Predicting Airline Ticket Prices in the United States

Harshit Arora, Patrick He, Halle Steinberg, Sean Yu

Team 1

ISOM 676 Machine Learning II

Emory University

April 13th, 2020

Business Understanding

Except for trade and healthcare, there are few industries more crucial to an ever-changing, dynamic global economy than the airline industry. Airlines are responsible for merging cultures and moving customers from all over the world, bringing people near and far together at an efficient and affordable rate. Today, our world is being painfully stricken by COVID-19, a respiratory illness impacting millions. The airline industry has been one of the most severely affected around the world, as the airlines are facing the new challenges of a sharp decline in demand and uncertainty about when this will all end. The striking decrease in demand has led to many major airlines cutting costs as they see their revenue and earnings continue to drop by the day (Forbes.com). Now more important than ever is the ability to carefully keep customers on-board, both physically and in terms of continuing loyalty, and one way airlines can continue to do this and respond successfully to this crisis is by offering customers appropriate flight prices.

We don't know when this crisis will end, as it appears the situation gets worse by the day. However, as employees at one of the world's top travel agencies, our responsibility of securing the best travel options for our customers remains. This crisis will end, though, and when it does, we need to be ready for the (hopefully) large surge in demand from people all over the world looking to celebrate and jump on a plane to the most exotic and appealing destinations possible. These people have been hit hard, though; in the United States, many have lost their jobs, enduring a new normal that might include living paycheck-to-paycheck. We need to be able to provide those that may be struggling a realistic look at their opportunities to travel and ensure them that their travel will be worth it when they are able to do so. Our

customers are still looking for a high-quality experience at the most convenient prices, and honesty and accuracy in these prices will be more important than ever once world-wide airline travel picks back up.

Obviously, these are unprecedented times, and our data will likely not be entirely reflective of the current world since it is from 2018. Our hope remains, however, that times will return to normal sooner rather than later, and we will be able to get our customers traveling by plane at their leisure. Our goal is to predict the price of certain flight tickets depending on the time of the year, route, airline carrier, among other things, in order to provide our customers with options that fit their needs and current budget. By predicting what price a ticket might be sold at, we can advise our customer whether they should buy tickets now, or perhaps travel at a different time of the year when circumstances might change.

Tackling the issue of airline prices would significantly add value to our agency's business, as it would give us insight into what features might ultimately impact the price of certain flights, and whether or not there are differences among different carriers in the industry. With our data and insights, we can see whether certain airlines charge higher than others, if there are surges at different times of the year, and how distance between destinations plays a role in pricing. We know that airline prices are dynamically changing, sometimes on a day-to-day basis, so as travel agents, we can do our best to ensure our customers that they are getting the best price they can when they are traveling by plane. If we can determine with a high degree of confidence that certain trip routes at certain times of the year will be offered at an affordable

price to our customer, we can keep our customers (and their wallets) at ease and hopefully allow them to find some relaxation and excitement through travel after this crisis is over.

Ultimately, our goal is to explore various supervised learning data mining approaches to uncover solutions to questions revolving around airline ticket prices, allowing us to provide suggestions to our customers as to where they should fly, when they should fly, and which airline they should fly with, among many others. We will make sure our customers are getting the best deal for their personal needs and desires, a promise that will be crucial given today's current state of uncertainty and stress. We hope to bring our customers the best experiences they deserve and bring the world back together one flight at a time.

Data Understanding

Our goal was to use a relevant dataset that would provide us information on flight prices between multiple routes across all major airlines. Given the natural complexity in how flight routes are designed, we decided to limit our focus to only domestic flights in the United States. After thorough research of public datasets on Kaggle and Bureau of Transportation Statistics (BTS), the official data provider for all airline operations, we were able to find a dataset that provided us rich data on all major flight operations within the US throughout 2018. The variables found in our dataset are detailed below.

Variable Types	Variable Names	Description
Identifiers	ItinID	One of two identifiers used to indicate the order of the ticket booking
	MktID	One of two identifiers used to indicate the order of the ticket booking
Airport Identifiers	OriginWac	The area code for origin airport
	Origin	The city out of which the flight begins

	DestWac	The area code for destination airport
	Destination	The city out of which the flight ends
Code-Related Variables	MktCoupons	The number of coupons in the market for that flight
	Quarter	The quarter dates for the flight
	Miles	The number of miles traveled
	ContiguousUSA	Whether the flight is in the contiguous 48 USA states
	NumTicketsOrdered	The number of tickets purchased by the user
	AirlineCompany	A code that indicates the flight's airline company
	Origin_Region	A state code for origin area
	Dest_Region	A state code for destination area
Target Variable	PricePerTicket	The ticket price for the flight

Figure 1: Variable Descriptions

As part of understanding the information in each data instance, we first saw that we had close to 9.5 million observations, where each observation represented one purchase of tickets for a certain flight. None of the rows have any missing data or seemingly abnormal values. One thing to note here is that each instance is represented as a one-way flight. In the case of return trips, there are multiple rows for the same itinerary ID. This number makes sense, especially because it includes all commercial flights within the US for all airlines. Further, the distribution of ticket purchase transactions by the airline is fairly skewed towards four airlines (Southwest, Delta, American, and United), operating more than 50% of the total flights.

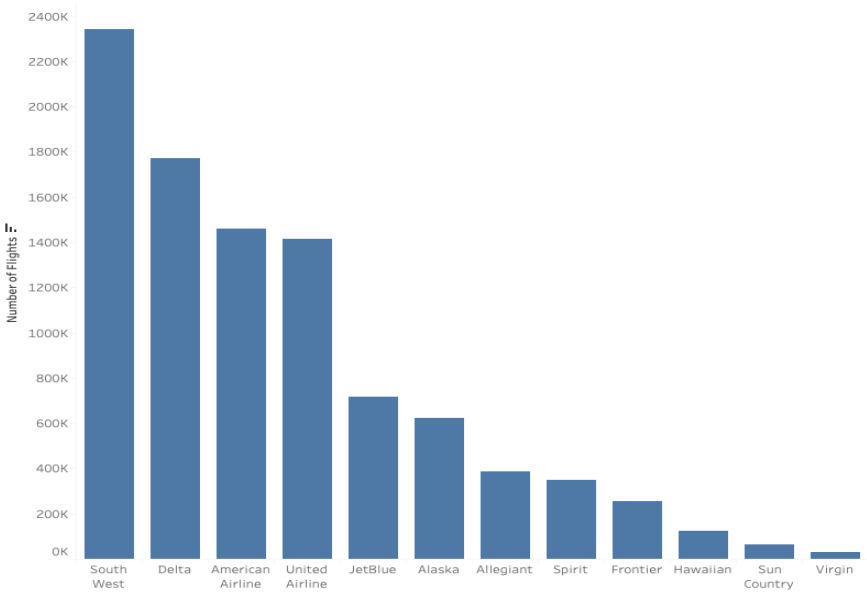

Figure 2: Rank of Airline Company by Number of Flights

	Miles	PricePerTicket ($)	NumTicketsOrdered	MktCoupons
Mean	1202.72	232.83	2.41	1.01
Median	1028	197.19	1	1
Min	11	50	1	1
Max	5095	1000	2	3
SD	697.61	143.72	2.88	0.13

Figure 3: Descriptive Summary of Numerical Variables

As expected, we see that PricePerTicket has a higher mean than the median, which indicates that the distribution of price could be right-skewed, and the density plot shown below confirmed this belief. There are many possible reasons for this to be the case, such as a few people purchasing expensive tickets at the last moment, or the presence of less common, exotic destinations that are typically more expensive to travel to than the vast majority of places. Given this would be our dependent variable, we wanted to keep this in mind when considering transformations that can improve model performance.

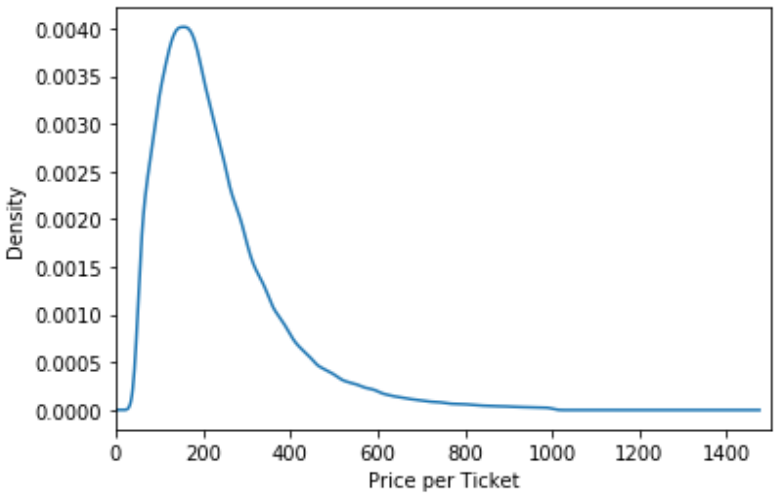

Figure 4: Distribution of PricePerTicket

Next, we wanted to better understand the kind of routes on which flights operate within the US. In reality, prices tend to vary hugely across routes, which, in turn, may depend on

several factors. That said, we saw that there were 242 different Origins and Destinations in the data, and this was because each of these represented a specific airport. Of course, we know that each state or city can have multiple airports. However, for the sake of a prediction model, we wanted to better understand if prices vary enough that each airport is treated differently despite being in the same state (e.g. JFK vs. LaGuardia airport within New York). If not, then it would be much more efficient from a data dimensionality and computational perspective to group airports by state or region.

Finally, while we did not have the actual date of each transaction, information on the quarter in which the flight was scheduled was available. We felt that this variable could be important, as this would allow us to control for some seasonality that may naturally exist in ticket prices due to varying demands throughout the year. Since our data only comes from one year, however, we know seasonality would not be able to be captured as thoroughly as if we had multiple years of data.

Data Preparation

Building what we have learned so far, it was important that we distinguished between categorical variables and continuous variables. Our categorical variables (Quarter, Airline, Origin, and Destination) were transformed into n-1 binary variables each using One-Hot Encoding, where n represents the number of levels or factors within each categorical variable. For improved performance, we also used LabelEncoder to transform all categories represented by text (e.g. American Airlines as AA) to numbers.

As mentioned earlier, it was also important for us to recognize the granularity of the Origin and Destination columns (airport vs. state vs. region) that would be used in any potential models we build. While we saw that using the data at the airport level would lead to high dimensional data (242 levels each for Origin and Destination), we felt comfortable using state level granularity as our exploration revealed that prices for similar routes (e.g. NY to CA) do not vary much across airports. This required us to seek additional data that provided a mapping between airports, states, and regions. This data was then augmented to the original dataset.

Granularity of Origin/Destination	Number of Levels
Airport Level	242
State Level	50
Region Level	6

Figure 5: Number of Levels for Each Level of Origin/Destination Granularity

Each destination and origin airport was mapped to its corresponding state, which was then assigned a region based on the map below. Our data also included some destinations that were in United States territories, such as Puerto Rico and Guam, and these were categorized as "Other". This gave us a total of 6 regions which were used in the model.

Figure 6: Map of Airport Regions

Next, since our dependent variable was found to be right-skewed, we decided to try transformations, such as the log transformation of PricePerTicket, in addition to the original variable available. The idea was to compare model performances on a hold-out sample before finalizing which model to use. Thus, data was also split into two sets of 70-30% training-validation data before any models were built on the prepared dataset.

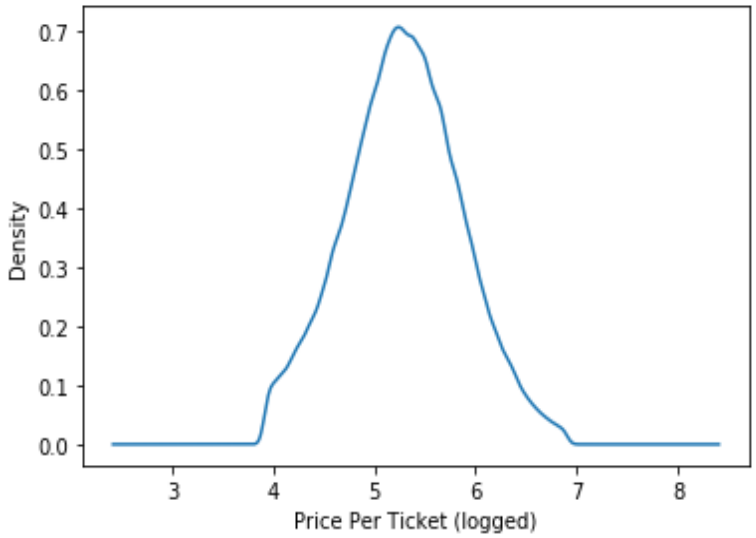

Figure 7: Distribution of Log(PricePerTicket)

Finally, we have several variables available, most of which make perfect sense to include in a model that predicts ticket price. Number of tickets purchased is one such variable. We considered if the value of this variable would be available when making predictions (i.e. understand if this could be a leakage variable, or is multicollinear with the price per ticket), but realistically, people know the number of people who want to travel when they decide to book their flights, and this number is not usually flexible. Thus, we concluded that we can keep this variable in our modelling process.

Modeling

We have a total of 9.5M instances of individual flight bookings. However, given all these bookings were made throughout 2018, we decided to use a random sample of 5M rows for our modelling process. This was done to ensure computational efficiency, since we were dealing with a relatively large dataset and hoped to perform hyperparameter tuning in all the models we would build.

Given that we were dealing with a regression task, we decided to start our modeling with a simple approach - Linear Regression. Because linear models require and work best when continuous variables follow a normal distribution, we used the transformed (log) price as our dependent variable. As a naive approach, we first used all available features to train the model on 70% of the data and test it on the remaining 30%. This gave us an RMSE of 212.47 on the train data and 228.03 on the test data. Clearly, there were several issues that needed to be dealt with. First, this was a clear case of overfitting. Second, we needed to see just how good or bad the model accuracy was.

To approach the first issue, we used regularization techniques (L1-Lasso and L2-Ridge). Lasso works well for models that are dealing with problems related to overfitting and allows for automatic feature selection by penalizing and making coefficients of uninformative variables zero. Between Lasso and Ridge, Lasso gave us a better overall performance, with an RMSE of 210.12 on the training data and 212.33 on the test data. We also approached feature selection using stepwise linear regression, but the model performance did not beat Lasso by itself.

We then tried Lasso, our best model thus far, on the same set of features predicting Log(PricePerTicket). By changing the scale of predictions back to linear scale, we were able to measure the performance on a more interpretable scale. Lasso with transformations performed slightly better, achieving an RMSE of 203.51 on training data and 204.17 on the test data. By intuition, it was possible that there could exist non-linearities in the data that can be better captured using higher order features and more sophisticated models. In particular, since Linear regression is a high bias modeling technique, we decided to consider more flexible approaches.

When considering a business task such as the one we currently face with predicting flight ticket prices, comprehensibility and ease of implementation are key. Regression Trees are among some of the most popular data mining tools, as they are easy to understand, easy to use, and computationally cheap. There are many practical advantages for using trees, so we decided to approach our regression task using a Regression Tree.

For Regression trees, we built two iterations of the model - one that uses an untransformed version of our target variable, PricePerTicket, and one that uses a log transformation. For our first Regression Tree model, we used the log transformation version of PricePerTicket. To ensure generalization performance, we used nested cross validation grid search to optimize various parameters for our model. The parameters that we optimized included the maximum depth of the tree, the minimum number of samples required at each leaf node, and the minimum number of samples required to split a node. We found that the optimal model used a maximum depth of 20 and a minimum of 5 samples required at each leaf node. We trained the optimal model on our training sample and tested using the test sample.

To evaluate performance on both the train and test data, we calculated RMSE on the non-logged scale for the dependent variable and saw a train RMSE of 113.08 and a test RMSE of 120.93. We can see already that this model is better at capturing the non-linearities in the data that our linear models could not, as the RMSE is much lower than before.

We repeated the same Regression Tree parameter optimization process, but instead using an untransformed version of Ticket Price, and found that the optimal model used a maximum depth of 20 and a minimum of 5 samples required at each leaf node. Again, we trained on the training sample and tested on the test sample and saw a training RMSE of 115.26 and a test RMSE of 121.62. We can see by comparing these two models that the transformed Regression Tree model has better predictive ability due to its lower RMSE.

So far our best model has been the transformed Regression Tree, but we know that there are techniques to combine various "best" models to improve performance. We know that meta-modeling techniques are procedures that involve some specific data mining technique as a subroutine. There are many ways to use them in predictive analytics, but one of the most common ways is through ensemble methods. The idea behind this method is to combine models, which yields diverse, independent predictions that may lead to improved outcomes. Predictive accuracy and generalization performance are typically improved through ensembling methods, so we thought it would be a good idea to use these methods in our analysis.

Random Forest is a popular ensemble method algorithm that builds many "random" trees and takes the average of the predictions for each tree as the final predictions for the model. It is an example of bagging, whereby k different tree models are trained on different

bootstrap samples (i.e., sample with replacement), and each tree is built using only a random subset of m attributes, where m is less than the total number of features, to calculate each split. This method tends to reduce variance in the predictions, thus reducing the possibility of overfitting. Since we've already tried two different Regression Tree models and we understand the ease of comprehensibility that they provide, we decided to try to optimize the parameters of a Random Forest model in order to take advantage of both the benefits of trees and those that come with ensemble method approaches.

In order to ensure generalization performance, we used nested cross validation to perform a grid search in order to optimize parameters for the Random Forest algorithm. For the first model, we used the log transformation of Ticket Price as our target variable. The optimal parameters were determined based on the root mean squared error (RMSE) of each potential model. We chose to optimize four different parameters - the number of trees to build in the forest, the maximum number of features to use for a split, the minimum number of samples required for a split, and whether or not bootstrap samples should be used when building the trees. Our optimal model ended up building 30 trees, using the default "auto" for the number of features used for a split, using a minimum of 8 samples required to split, and using bootstrap samples to build the trees. We trained the optimal algorithm on our training sample and tested using the test sample. To evaluate performance on both the train and test data, we calculated RMSE on the non-logged scale for the dependent variable and saw a train RMSE of 119.46 and a test RMSE of 122.13.

Again, we wanted to compare the difference in performance between the best Random Forest model with our transformed (logged) dependent variable and the best with the untransformed dependent variable. Using the untransformed target variable now, we used the same grid search technique as above, testing the same parameter options, and found that the best model resulted in building 40 trees, using the default "auto" for the number of features used for a split, using a minimum of 4 samples required to split, and using bootstrap samples to build the trees. Again, we trained the best model on our training data and predicted on the test data. We looked at performance on both the training and test data and saw a train RMSE of 116.41 and a test RMSE of 122.82. In the case of our Random Forest models, the transformed version of the dependent variable yields a slightly better result. However, our best yet is still the transformed Regression Tree.

Finally, given relatively good performance of tree-based models on our dataset, we decided to try XGBoost (Extreme Gradient Boosting), a decision-tree based ensemble algorithm that uses the gradient-boosting framework. With optimized gradient boosting, multiple trees are built in parallel and regularization techniques (L1 and L2), tree pruning (optimizing maximum depth parameter), and built-in cross validation is performed to improve model performance. In our case, several parameters were optimized on the models with a transformed (logged) dependent variable and with the untransformed dependent variable. Specifically, the parameters optimized were maximum depth, learning rate, minimum child weights, and 500 estimators were used. The model using the transformed dependent variable performed better, with an RMSE of 119.52 on training set and 120.56 on the test set.

Providing reliable ticket price results to our customers is our ultimate goal. By optimizing model performance and minimizing RMSE, we can ensure to the best of our abilities that our customers are getting honest and true price listings, allowing them to put their trust in us as a travel agency. For this reason, we needed to make sure we are paying close attention to detail by accurately and efficiently evaluating the performance of each one of our models.

Evaluation

After a thorough model building process, where we tried some very different modeling techniques each with its unique set of pros and cons, we wanted to select what the "best" model would be for our prediction task. While maximizing model performance (measured by minimizing RMSE, MAE, MAPE) is by far the biggest criteria, we wanted to be mindful of other factors, such as model comprehensibility, computational efficiency, and generalized performance.

Model	RMSE (test)	MAE (test)	MAPE (test)	Computational Efficiency (Based on time taken to train)
Linear Regression (with Lasso Regularization)	212.33	196.03	61.32	High
Regression Tree (Untransformed)	121.62	85.67	38.81	Medium
Regression Tree (Transformed)	120.93	82.81	43.92	Medium
Random Forest (Untransformed)	122.82	85.77	44.20	Low
Random Forest (Transformed)	122.13	83.01	38.93	Low
XGBoost (Untransformed)	121.03	81.46	38.44	Medium

XGBoost (Transformed)	120.56	81.30	37.81	Medium
XGBoost + DT Ensemble	120.11	81.22	37.77	Medium

Figure 8: Performance Metrics for Each Model

Clearly, tree-based models that were able to capture the non-linearity in the data and used parameter optimization performed better than linear models and lazy and slow learners such as K-NN (did not compute in time). Based on our performance measures and other factors, we decided that we would proceed with an Ensemble of XGBoost and Regression Trees as our final model, where the price prediction (on a transformed log scale) is averaged from the two models. As an add-on step, we stress-tested this model on additional, mutually exclusive test sets that were not used to train or test the models earlier (for computational reasons) to ensure that the model can handle more variances in the data and perform consistently well. We found that the model had an RMSE range of 120.05 to 120.38 in all of these test sets, which was well within our expected range.

Our overall goal is to use a model that can help potential tourists identify the best times to travel based on an optimal predicted price. Thus, it is important to address the customer needs by understanding which of the "features" (travel conditions) they are flexible on and which features are not up for negotiation. For instance, if a customer definitely wants to travel from New York to Los Angeles but is somewhat flexible on airline carrier and time of the year, our model can recommend the ideal options at the best price. With other future additions, such as ability to recommend the number of days in advance to book a flight, and time of the week/day to make a booking, we can potentially beat existing services into providing customers the best options with little research from their end. At the same time, there are errors

associated with the model (38% MAPE as of now), which does bring a risk of losing customers if we predict a price too high, when in reality, the flight is available for cheaper. In this case, customers won't end up using our services. While a tangible return on investment is hard to calculate right now, such a service can likely be used on a subscription basis and help frequent travelers save money in the long term.

Deployment

In today's unprecedented and uncertain world, as travel agents, we don't know when we can start helping our customers find travel experiences again. We hope that this time will be sooner rather than later, and when it is time, we want to be able to help our customers get back on planes and travel as easily and stress-free as possible. By taking action now, we can prepare for what's ahead. We will begin by using our new model to start predicting airline prices right away. Based on current government actions and results we have seen in China and other countries, our hope is customers will start traveling again when the summer comes. We will use our model to account for flight features that we know now go into predicting flight prices. With our model, we are able to give our customers options for the features they are flexible on - whether they want to travel in the winter vs. the summer or whether they need a direct flight or are fine with connections. Our model can provide each customer with the best flight price based on his or her "optimal" features.

In terms of practical use of our model results by our customers, we will construct and implement an online tool where our customers can input their own travel information - how many people will they be traveling with, when do they want to travel, where will they be flying

out of - in order to see a spread of options by airline, destination, and price point. We will create a hands-on, easy-to-use tool that customers can use at their convenience and receive results that are linked directly to a booking page. Our customers will be able to research and book flights in the matter of clicks, in turn enhancing customer experience and, ultimately, satisfaction.

People will be very tentative to travel once the pandemic dies down and things start to return to normal. We know this, and we can prepare for this. Essential business travel will start to return first - likely business travelers who are only purchasing one ticket for short-distance flights. We can use this information to target these travelers as the first to use our tool. We can offer them unique incentives by partnering with airlines to provide discounts or benefits as they start traveling again. Similarly, once the summer months hit, we hope flyers will start to book vacation travel, which we can prepare for by offering prices for popular vacation destinations and origins from popular flight hubs, such as Atlanta, New York City, Chicago, or Los Angeles. We can prepare for some things to return to normal, but we need to be wary of tentativeness and uncertainty that flyers will have to begin with. Understanding that these will be the first travelers to return to airplanes, we can pair with participating airlines to offer these travelers deals and promotions when using our interactive tool.

We will need to take careful consideration to understand how this pandemic might change our predictions, but for now, having a baseline ticket price to offer our customers will be a good starting point. Our firm, and other firms like us, needs to understand that this model takes into consideration data from a fairly uneventful timeframe - 2018 did not see any world-

wide pandemics or massive economic events that could have accurately modeled what 2020 has seen. We need to understand that we will need to handle the next few months with caution and care when using our model. If we avoid these precautions and provide prices to our customers based solely on our 2018 data, there is the possibility that we are providing too high or too low of prices, and our customers might not be willing to travel or they might leave our travel agency. By pricing too high or too low, we risk losing our loyal customers and any new business they might bring in. Current research shows that cheaper flights and flexible cancellations may be around for a while, as air passenger traffic is down nearly 90% right now. A new report from an analyst at Stifel, an investment banking firm, predicts that air travel demand will not return to pre-outbreak levels until at least mid-2021 (Insider.com). We know that travel will eventually rebound, but the question of "when" remains. That being said, on the bright side, we do hope that the economy and airline industry will begin to return to "normal" by the end of the year, hopefully seeing a rather uneventful 2021, 2022, and so on.

As we've mentioned, generating an exact return on investment of our model is difficult to do, as we don't have spending on a customer basis, and we won't know exactly what the market will be like once the pandemic is over. We would expect that our return would be a function of a customer's continued loyalty in using us as his or her preferred travel agency and the amount of business we might gain from loyal customers' word of mouth. If we were to provide dishonest or inaccurate pricing estimates to our customer, we risk losing that customer to another travel agency. If we can provide accurate estimates, however, we can maintain and increase customer retention. These customers would be more likely to book with us again and again, driving up our revenue and profits. If we could obtain airline ticket spending at a

customer level, we could determine how much our most loyal customers are spending with us and generate a baseline before and after our model implementation.

Future Improvements

By using relevant, reliable and up-to date datasets and methodology, we have successfully built an easy to implement airline pricing prediction model that can predict prices at a flight-level granularity. Airline pricing, however, is very complex and dynamic, and the variables we are able to include in our model so far are only a subset of those important factors that tend to affect airline prices. For instance, airline pricing can also vary by the number of days before the departure date that a flight is booked, booking details such as time of day and day of the week, booking agent website (different agents have varying prices as well). Further, our model is built on data for Economy Class seats (or equivalent for some airlines) but some customers may prefer Business or First Class tickets as well, data for which wasn't available yet. Finally, external factors such as weather and rare events (e.g. COVID-19) can massively impact airline prices as well, as we are currently seeing in today's environment. Coronavirus has taken over many businesses across the world, and one of the most affected industries has been the airline industry. Having the ability to include some of these factors and having data for additional years (e.g. 2019) can help us build a more comprehensive model that can be used to make more reliable predictions and also help us understand the relative importance of factors in ticket pricing.

Works Cited

Hoeller, Sophie-Claire. “Cheaper Flights and Flexible Cancellation Fees Might Be Here for a While. Here's What Air Travel May Look after the Pandemic.” *Insider*, 7 Apr. 2020, www.insider.com/how-coronavirus-will-affect-travel-future-2020-4.

Wyman, Oliver. “How COVID-19 Is Transforming Global Aviation's Outlook.” *Forbes*, Forbes Magazine, 6 Apr. 2020, www.forbes.com/sites/oliverwyman/2020/04/06/how-covid-19-is-transforming-global-aviations-outlook/.

Team Member Contributions

Harshit Arora - primarily worked on report, presentation, and model building

Patrick He - primarily worked on EDA and model building

Halle Steinberg - primarily worked on report, presentation and model building

Sean Yu - primarily worked on model building