)
Price	Numerical

Table 1. Attribute Descriptions

As shown in Table 1, we have 5 attributes that are categorical and one attribute that is numerical, and our target value is price which is a continuous variable. None of the attributes in the dataset contain missing values. The dataset consists of historical observations organized as individual level data, where each column contains one variable and each row is an observation for an individual, as opposed to aggregated data. The dependent variable has observations that show the cases where the events of our interest did happen, all the independent variables are relevant to the airline ticket price by using institutional knowledge and all the predictors are exante.

Figure 2 shows the a sample graph of the word count used in our sentiment analysis. This dataset does not contain any missing values but those contain many punctuation characters that will be removed in our data cleaning pre-processing to conduct our analysis.

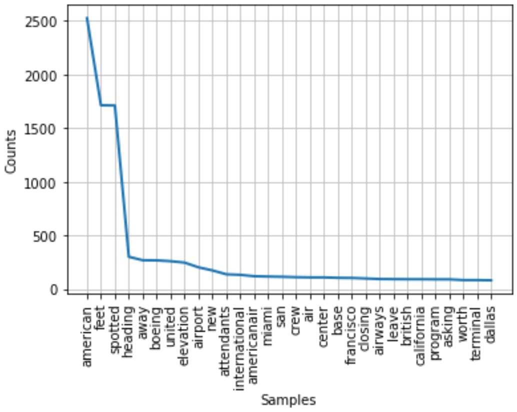


Figure 2.Word Count Graph

System Design

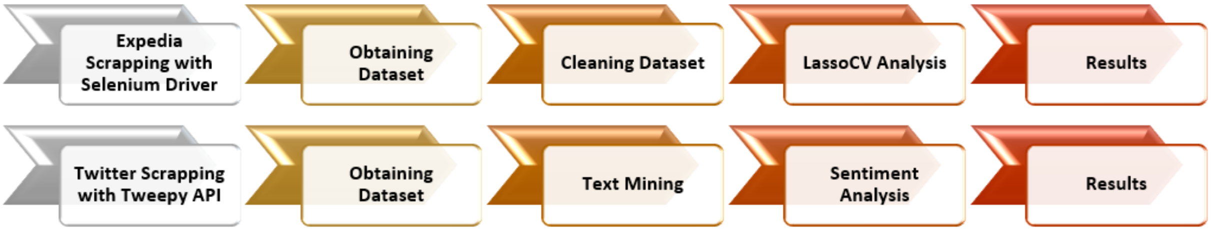


Figure 3. System Design Flowchart

As described in Figure 4, we performed two different analyses in our project. First one was performing Expedia scraping by utilizing Selenium driver to mimic user clicks and obtain our desired dataset, then cleaning the data to perform the LASSO Analysis. In our second analysis, we used Tweepy API to obtain all the reviews from the different airlines of our original dataset, then we implemented Sentiment Analysis after cleaning the dataset to obtain our results.

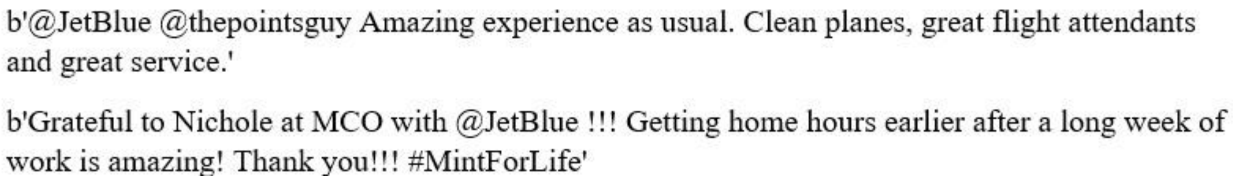
System Implementation

The first tool used was Selenium WebDriver for scraping data from Expedia.com. The first step was to utilize the searching flights feature where we inserted specific details about the flight. We set the following parameters:

- “Leaving from” as “New York (NYC - All Airports)
- “Going to” as “Los Angeles (LAX - Los Angeles Intl.)
- “Departing Date” as “Dec 23”
- One-way
- 1 traveler
- Economy

The second step was to put together a code that automatically runs every hour for eight hours and sends back the dataframe as a result. After running the code during the 8-hour window we combined the 8 data frames collected into 1 big dataset. This dataset contains the flight ticket information of 793 observations with 7 attributes. There are 6 independent variables and 1 dependent variable. Our dependent variable is price while the independent variables are crawl_time, dep_times, arr_times, airlines, durations, and stops. We exported all the crawled data into a csv file.

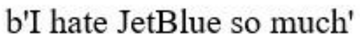
In the second tool used, Tweepy API, we input the airline names scrapped using the first tool and we obtain all tweets where the names of the airlines were mentioned. We later use that output to do the sentiment analysis and get the negative and positive sentiments to be evaluated and reported.



b'@JetBlue @thepointsguy Amazing experience as usual. Clean planes, great flight attendants and great service.'

b'Grateful to Nichole at MCO with @JetBlue !!! Getting home hours earlier after a long week of work is amazing! Thank you!!! #MintForLife'

Figure 4. Example of a positive-attitude tweet



b'I hate JetBlue so much'

Figure 5. Example of a negative-attitude tweet

Data exploration

For exploratory analysis, we draw the side-by-side boxplots display their relevance relationships with the dependent variable price as shown in Figures 6 to 10 below.

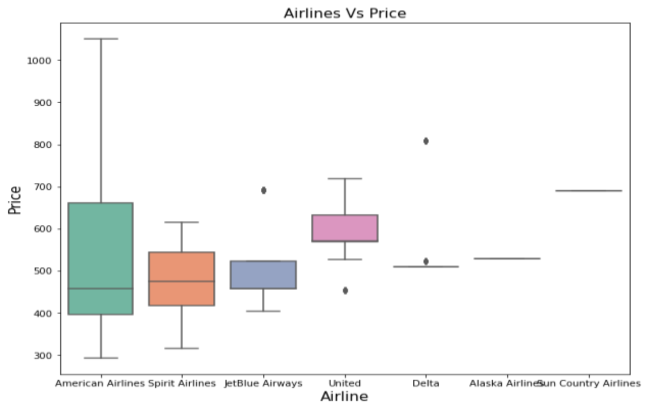


Figure 6.Side-by-side Boxplots Airlines vs. Price

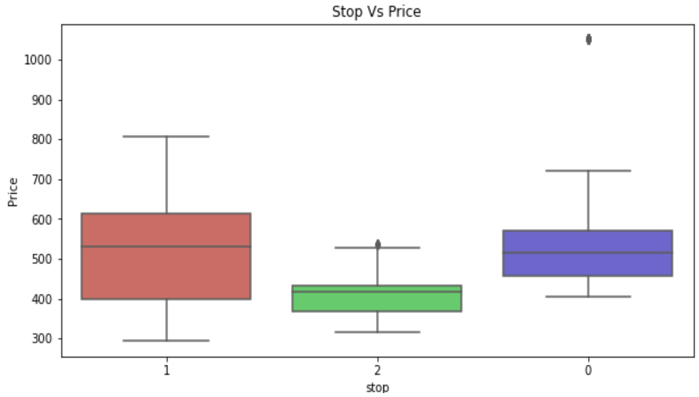


Figure 7. Side-by-side Boxplots Stops vs. Price

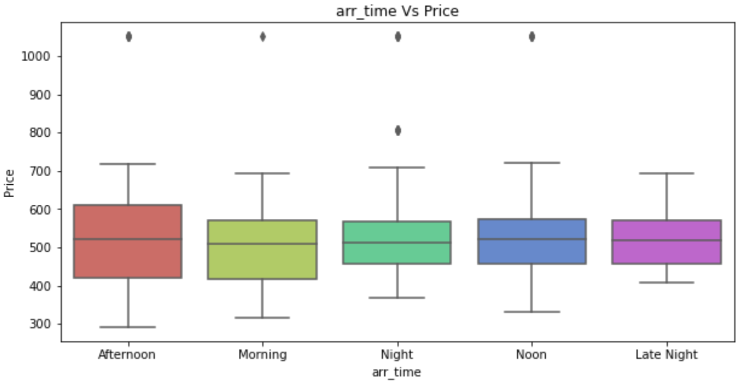


Figure 8. Side-by-side Boxplots Arrival Time vs. Price

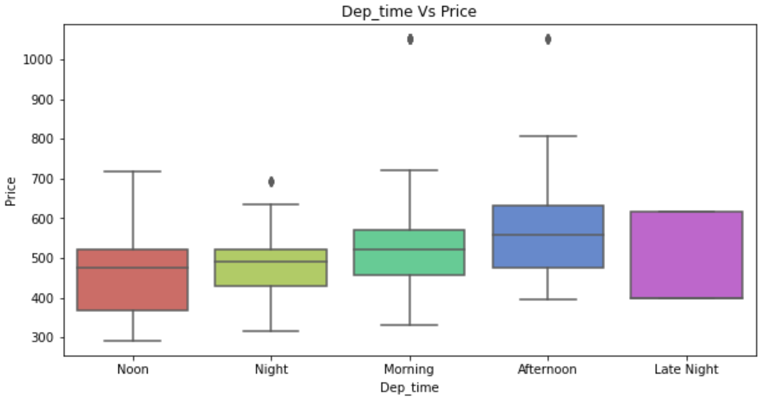


Figure 9. Side-by-side Boxplots Departure Time vs. Price

Data Cleaning

For our dataset we had the following characteristics:

- No missing values
- For time attributes (dep_times, arr_times), we first convert the 12-h format (am/pm) into a 24-h format. Then categorize them as “morning (5am-12am)”, “afternoon (12pm-17pm)”, “evening (17pm-0am)”, “late night (0am-5am)”.
- For durations, we convert the hour, minute format into seconds for consistency.
- Before running Regression, we calculate the Variance Inflation Factor (VIF) for each explanatory variable.

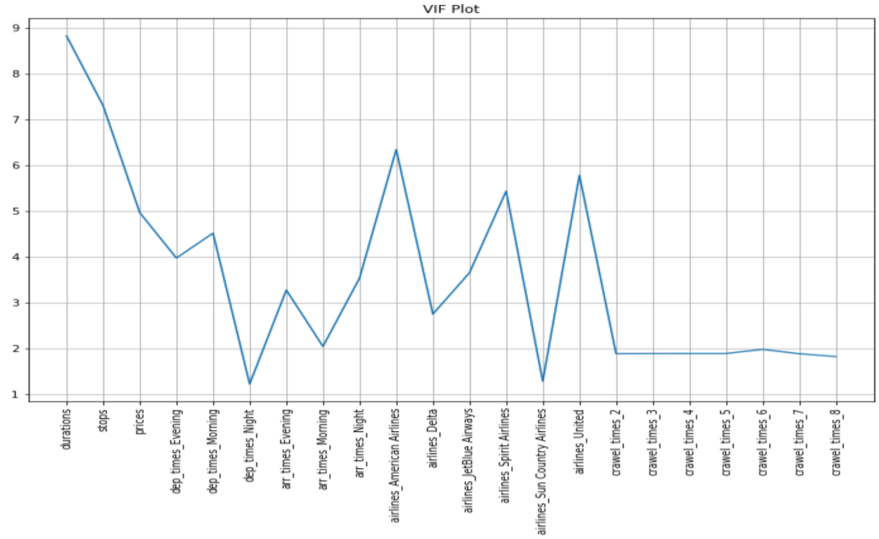


Figure 10. Variance Inflation Factor (VIF) Plot

- Since no VIF exceeds 10, no serious multicollinearity is detected. Therefore, we decide to include all explanatory variables in our regression.

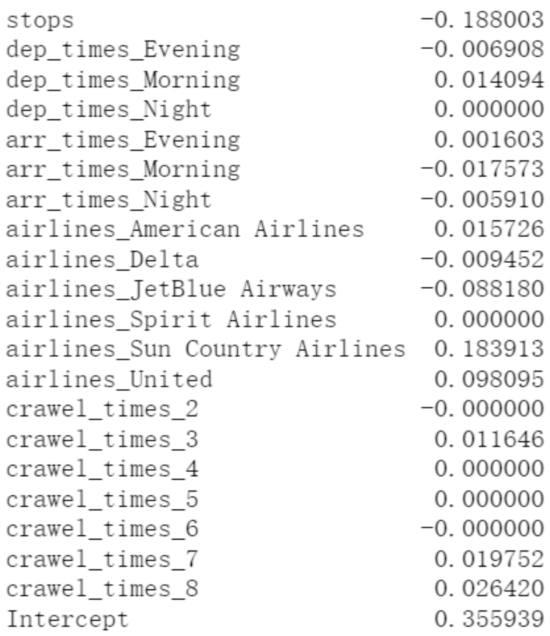
Lasso Regression Analysis

- We used Lasso Regression to find the most important and relevant predictors.
- Predictors are standardized before conducting Lasso, so that the differences between measurement scales don’t affect the penalization of the coefficients.

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

- By increasing the λ (the penalty level alpha), we can tell from looking at the formula above that the slope of the regression will decrease. When λ = 0, the Lasso Regression is the same as Linear Regression. Along with the increase of penalty level, the predicted values approach the mean values of the original data. Therefore, only important predictors are included in the regression since the variance explained by the model decreases.

- In LassoCV, the model performs grid search using k-fold cross validation method to find the optimal penalty level. We use the coefficients returned by the LassoCV model to identify the effect of each predictor on flight ticket price.
- However, we should notice that the coefficients of some predictors return a value of 0. It doesn't mean these predictors are not meaningful, it only suggests they don't add enough importance to our context.
- The following figure displays the coefficient of each predictor returned by LassoCV model:



```
stops                            -0.188003
dep_times_Evening                -0.006908
dep_times_Morning                 0.014094
dep_times_Night                   0.000000
arr_times_Evening                 0.001603
arr_times_Morning                -0.017573
arr_times_Night                  -0.005910
airlines_American Airlines        0.015726
airlines_Delta                   -0.009452
airlines_JetBlue Airways         -0.088180
airlines_Spirit Airlines          0.000000
airlines_Sun Country Airlines     0.183913
airlines_United                   0.098095
crawel_times_2                   -0.000000
crawel_times_3                    0.011646
crawel_times_4                    0.000000
crawel_times_5                    0.000000
crawel_times_6                   -0.000000
crawel_times_7                    0.019752
crawel_times_8                    0.026420
Intercept                         0.355939
```

Figure 11. Predictors with their respective coefficients and Intercept

- In the end, we conduct the necessary analysis to rank the predictors based on their importance. As per Figure 11, the mot important predictor is stops.

Twitter Scraping

- We got tweets by searching keywords: "American Airlines", "Spirit Airlines", "JetBlue", "United Airlines", "Delta Airlines", "Alaska Airlines", "Sun Country Airlines"
- Set maximum tweets searching upper bound to 5000 (even though the more tweets the better for further sentimental analysis, we can't scrape more because Tweepy API has rate limits)
- Output the tweets for each airline to csv files. Table 2 shows a summary of scrapped tweets per airline.

Airline Name	Number of Tweets Collected
American Airlines	5000
JetBlue	4628
United Airlines	4463
Spirit Airlines	1694
Alaska Airlines	1508
Delta Air Lines	1046
Sun Country Airlines	70

Table 2. Number of Scrapped Tweets per Airline

By just comparing the number of the tweets we scraped from Twitter, we can tell how popular each airline is. For example, American Airlines reaches the maximum search to 5000 tweets, making it the most popular airline since people have a lot of discussions about it on Twitter. Sun Country Airlines has only 70 tweets, which means it is not a very well-known airline.

Text mining

- We conducted text mining for each airline’s tweets for further sentimental analysis
- Package used: nltk
- Stop words removed:
    - HTML-related words: "br", "href", "http", "https", "b", "amp", "intl", "et", "ne", "l", "int"
    - High frequency but useless words: "airlines", "airline", "flights", "flight", "spotted", "feet"

Text Scraping

- We used the SentimentIntensityAnalyzer from the nltk.sentiment package
- The portion of neutral-attitude tweets is relatively high thus the rows whose compound sentiment score is 0 were removed to just focuse at the ones with an extreme attitude.
- Finally, we generated the Word Clouds for both positive and negative tweets keywords for each airline.

Evaluation

I. LASSO Regression-Residual Analysis

- Residual = Observed - Predicted.
- The residuals are symmetrically distributed and tend to cluster toward the middle of the plots. Generally, we don’t see any patterns in our residual plots, which suggests that our model doesn’t suffer from autocorrelation.

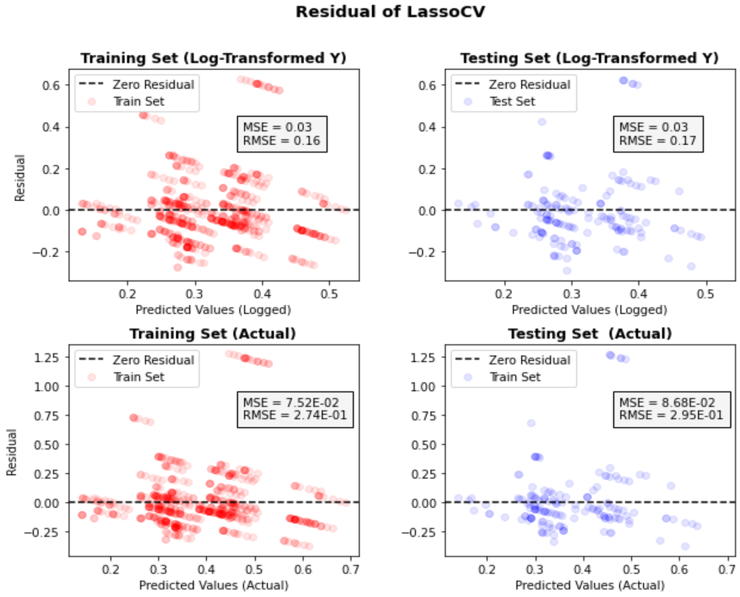


Figure 11. Residual of LassoCV

Training Curve:

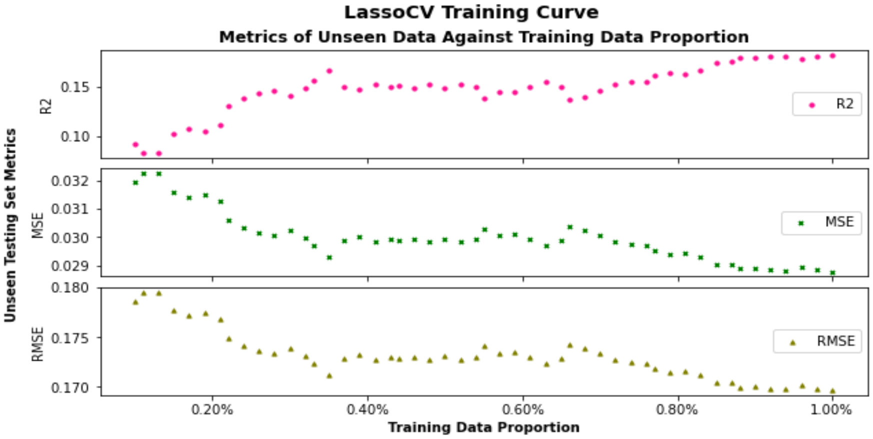


Figure 12. LassoCV Training Curve

- From the training curve above, we can tell the effect of our predictors on flight fare is relatively low (R-squared around 20%). It suggests the major factor determining the price of a flight is not included in our independent variables.

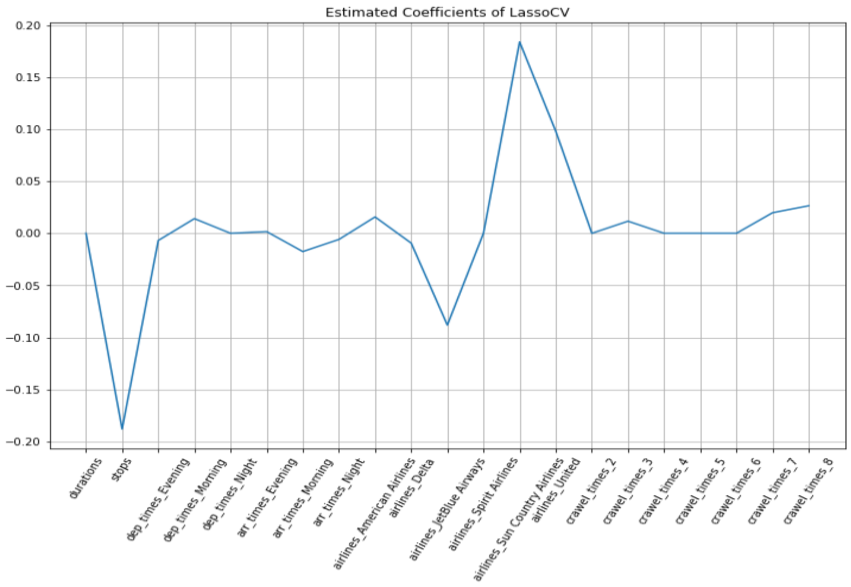


Figure 12. Estimated Coefficients of LassoCV

Coefficient Analysis

- However, among the factors listed in our dataset, the number of stops has the most negative correlation with the flight fare. As the number of stops increases, the price of a flight would decrease.
- The variable "Airline_Sun County Airline" has the most positive correlation with the flight fare. It advises that Sun Country Airline tickets on average are more expensive than other airlines.
- Crawl_time_8 and crawl_time_7 have the 2nd and 3rd most positive correlations with the flight fare, which advises that for customers, buying flight tickets around 7pm-8pm is a bad option.

II. Text Scraping- Sentimental Analysis

- Sentiment compound score for each airline:

Airline Name	Avg. Sentiment Score (compound)
United Airlines	0.34628625
Sun Country Airlines	0.283522222
Alaska Airlines	0.265913889
Delta Air Lines	0.161073718
JetBlue	0.142553091
American Airlines	0.131107022
Spirit Airlines	0.098653163

Table 3. Average Sentiment Score for Each Airline

All airlines have an overall positive sentiment score. United Airlines has the highest score, and Spirit Airlines has the lowest.

Conclusion and Future Direction

Recommendations to purchase cheap flight tickets in Expedia based on our sample data:

- If not time sensitive, try to add as many stops as possible to your trip. The more stops, the cheaper the ticket.
- Try to avoid Sun Country Airline, if possible, since it does not offer cheap airfares.
- On average, JetBlue Airways offers the cheapest flight tickets for both direct flights and transfer flights.
- Search flights during daytime, the earlier in the day, the better.
- Departure time and arrival time do not actually affect the ticket price, they do have some correlation, but they are not determining factors.
- United Airlines is the best choice because it has no positive effect on price and has the best sentiment score.

From our sentiment analysis we can conclude that airlines with cheaper fare do not have the best reviews. However, in practice, many customers do not care about reviews when buying a cheap flight.

For the future, we want to provide the code to the clients so that they can visualize the rate of change for the price of desired flight. Since our code can scrap Expedia effectively, customers will be able to make searches for any flights and obtain the best price possible.

References

Morries, L(2022) Visualizing Regression Training and Residual Plots
https://www.kaggle.com/code/leekahwin/visualizing-regression-training-and-residual-plots

Tweepy APIv2 Rate Limits
https://developer.twitter.com/en/docs/twitter-api/rate-limits#v2-limits