Group Project 2: A Price Optimization Model for Delta Air Lines, Inc.

Logan Donley, Kshitiz Kharel, Cheng-Yu Lee, Hitali Shah, and Idrissa Stephen
ISM 6136 S21: Just Mining Our Own Business Group
May 1, 2021

Background: Delta Air Lines Price Optimization Model

Delta Air Lines, Inc. (NYSE: DAL) is one of the largest commercial airlines in the United States of America, serving nearly 200 million global customers each year (Delta Air Lines, 2021). Consistently rising customer demand – both prior to the COVID-19 pandemic and now as U.S. and other global customers return to a new state of normalcy – offers the firm an opportunity to improve their profitability and competitiveness by incorporating a data mining price optimization model into their business operations. Price optimization models are powerful tools for helping firms calculate variations in demand, craft targeted marketing and promotion strategies, manage dynamic inventory levels, and keep up with rapidly changing customer needs (Bain & Company, 2018). Implementing such a model at Delta will allow the firm to more effectively adjust their pricing strategy to maximize profit. Also, firm analysts will be able to better simulate how different consumer segments will respond to changes in pricing, which will allow for more accurate consumer targeting at different price levels. Optimized pricing will directly contribute to higher levels of customer satisfaction by reducing price volatility in a complex and dynamic market.

A data mining application to optimize price must necessarily attempt to account for the firm's overall pricing strategy and business model as well as the value of the product for the consumer and the cost of providing the product. Therefore, the model is based on carefully designed data mining and machine learning approaches to ensure meaningful input variables that will produce useful outputs. Ideally, this would involve accessing comprehensive historical information that includes past promotions, product and service prices, product sales volume, economic conditions impacting the industry, prices of key competitors, fixed and variable costs, seasonal trends, and product availability (Bain & Company, 2018). The ideal model would also function best if the decision-making for design and implementation were to be conducted cross-functionally with input from all relevant areas of the business. Finally, it will be important to clarify the value of the data mining model for this business opportunity to ensure it fits with the firm's long-term strategy.

In the absence of ideal circumstances and with limited access to all the desired data as articulated in the first part of the project, we have made a few revisions to the problem statement based on the data available over the internet to do the analysis.

Motivation

As previously mentioned, setting transportation fare prices that meet the unique needs for every traveler is an impossible endeavor, but the airline industry has been on the leading edge of revenue management technology for more than four decades. Dynamic inventory pricing, demand forecasting, fare restrictions, and the development of increasingly sophisticated

optimization tools have been foundational practice at firms like Delta since the 1980s, making the question of how to price a plane ticket one of the most consistent and compelling problems in the industry. However, the emphasis of these analytics efforts has been on pricing of core tickets in response to an endlessly changing environment.

To realize the true potential of optimization technology, we posit that airlines must seek to adopt predictive machine learning models that consider not only available inventory and consumer demand but also other factors like competitive dynamics, flight velocity, traveler demographics, geographic constraints, and ancillary services, among other considerations. Historical methods of reactive analysis will no longer suffice for firms to manage the nuanced pricing that is critical to meet profit and loss targets.

As a starting point to enable this transition, we believe the data needed for a data mining-driven approach to price optimization already exist in airline industry databases. This historical transaction information about travel patterns, booking behavior, sales prices, inventory levels, flight details, and other price-influencing attributes already exist in Delta's databases. In lieu of actual data from the airline and a dedicated business intelligence pipeline, we will use publicly available datasets and analysis tools to further articulate the data mining and machine learning price optimization approach discussed.

Description of Dataset

The data source for our project is Airline Origin and Destination Survey (DB1B) provided by the Bureau of Transportation Statistics. It is a 10% sample of airline tickets from reporting carriers collected by the Office of Airline Information of the Bureau of Transportation Statistics. Data includes origin, destination and other itinerary details of passengers transported. This is a close representation of the essential attributes for the optimization model that we initially proposed.

Three different categories of table are available in DB1B: Coupons, Markets, and Tickets.

DB1BCoupon

This table provides coupon-specific information for each domestic itinerary of the Origin and Destination Survey, such as the operating carrier, origin and destination airports, number of passengers, fare class, coupon type, trip break indicator, and distance.

DB1BMarket

This table contains directional market characteristics of each domestic itinerary of the Origin and Destination Survey, such as the reporting carrier, origin and destination airport, prorated market fare, number of market coupons, market miles flown, and carrier change indicators.

DB1BTicket

This table contains summary characteristics of each domestic itinerary on the Origin and Destination Survey, including the reporting carrier, itinerary fare, number of passengers, originating airport, roundtrip indicator, and miles flown.

Although the three databases are a source of extensive information, they were merged to create a unique data source. We used Microsoft Power BI to merge the datasets and create a master dataset. Consider the following tables to understand the relationships between the tables:

	Table Type		
Field	Coupons	Market	Ticket
Itinerary ID	Y	Y	Y
Market ID	Y	Y	
Year	Y	Y	Y
Quarter	Y	Y	Y
Passengers	Y	Y	Y
Fare Class	Y		
Market Fare		Y	
Roundtrip Fare			Y
Distance	Y	Y	Y
Origin	Y	Y	Y
Destination	Y	Y	
Carrier			Y

Field	Combined table
Itinerary ID	Y
Market ID	Y
Year	Y
Quarter	Y
Passengers	Y
Fare Class	Y
Market Fare	Y
Roundtrip Fare	Y
Distance	Y
Origin	Y
Destination	Y
Carrier	Y

Apart from these fields, we used the Local Area Transportation Characteristics by Household (LATCH) dataset publicly available from the Bureau of Transportation Statistics. This dataset clusters U.S. states based on the total population, household counts, median household income and various other socio-economic factors. We use the clusters to create a separate variable to indicate the geographic origin of the flight. Our final dataset has the following fields:

1. ORIGIN (Origin Airport Code)
2. DESTINATION (Destination Airport Code)
3. PASSENGERS (Number of Passengers)
4. ROUNDTRIP (Round Trip Indicator, 1=Yes)
5. FARE CLASS
    a. C - Unrestricted Business Class

   b. D - Restricted Business Class
   c. F - Unrestricted First Class
   d. G - Restricted First Class
   e. U - Unknown
   f. X - Restricted Coach Class
   g. Y - Unrestricted Coach Class
6. DISTANCE_IN_MILES (Itinerary Miles Flown/Track Miles)
7. FARE (Itinerary Fare Per Person, **DEPENDENT VARIABLE**)
8. CARRIER (CARRIER CODE, Delta Airlines Inc. = DL)
9. ORIGIN_CLUSTER (State of origin designated by cluster in LATCH)

A snapshot of how the data would be modelled and transformed is as follows:

ORIGIN	DEST	PASSENGERS	ROUNDTRIP	FARE CLASS	DISTANCE (MILES)	FARE (USD)	CARRIER	ORIGIN CLUSTER
JFK	LAX	195	0	X	2475	448	DL	1
JFK	LAX	168	0	X	2475	358	DL	1
LGA	ORD	151	0	X	733	88	DL	1
JFK	LAX	151	0	X	2475	98	DL	1
LGA	ATL	146	0	X	762	83	DL	1
LGA	MCO	131	0	X	950	78	DL	1
JFK	SJU	115	0	X	1598	74	DL	1
ORD	LGA	112	1	X	733	176	DL	2
LGA	ORD	112	1	X	733	176	DL	1
ORD	LGA	108	1	X	733	236	DL	2

The following data filters are applied to the dataset:

- Only considered data for Carrier = DL (i.e., Delta Airlines Inc.)
- All zero and null values for fares were removed
- All round-trip fares having values less than $50 and greater than $5,000 were removed
- All round-trip distances less than 200 miles and greater than 10,000 miles were removed

Our final dataset contains approximately 67,000 records.

Solution Methodology and Evaluation Metrics

The dataset was divided into training and testing sets by randomly selecting 80% of the data for training and the remaining 20% for testing. We ran Boosted Decision Tree Regression and Decision Forest Regression on the training set to train the model. The trained model will be used

to predict the fare prices for the testing set. The training, testing, and implementation of the models was conducted using Microsoft Azure Machine Learning Studio (classic). The following steps were implemented in Azure:

1. *Import our processed final dataset* – data processing, cleaning and transformation of raw data is done in Excel and Power BI which is covered in the Description of Data segment.
2. *Split the data* – randomly 80% as training and 20% as testing set.
3. *Train model module* – to train the model, left input is Boosted Decision Tree Regression or Decision Forest Regression, and right input is the training set.
4. *Score Model module* - to make the prediction on the test dataset, left input is the train model module and right input is the testing set.
5. *Evaluate Model module* - to compute the regression performance, takes the score model module as input.

The model from the Boosted Decision Tree Regression and Decision Forest Regression will be evaluated using the following metrics:

1. *Mean Absolute Error (MAE)* measures how close the predictions are to the actual outcomes; thus, a lower score is better.
2. *Root mean squared error (RMSE)* creates a single value that summarizes the error in the model. By squaring the difference, the metric disregards the difference between over-prediction and under-prediction.
3. *Relative absolute error (RAE)* is the relative absolute difference between expected and actual values; relative because the mean difference is divided by the arithmetic mean.
4. *Relative squared error (RSE)* normalizes the total squared error of the predicted values by dividing by the total squared error of the actual values.
5. *Coefficient of determination (R squared)* represents the predictive power of the model as a value between 0 and 1. A value of 0 means the model is random (i.e., explains nothing); 1 means there is a perfect fit.

Final Model building in Azure

As previously mentioned, we have implemented two algorithms, Boosted Decision Tree Regression and Decision Forest Regression. For each of the model, we conducted two experiments to evaluate the parameters that provide the least erroneous solution:

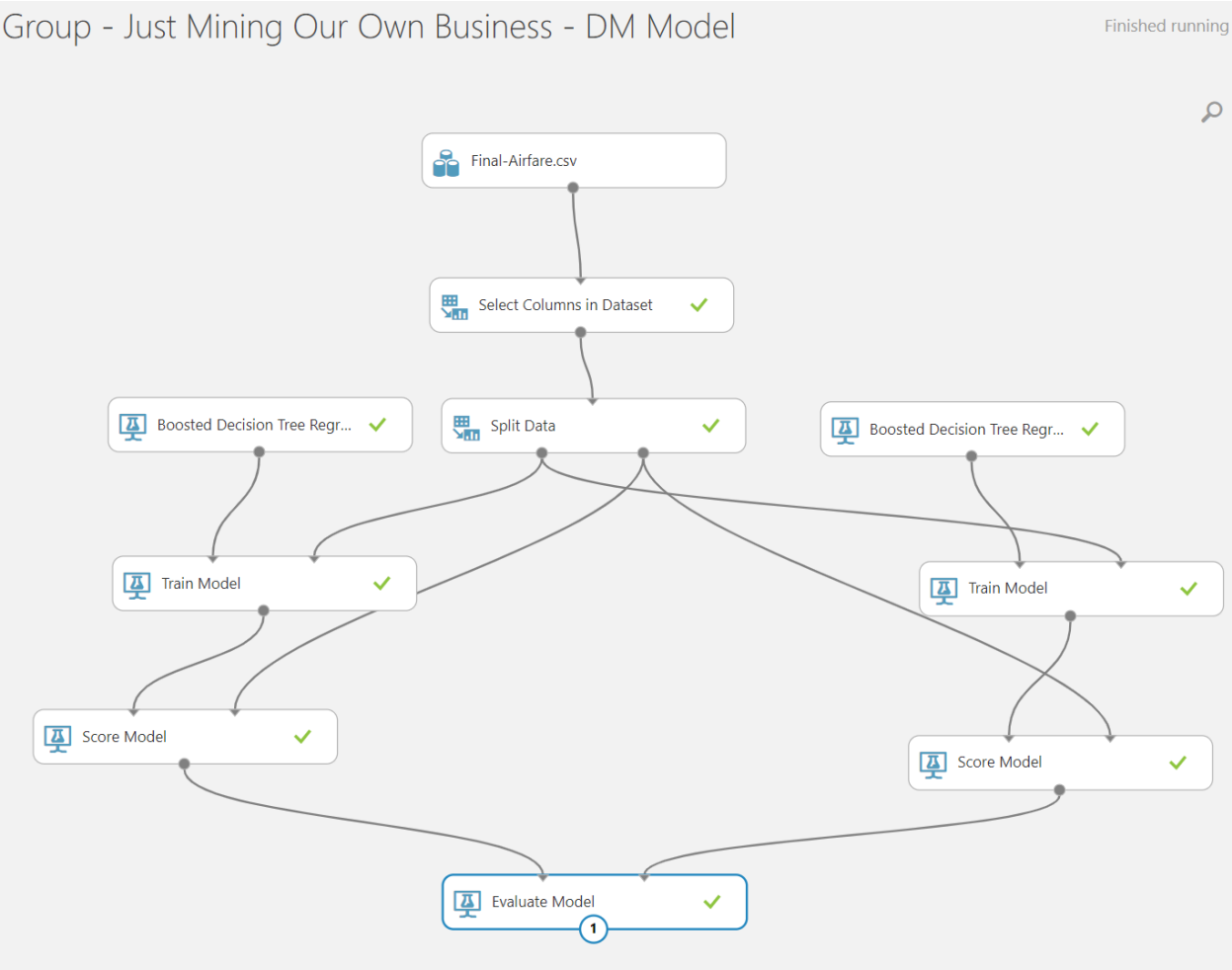
Boosted Decision Tree Regression

- Experiment 1: All variables selected, and parameters were as follows:

- Max # of leaves: **10**
- Min # of training instances: **15**
- Learning rate: **0.2**
- Number of trees: **10**

- Experiment 2: All variables selected and all parameters as default

A snapshot of how the experiments is modelled in Azure is as follows:



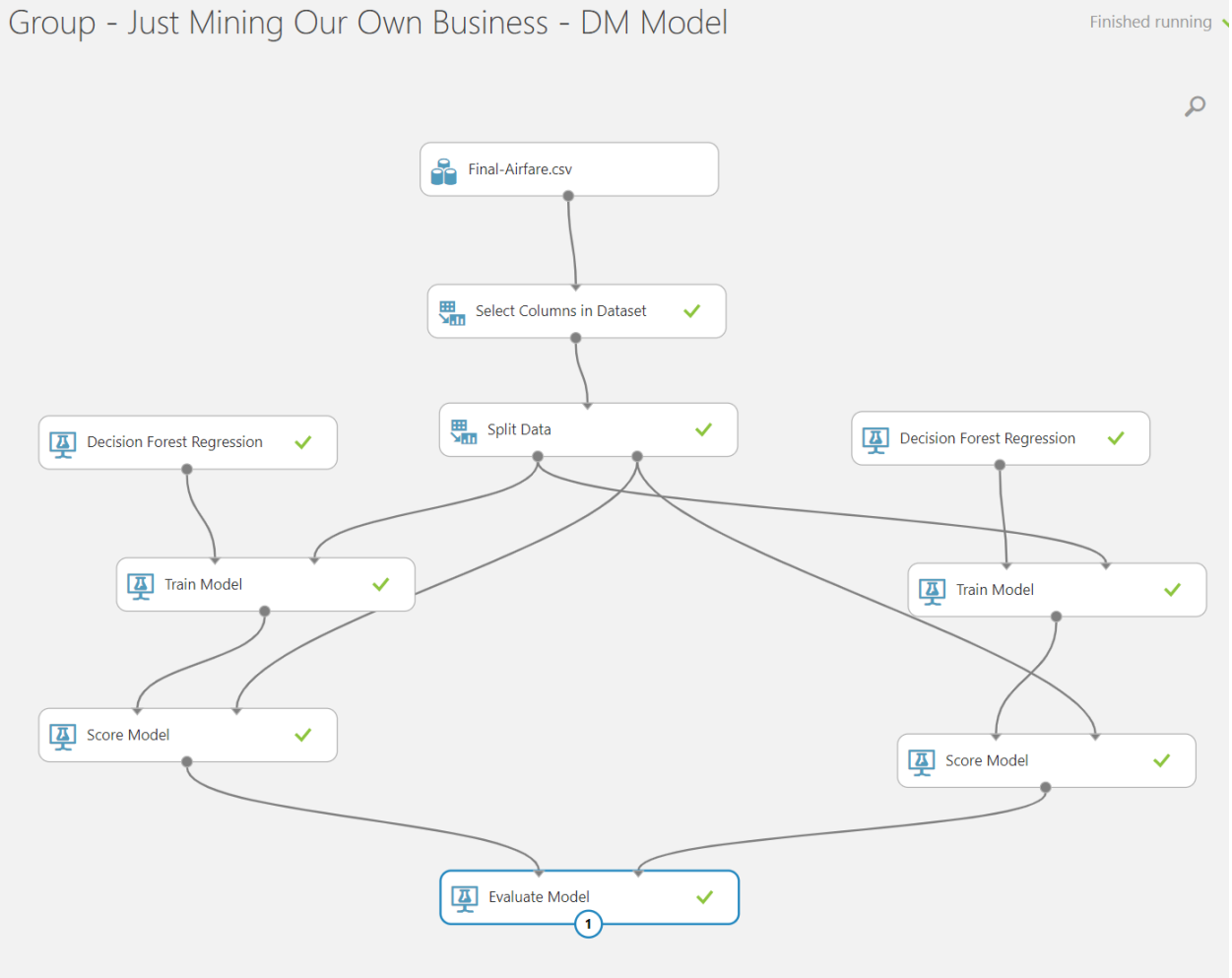
Boosted Decision Tree	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Relative Squared Error	Coefficient of Determination
Experiment 1	90.654	136.758	0.310	0.083	0.917
Experiment 2	38.423	91.259	0.131	0.037	0.963

Experiment 2 performed better than the Experiment 1 in terms of our evaluation metrics.

Decision Forest Regression:

- Experiment 1: All variables selected, and parameters were as follows:
  - Number of decision trees: **10**
  - Maximum depth of the decision trees: **16**
- Experiment 2: All variables selected and All parameters as default

A snapshot of how the experiments is modelled in Azure is as follows:



Decision Forest Regression	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Relative Squared Error	Coefficient of Determination
Experiment 1	54.449	113.742	0.186	0.057	0.943
Experiment 2	55.656	114.369	0.190	0.058	0.942

Experiment 2 performed slightly better than the Experiment 1 in terms of the evaluation metrics.

Summary of models

Comparing the evaluation metrics from both models:

	Boosted Decision Tree Regression	Decision Forest Regression
Mean Absolute Error	34.423	55.656
Root Mean Squared Error	91.259	114.369
Relative Absolute Error	0.131	0.190
Relative Squared Error	0.037	0.058
Coefficient of Determination	0.963	0.942

As shown in the above table, the Mean Absolute Error with Boosted Decision Tree Regression is approximately 35, indicating that our mean predicted airfare is only $35 away from the mean actual airfare. RMSE with the Boosted Decision Tree Regression is also less as compared to the Decision Forest Regression.

A worrying factor here is that with Boosted Decision Tree Regression the error rates are quite low suggesting that probably there can be a chance of overfitting the model. To avoid that we tried to experiment with other model parameters keeping the number of trees to the half of the data points available.

Recommendations

This model provides projected airfare for a given itinerary based on various attributes with high accuracy levels. If the firm's analysts can consistently translate the target price predictions into actionable recommendations, a high-performing model as described above would provide strategic advantages in the ability to manage aircraft usage, target marketing messages, and proactively set competitive prices. In effect, the model ought to contribute to the firm's dynamic pricing strategy by helping to highlight unique combinations of price-influencing factors that proactively drive customer demand and maximize profit margins. Once integrated with existing systems, the model has the potential to make pricing changes more efficient and less frequent, or even automate the process completely.

To effectively implement the proposed model at Delta Air Lines, a key recommendation would be to integrate the model with their existing Online Transaction Processing (OLTP) system and Global Distribution System (GDS). Using the OLTP to mine and create datasets for the price

optimization model provides the opportunity for continuous refinement of the model and will ensure that the optimized price outputs can be generated and updated for customers in near real time. Some consideration will need to be given as to what events trigger a price change and how often the optimization process can be efficiently acted upon (i.e., how quickly the other systems can adjust to new target prices for a particular customer profile). Integration between the model and the GDS will effectively allow Delta to lead the industry into the next generation of price optimization techniques, as it would expose the model to an infinite amount of external firm data about consumer travel preferences and priorities. For example, the model could be used to optimize not only the price of a plane ticket, but also dynamically predict ideal bundle prices for a customer's flight (with ancillary costs such as excess baggage and premium meals already accounted for), rental car/rideshare, and lodging based on real time data.

Maximizing these existing systems with the proposed data mining approach will give Delta an unparalleled strategic advantage by aligning prices with customer value rather than customer demand.

References

Bain & Company, Inc. (2021). Price Optimization Models. Retrieved from https://bain.com/insights/management-tools-price-optimization-models.

Delta Air Lines, Inc. (n.d.). Delta and CLEAR partner to improve airport experience. Retrieved from https://news.delta.com/delta-and-clear-partner-improve-airport-experience.

Delta Air Lines, Inc. (2020). Delta launches interactive travel requirements map to take more stress out of planning your next trip. Retrieved from https://news.delta.com/delta-launches-interactive-travel-requirements-map-take-more-stress-out-planning-your-next-trip.

Delta Air Lines, Inc. (2021). Investor relations. Retrieved from https://ir.delta.com.

Delta Air Lines, Inc. (2021). Privacy policy. Retrieved from https://delta.com/us/en/legal/privacy-and-security.

IBM Corporation. (2021). Optimization modeling. Retrieved from https://ibm.com/analytics/optimization-modeling.

Panigrahi, G. (2019). Airline fare prediction. Retrieved from https://kaggle.com/gpanigrahi/airline-fare-prediction.

Evaluate Model in Microsoft Azure (2019) Retrieved from https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/evaluate-model

DB1B data. Retrieved from https://www.transtats.bts.gov/