TIME SERIES FORECASTING OF DELTA AIR LINES STOCK PERFORMANCE USING ELASTIC NET REGRESSION

CITADEL SUMMER INVITATIONAL DATATHON

Sanjit Kobla
skobla@sas.upenn.edu
University of Pennsylvania

Arjun Agrawal
arjunag@seas.upenn.edu
University of Pennsylvania

Angelina Zheng
azheng@wharton.upenn.edu
The Wharton School

Pranav Ramesh
pranavramesh@college.harvard.edu
Harvard University

July 30, 2023

1 Executive Summary

Our team was motivated by the fact that companies do not release their quarterly earnings until between two and six weeks after a quarter ends. So, we posed the question: Can we find a relationship between monthly data and Delta Air Lines' post-earnings release quarterly stock price? We have two major findings: (1) there exists a strong and statistically significant relationship between our collected data and Delta's quarterly stock price and (2) we are able to accurately predict Delta's quarterly stock prices with our monthly data.

1.1 Key Finding One: Monthly Data is Strongly Related to Delta's Quarterly Stock Price

The results of our most pared down analysis are shown in the graph below. Our analysis demonstrates that there is significant impact from these four inputs:

- *CPI (Consumer Price Index):* As CPI increases, Delta's stock price increases.
- *American Express's Stock Performance:* As American Express's stock price increases, Delta's stock price increases.
- *Delta's Previous Stock Price:* As Delta's stock price from the day before the prediction date increases, the predicted day's stock price increases.
- *Jet Fuel Price:* As the price of jet fuel increases, Delta's stock price decreases.

These strong and statistically significant correlations culminate in the prediction graph shown in Figure 1, which closely aligns with the actual values of the stock. The model developed for this analysis is a multivariate regression model with an R^2 value of 0.998, indicating that approximately 99.8 percent of the variability in Delta's quarterly stock price can be explained by our model.

The choice of these four factors in the model is well-founded and logical. Aviation CPI and American Express's stock performance are directly related to Delta's revenue, encompassing profit from airline ticket sales and returns from their loyalty program. Meanwhile, the price of jet fuel is a key driver of fluctuations in Delta's costs. Finally, considering Delta's stock price from the day before the prediction accounts for extraneous market factors at the time of forecasting. Overall, these four factors form a coherent and statistically significant set of inputs, enabling us to make accurate predictions of Delta's quarterly stock price.

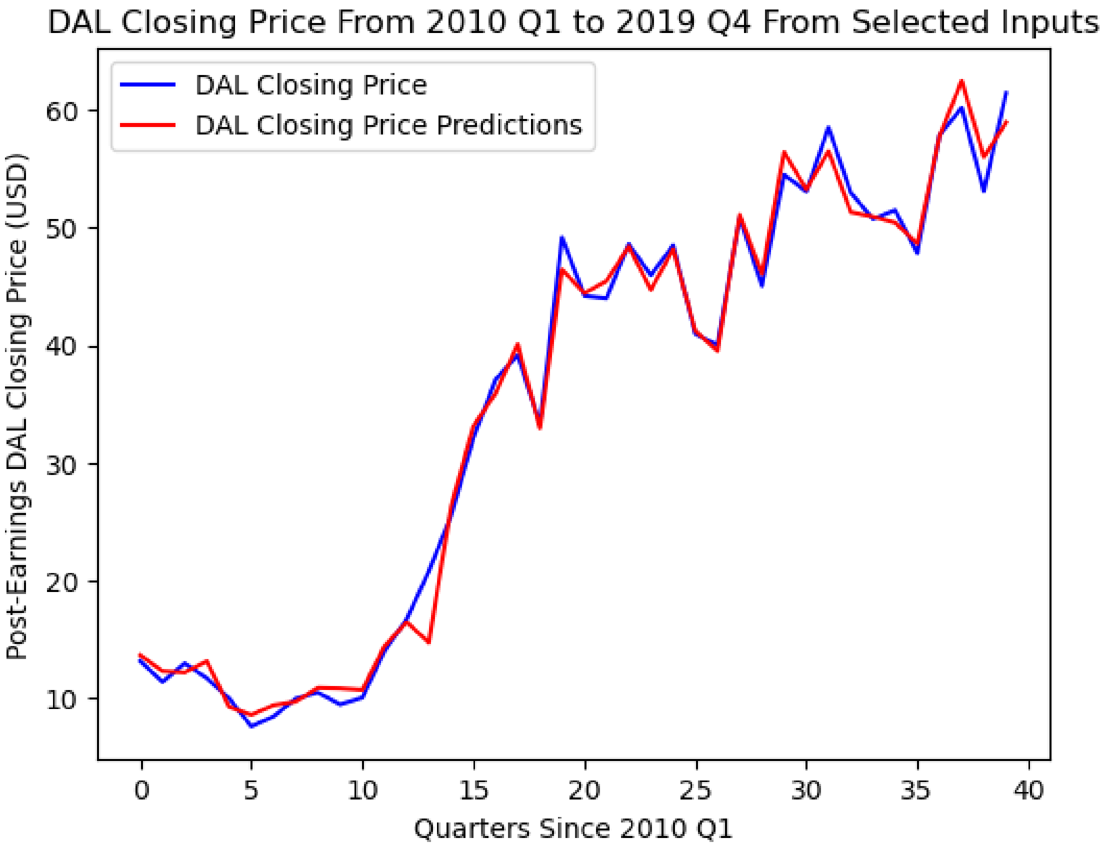

Figure 1: Delta's Stock (DAL) Price by quarter from 2010 to 2019 predicted by four selected inputs

1.2 Key Finding Two: Monthly Data can Predict Delta's Quarterly Stock Price

Using our model from Key Finding One, we created a prediction model for what Delta's quarter end stock price will look like for the next three quarters after the previous data is put into our model. For example, a prediction of Delta's 2018 Q3 through 2019 Q1 stock prices based on training data from 2010 Q1 through 2018 Q2 are shown below in Figure 2.

This model only relies on the four inputs explained from the first key finding and has R^2 values as high as 0.926, which means that we will be able to very well account for the variation in stock price with our prediction model to a degree of 92.6 percent.

Through coefficient significance tests and time-series analysis, we have been able to causally link four input data sets with the stock price of Delta each quarter and use these input data sets to predict Delta stock prices for future quarters. This model, if expanded upon with a larger data set, can be applied to a wide array of different airlines and can be used to predict future Delta stock prices in each quarter.

2 Technical Exposition

2.1 An Overview of the Methodology

We began with the question: can we find data that correlates with Delta Air Lines' (DAL) stock price, and if so, can we use this to predict Delta's stock price and determine if these relationships are causally related? We chose Delta Air Lines because, according to the database Statista, it is the third largest airline in the US. This means that the airline is large enough to not be heavily impacted by small events such as sharp weather fluctuations in a subset of the continental US and there would be enough publicly available data for us to work with. Furthermore, we chose only one airline,

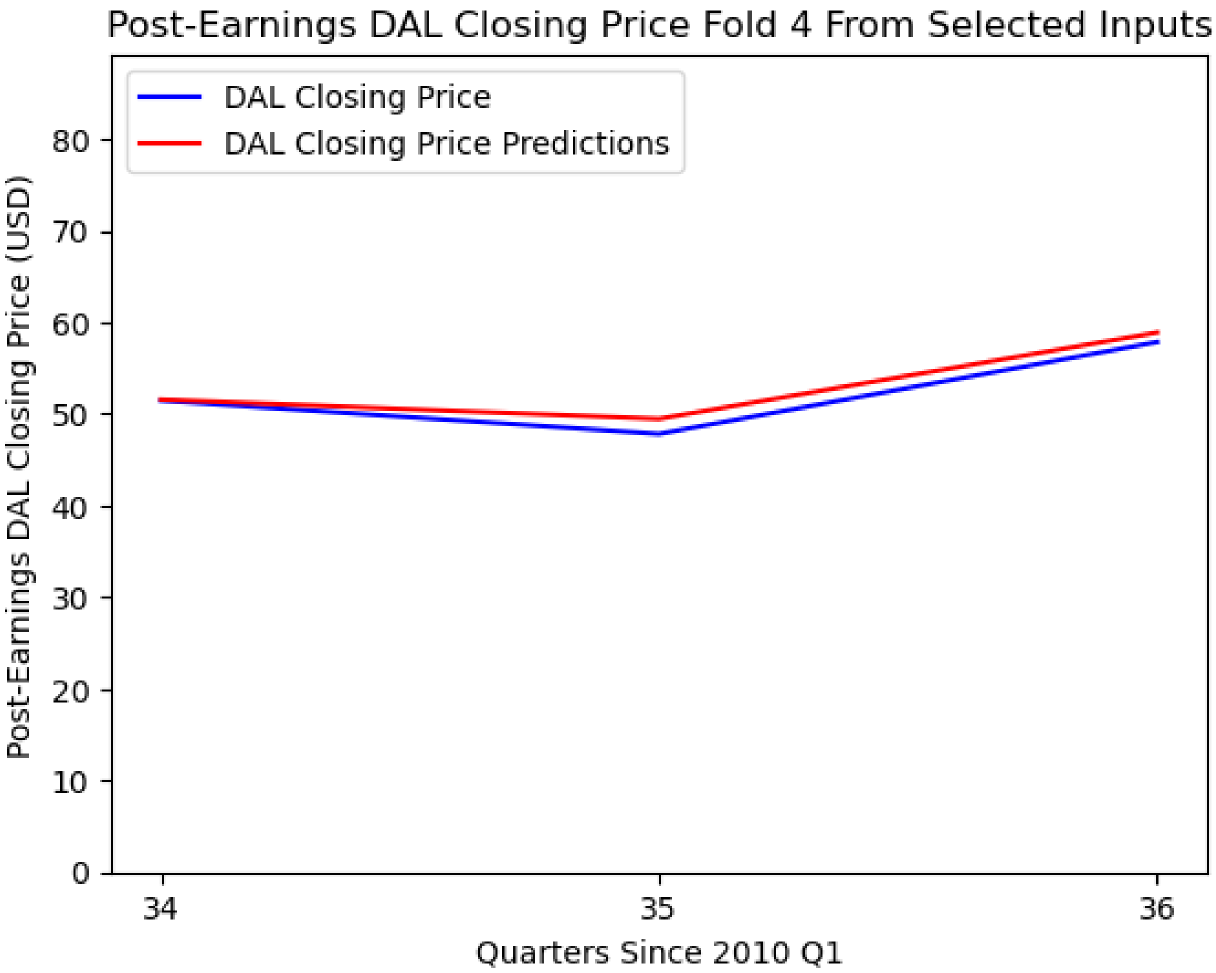

Figure 2: Selected Inputs Predictions for 2018 Q3 - 2019 Q1

Delta, to have more specific and nuanced predictions than generalized predictions for a whole industry that is highly dependent on many different macroeconomic factors.

In the following technical exposition, we begin by discussing our initial data exploration with the provided data sets, and why we shifted away from them and the time frame in favor of monthly and quarterly statistics over the 10 years between 2010 and 2019. We then discuss our new data sets that we collected and cleaned, and the intuition behind using them. We continue on to show our results of correlation, prediction, causation, and feature selection for predicting Delta Air Lines' stock price each quarter between 2010 and 2019 right after quarterly earnings are released.

2.2 Exploratory Data Analysis

We began by exploring the relationship between monthly data for Delta derived from the provided data sets to determine whether there exist any preliminary correlations and if we could extrapolate it to a broader time range based on the statistical significance of the correlations. Using the `events_US` data set, we calculated the number of public events that occurred each month. Using the `flight_traffic` data set, we calculated monthly flight traffic data by the unique route, distinguishing between pairs with flipped origin and destination airports. The data includes the total number of flights, cancelled flights, and diverted flights, and the total number of minutes delayed by delay type by month. Finally, using the `fares` data set, we calculated quarterly revenue by unique route and recorded the distance of each route. Given we only had quarterly revenue by route, when we merged the data set with our monthly flight traffic data set, we calculated monthly revenue by route by dividing the quarterly revenue by the total number of flights on that route within that quarter, and then multiplying by the number of flights on that route within each month, which is essentially a weighted average of the revenue by the relative number of flights.

In addition, we calculated the total distance traveled monthly on each route by multiplying the number of flights on that route in that month by the distance traveled stored in our fares data set. Subsequently, we synthesized twelve months of data for Delta, creating a data set with the total number of flights, canceled flights, and diverted flights, total number of minutes delayed by delay type, total distance travelled, and total revenue by month. Then, we created a multivariate regression with the revenue as the target variable and the rest of the data set as the input variables in order determine if revenue was correlated to these variables. The model yielded a coefficient of determination of 0.946, proving that there was indeed a correlation between the provided monthly data and monthly revenue. This model is shown in the image below in Figure 3.

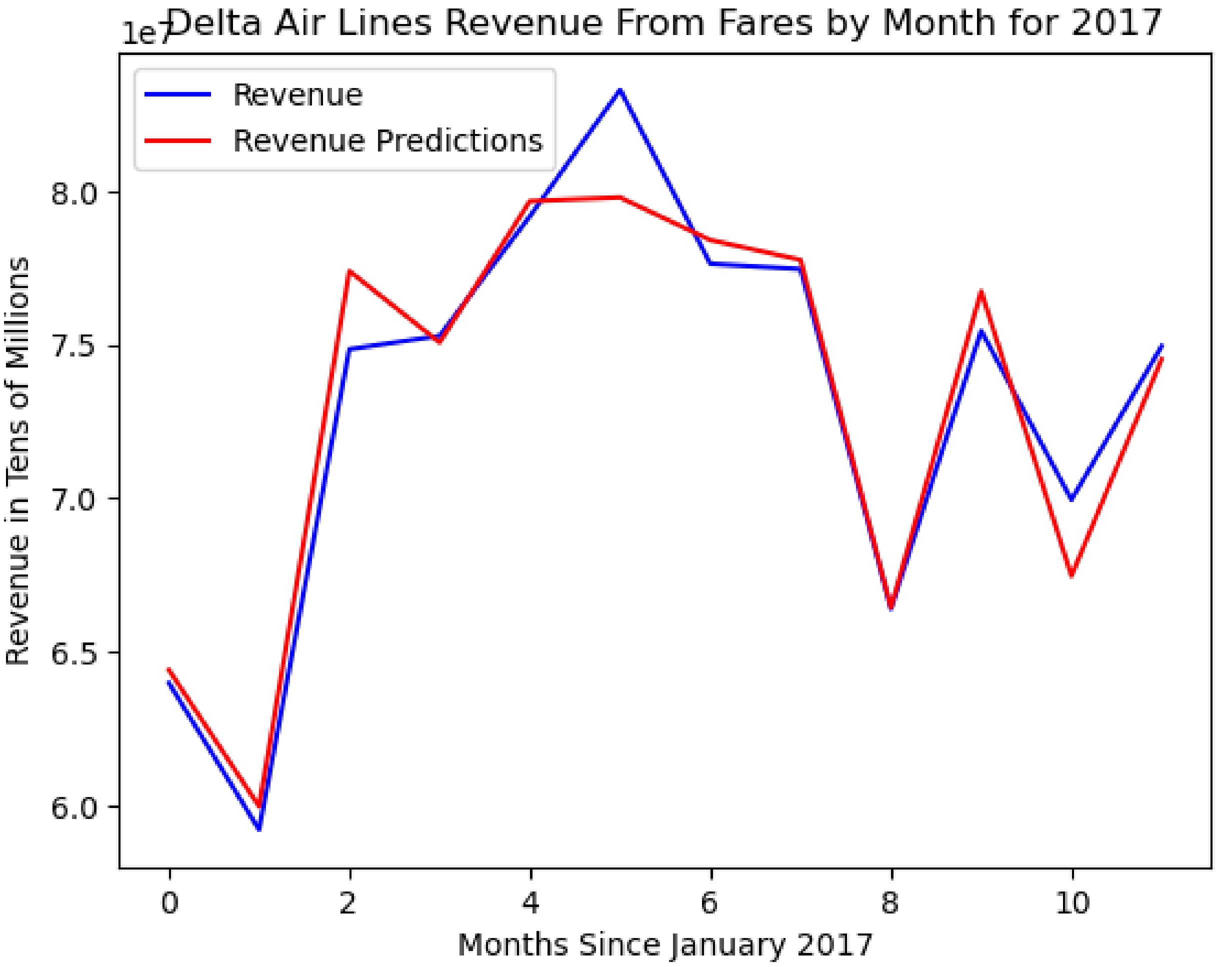

Figure 3: Graph of predicted vs. actual revenue by month for 2017

Given the limited time frame of the data provided, we looked for our own data sets and looked at the Delta stock price over a longer period of time to see if we could determine its value. We chose the time period of 2010 to 2019 because it was after the 2008 financial crisis and before the Covid-19 pandemic, so the large shocks that resulted from these events could be ignored in our analysis. This time frame allows us to have a statistically significant sample size, with 120 data points for monthly data and 40 data points for quarterly data. Delta's stock price has large changes whenever the company releases financial earning information, which occurs after every quarterly earnings release. These quarterly releases include information such as Revenue Passenger Miles (RPM), Passenger Revenue per Available Seat Miles (PRASM), and Cost per Available Seat Mile (CASM). The rest of the fluctuations in Delta's stock prices are more heavily governed by other market factors. Therefore, we collected additional data sets that we hypothesized could be strong predictors of Delta's stock before they release their quarterly earnings, to predict the value of the stock immediately after traders —and the public—receive new financial information. This would be a ready extension from our data analysis with the given 2017 data sets to a stronger analysis of Delta's large stock price fluctuations.

2.3 Data Acquisition

We collected the following data:

1. Monthly Available Seat Miles (ASM) for Delta Air Lines from the U.S. Department of Transportation's Bureau of Transportation Statistics
2. Monthly Revenue Passenger Miles (RPM) for Delta Air Lines from the U.S. Department of Transportation's Bureau of Transportation Statistics
3. Monthly Passenger Volume for Delta Air Lines from U.S. Department of Transportation's Bureau of Transportation Statistics
4. Monthly Average Airline Fares (CPI) in U.S. Cities (not seasonally adjusted) from the U.S. Bureau of Labor Statistics
5. Monthly Delta Air Lines Google Search Activity from Google Trends
6. Monthly Kerosene-Type Jet Fuel Prices: U.S. Gulf Coast in Dollars per Gallon (not seasonally adjusted) from Federal Reserve Economic Data
7. Monthly Number of Total Employees in the Trade, Transportation, and Utilities Industry (not seasonally adjusted) from Federal Reserve Economic Data
8. Monthly Average Hourly Earnings of All Employees in the Trade, Transportation, and Utilities Industry (not seasonally adjusted) from Federal Reserve Economic Data
9. Monthly Producer Price Index (PPI) for Scheduled Passenger Air Transportation (not seasonally adjusted) from Federal Reserve Economic Data
10. Quarterly American Express Pre-Delta-Air-Lines-Earnings Release Stock Closing Prices from Yahoo! Finance
11. Quarterly Delta Air Lines Passenger Revenue per Available Seat Mile (PRASM) and Cost per Available Seat Mile (CASM) from Delta SEC Filings
12. Quarterly Delta Air Lines Pre-Earnings Release Stock Closing Prices from Yahoo! Finance
13. Quarterly Delta Air Lines Post-Earnings Release Stock Closing Prices from Yahoo! Finance

All of this data was taken from the above respective websites, exported as in CSV or Excel format, and then loaded into Python to be cleaned and analyzed. The only exception is the stock data, where we first manually found the dates where Delta released their quarterly earnings, since these dates were a few days to a month before the release of SEC filings, this process could not be automated without access to more information. Using these dates, we then took the stock values on that day for the Delta Post-Earnings Stock Closing Prices data set and the prior trading day for the American Express and Delta Pre-Earnings Stock Closing Prices data sets through Yahoo Finance's historical stock data search feature.

The first three metrics, ASM, RPM, and Passenger Volume, give a measure of the amount of people using Delta Air Lines, which we would expect to drive Delta stock, as more people flying Delta would increase Delta's revenues. In the same vein, Google search activity for Delta would be a predictor of the number of people flying Delta Air Lines. Coupled with these factors, the monthly average air fare would give the cash revenue that Delta gains from passenger flights, since a higher average ticket price causes a higher revenue for the same amount of miles flown.

The price of jet fuel was chosen as a feature because it is a major indicator of Delta's operating costs. We then estimated Delta's labor costs from the total number of employees in the sector (as a metric to capture employee seasonality), the average hourly wage of these employees, and the Producer Price Index for scheduled passenger air transportation.

We also collected data on the value of American Express stock (AXP) the day before Delta reported quarterly earnings. We believe that this information is useful for predicting the value of Delta stock the next day because a large portion of the value of an airline is held in their loyalty program. Delta has a market capitalization of 29.42 billion USD (Google Finance, 2023), while the value of Delta's loyalty program is around 20 billion USD (Point Loyalty, 2023). Therefore, a large portion of Delta's value is tied to the value of their SkyMiles frequent flyer program. This is also supported by the fact that many large airlines use their frequent flyer programs as collateral when getting loans during the Covid-19 Pandemic. Thus, Delta effectively acts as a depositor to American Express on the value that is stored in their loyalty program. Delta partners with American Express as the credit card for these loyalty miles, and as a result, the value of their holdings in SkyMiles increases and decreases based upon the returns they get from holding that value with American Express. Thus, as American Express stock prices increase, we expect Delta to have a return on the value it has stored in SkyMiles.

We also collected quarterly data on Delta's Passenger Revenue per Available Seat Mile (PRASM) and Cost per Available Seat Mile (CASM), which are industry metrics that are correlated with the success of an airline and, by extension, its stock value. We later demonstrate that these quarterly metrics can be predicted by the aforementioned monthly metrics, which leads us to believe that these monthly metrics would be a good predictor of Delta's stock.

Finally, collected data on Delta's stock price the day before and the day of Delta earnings report releases. By having the stock price the day before, we have a benchmark to determine the value of Delta's stock combined with the other features. The stock price the day of earnings that we collected acts as the true value that we are trying to correlate to and ultimately predict.

2.4 Data Wrangling

To compose our data, we first took the aforementioned supplementary data sets and filtered each of these data sets to the desired dates. We then converted primitive representations of dates to universal `datetime` values and merged these data sets into one single data set of data from 2010 to 2019 segmented by month. Next, because we wanted to segment by quarter and not by month, we grouped the data into chunks of three months (to represent one quarter), taking averages of the monthly values. We then merged this modified data set with quarterly PRASM and CASM, pre-Delta-earnings AXP closing price, pre-earnings DAL closing price, and post-earnings DAL closing price data to produce our final quarterly data set.

2.5 Stock Prediction Model

We now discuss how we built our stock prediction model. We first report the relationship between monthly data and the quarterly airline industry metrics of PRASM and CASM. We then demonstrate a direct correlation between all of our collected monthly data and Delta's stock price. We use this correlation to build a prediction model of stock price, by splitting our data into training and test sets and using time-series cross-validation. We conclude by distilling our correlation into causation through causal inference to determine which factors most heavily affect Delta's quarterly stock price and determining our final feature set.

2.5.1 Monthly Data Correlation to Quarterly PRASM and CASM

We began by running two regressions. We used the following data as input variables, with the monthly variables transformed to averages of the three months within the quarter: passenger volume, RPM, ASM, Average Airfare Costs (CPI), Google Search Activity, American Express's stock price before Delta's earnings are released, and Delta's pre-earnings stock price. We ran these input variables into a multivariate regression with the Quarterly PRASM as the target variable. Similarly, we used the following data as input variables to a multivariate regression with Quarterly CASM as the target variable: passenger volume, Producer Price Index for Airlines, Jet Fuel Prices, Hourly Earnings, Number of Employees.

The multivariate regression for PRASM created the following predictions depicted in Figure 4, which is graphed below against the true Quarterly PRASM values. As we can see from the graph, our model lines up very well with the actual PRASM, capturing the seasonal rise and dips in PRASM as well as nonseasonal variations, as the concavity of the prediction over multiple quarters lines up with the concavity of the actual PRASM. This visual inspection is supported by a R^2 value of 0.844, which means that we have a strong correlation between our variables and PRASM.

Similarly, the multivariate regression for CASM created the following predictions depicted in Figure 5 for CASM, which are graphed below with the true quarterly CASM values. As we can see from the graph, our model from input data lines up well with the actual CASM, with the exception of one high outlier in the actual CASM value near the center of the graph. Thus, the CASM prediction lines up well with the true CASM, capturing both seasonal variations and long term trends in concavity. Our visual intuition is supported by a R^2 value of 0.698, which means that we have a moderately strong correlation between our variables and CASM.

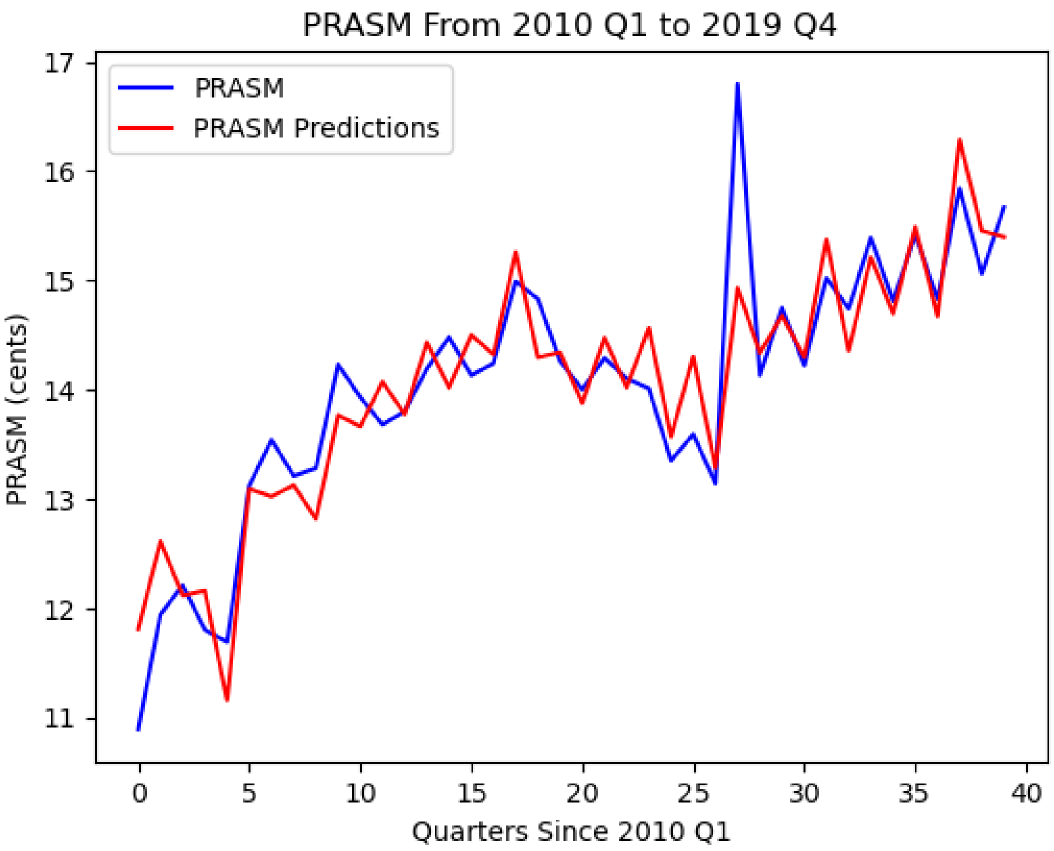

Figure 4: Graph of predicted vs. actual PRASM by quarter from 2010 to 2019

Therefore, our input variables are able to effectively predict PRASM and CASM. Since PRASM and CASM together essentially give the airline's profit per available seat mile, we now know that our data can be used to find a relationship with the profit and thus the stock price of Delta Air Lines.

2.5.2 Monthly Data Correlation to Stock Price

Now that we know our input variables are reasonable measures of an airline's profits and by extension, performance, we can begin to construct a correlation between our variables and Delta's stock price itself. As previously discussed, we want to find relationships for Delta's stock the day Delta releases its quarterly earnings reports because high volume and volatility in Delta stock occur during these days.

Since we are trying to predict the quarterly average stock price, we bin all our monthly data into quarterly data as done in the correlation to PRASM and CASM and leave the stock data as is. With this quarterly data, we run a regression between our quarterly inputs and the closing price of Delta's stock the day earnings are released. We graph the model's predictions for Delta's stock price below with Delta's actual stock price.

Our prediction is extremely close to the actual stock price, which implies that our data are capable of capturing seasonal variability, long-term trends, and data spikes (such as the downward spike in the middle of the graph where our prediction almost exactly coincides with the real value). Furthermore, we have an R^2 of 0.994, which means there is a very strong correlation between our features and Delta's quarterly stock price.

2.5.3 Monthly Data Prediction of Stock Price

Until now, we have only demonstrated correlations between our data and Delta's stock price. We now use our data to predict Delta's stock price using a model trained on prior data, not fit to the entire data set. We maintain the same inputs and outputs as before, but we now use time-series cross-validation, which entails successively enlarging the training set

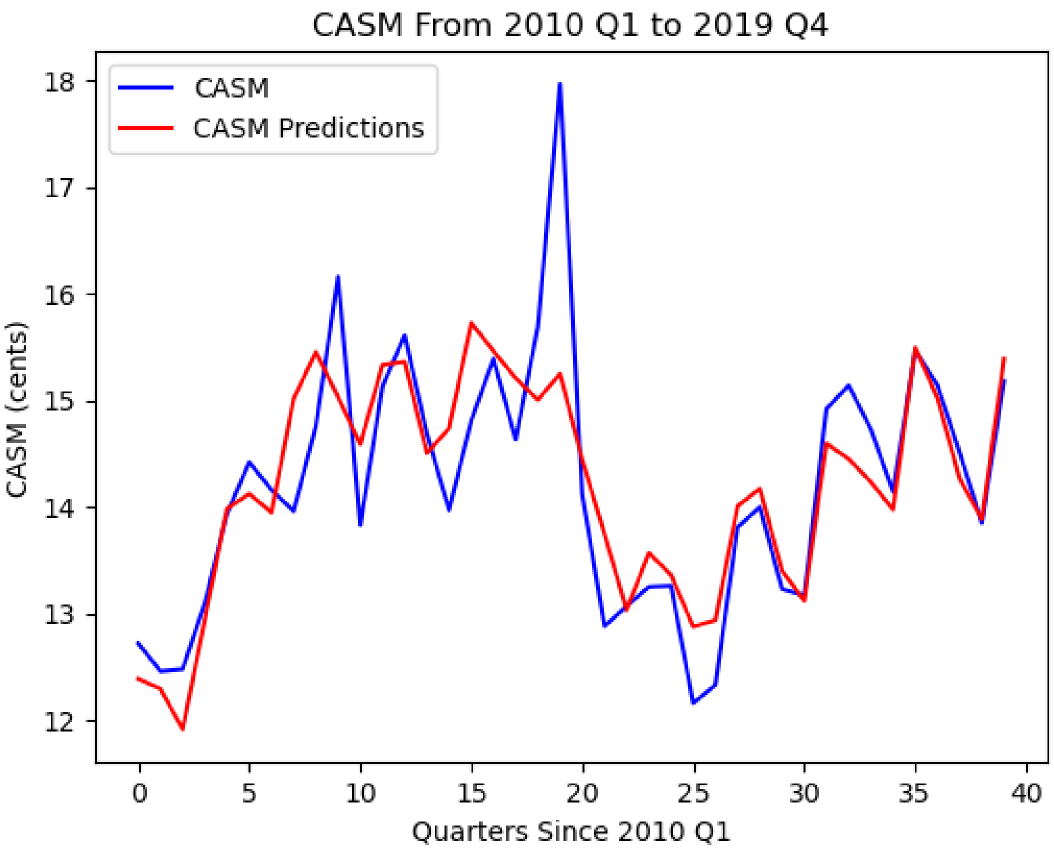

Figure 5: Graph of predicted vs. actual CASM by quarter from 2010 to 2019

of prior data and predicting a set number of future quarterly stock prices. We begin by taking the first 25 quarters of information and predict the next 3 quarters of stock price. We choose to predict for three quarters ahead because R^2 could not be easily and effectively captured for only a one or two quarter prediction set and we used R^2 as our metric to compare predictions. We then take these next three quarters and add it to our training set and repeat to predict the following three quarters of stock price. We repeat this process until we have predicted the final three quarters of 2019 (the end of our data set). We picked the hyperparameter of three quarters forward prediction because it allowed us to have a prediction that was far enough ahead without compromising our strength of prediction.

In our prediction model, we elected to use Elastic Net Regularization over a simple multivariate regression. We do this to simultaneously account for the L_1 and L_2 norm penalties. This makes our model better than a Lasso prediction because Lasso tends to select one term from a group of highly correlated variables, which many of our inputs are, and ignores the others. Our Elastic Net contains a quadratic penalty from the Ridge regression to overcome this. However, our Elastic Net performs better than just a Ridge regression because it also uses a shrinkage similar to a Lasso regression, which the Ridge regression lacks.

Graphs of the actual and predicted Quarterly Delta Stock Price are shown in Figures 7 through 11 for each of our training-test combinations. The fold number represents the current iteration of the Elastic Net Regularization that we are lopping through:

The R^2 values of these graphs are 0.674, 0.755, 0.846, 0.929, and 0.620 respectively. This implies that as we get more info in our test set, we achieve a better prediction of what the stock price is going to be, reaching R^2 values in the 0.9 range. In line with the R^2 values, the MAE of each of the predictions are: 2.57, 2.04, 0.952, 0.876, and 2.26 USD respectively. Additionally, for the prediction with the highest R^2 value, the MAE is 0.876, so our stock price predictor can become accurate to less than a dollar error for a stock that is worth between 50 to 60 dollars, which means that we can be accurate to within 2 percent error. The exception would be the last graph that has an R^2 of 0.620. After

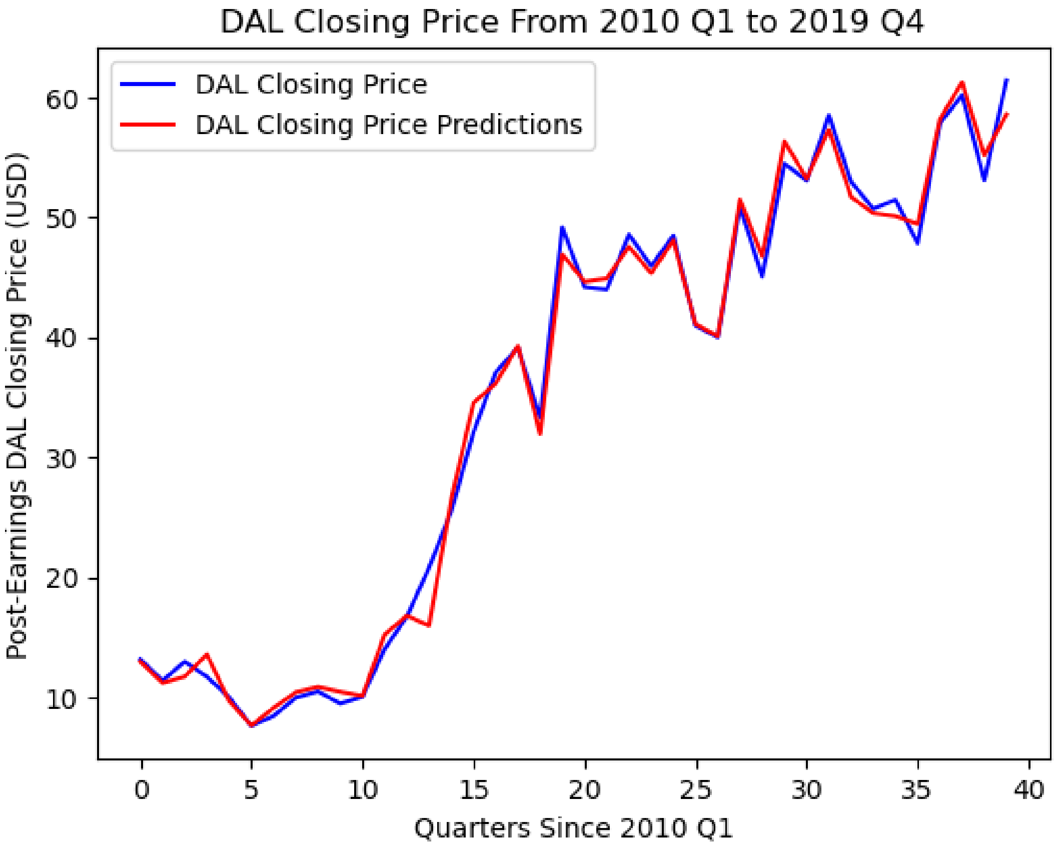

Figure 6: Graph of predicted and actual Delta stock (DAL) closing price from 2010 to 2019

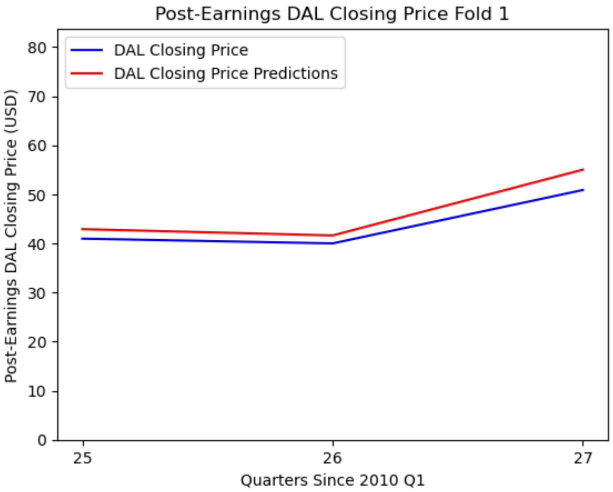

Figure 7: Predictions for 2016 Q2 - 2016 Q4

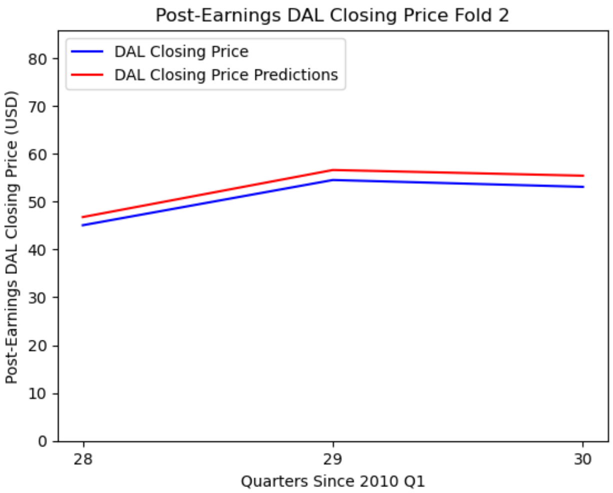

Figure 8: Predictions for 2017 Q1 - 2017 Q3

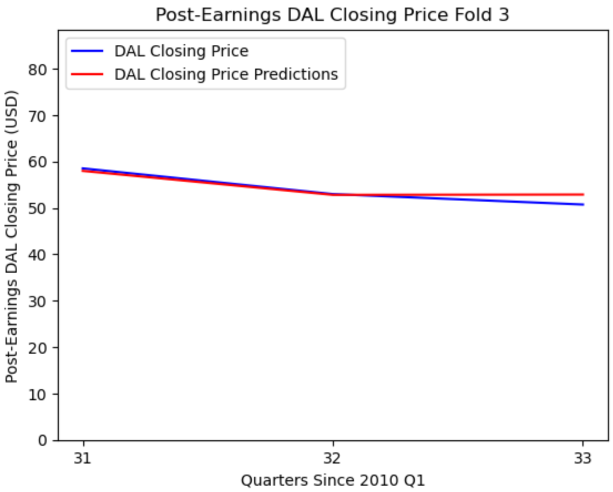

Figure 9: Predictions for 2017 Q4 - 2018 Q2

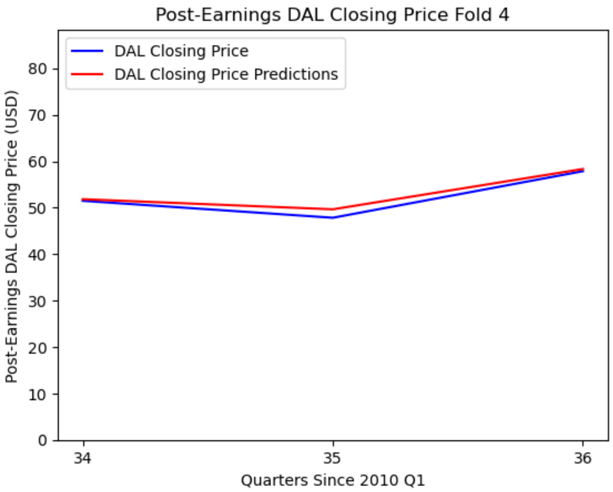

Figure 10: Predictions for 2018 Q3 - 2019 Q1

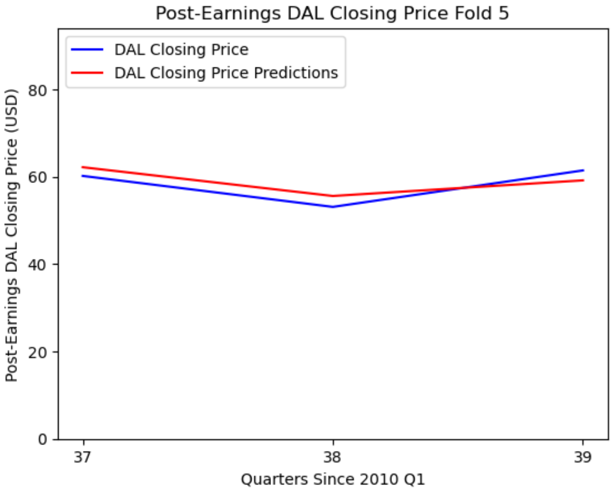

Figure 11: Predictions for 2019 Q2 - 2019 Q4

investigating this time period there appeared to have been unexpected growth in the airline industry that was not easily predicted; therefore, our model in the future will need more updating to better capture anomalous stock shifts.

As a whole, our model prediction is able to capture Delta's quarterly stock price effectively after the company releases its quarterly reports. This means that we can use our model to predict Delta's stock price change and act on that information to either sell or buy Delta stocks right before Delta releases their quarterly reports.

2.5.4 Monthly Data Causation of Stock Price and Feature Selection

We have many colinear inputs to our correlation and prediction models; therefore, we want to complete our analysis by parsing down our inputs in our correlation to just the inputs that are causally related to Delta's quarterly stock price. We create a set of inputs that are causally related to Delta's quarterly stock price by finding a subset of the inputs that minimizes the Akaike Information Criterion (AIC) for the multivariate regression that we discussed in Section 2.52. The AIC is a useful metric for time series analysis and small data sets and creates a metric to find the optimal balance between goodness of fit of the model and simplicity of the model.

Additionally, we chose to prioritize coefficient significance and AIC over Variance Inflation Factor (VIF) for each input in determining what to include in the final model. Reduction via VIF and AIC resulted in nearly the same set of inputs with the exception of VIF dropping the Delta Stock Price on the trading day before earnings were reported. We decided to use the AIC result, which included this input, because its coefficient was statistically significant. Thus, through all of the sub-lists of inputs into the multivariate regression, we picked the set of inputs that had the lowest AIC value. This set of inputs was: Available Seat Miles (ASM), the Previous value of Delta's stock, Consumer Price Index (CPI) for air travel, the value of American Express Stock, and the Price of Jet Fuel. However, ASM had a coefficient of $-2.0 * 10^{-7}$ with a p-value of .059, which is above the standard alpha-value of 0.05. Therefore, we also removed ASM. American Express stock has a p-value of 0.005, Delta's pre-earnings release stock has a p-value of 0.000, and Jet Fuel Price has a p-value of 0.021. All of these p-values are much less than the alpha-value of 0.05, which means that the relationship between these variables and Delta stock price is non-zero. We also kept CPI, even though its p-value of 0.06 is higher than 0.05 because its coefficient is 0.0153, which is significant. This leaves us with a final set of inputs of: the pre-earnings value of Delta's stock, the Average Airfare Consumer Price Index (CPI) for air travel, the value of American Express Stock pre-earnings, and the Price of Jet Fuel.

These four factors also make logical sense. Aviation CPI and American Express's stock capture many of the trends that affect Delta's revenue —Delta's profit from airline ticket sales and Delta's returns from the value held in their loyalty program. The price of jet fuel accounts for the greatest fluctuation in Delta's costs. The last input, Delta's stock price the day before, will account for extraneous market factors at the time of prediction. Therefore, these four factors form a logically sound and statistically significant basis set for inputs.

As a result of our feature selection, we reran the previous prediction model with our new data set and obtained a better level of predictions with only 4 of the original set of 11 input variables. Namely, we were able to get almost the exact same prediction metric, which means that our feature selection was applicable to our prediction model. The five graphs of prediction from selected input set are shown below in Figures 12 through 16. These graphs helps illustrate that the new model has better performance. For example, the predictions for 2016 Q2 to 2016 Q4 are much tighter with our new model than the old, and this is reflected through an increase in R^2 from 0.674 to 0.980.

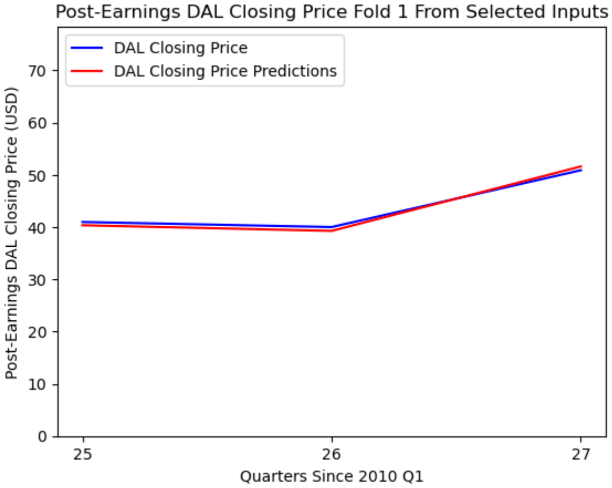

Figure 12: Selected Input Predictions for 2016 Q2 - 2016 Q4

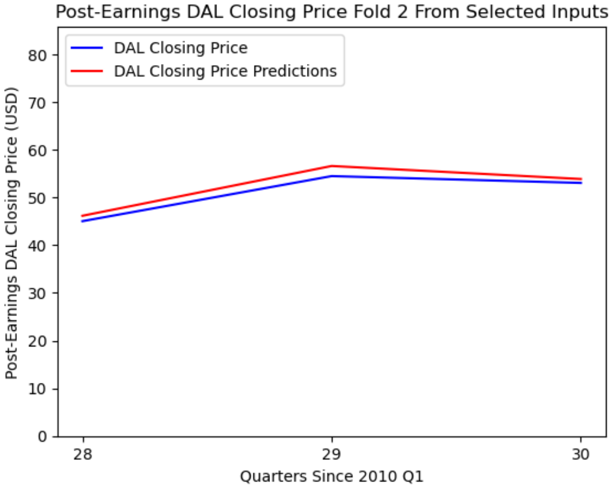

Figure 13: Predictions for 2017 Q1 - 2017 Q3

3 Conclusion

We have created a model that takes in four inputs and creates a regression model for Delta's Quarterly Stock Prices after every earnings report release. This model has an R^2 value of 0.998 and very closely matches the patterns in Delta's stock price. This model is statistically significant and can be used to predict future Delta quarterly stock prices with MAE values of less than a dollar. With better access to data and future research, this model can be extended to include other airlines and be made to predict price on smaller time-scales.

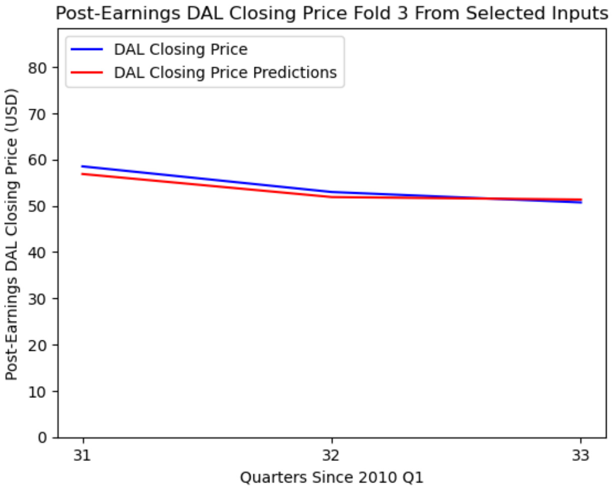

Figure 14: Selected Input Predictions for 2017 Q4 - 2018 Q2

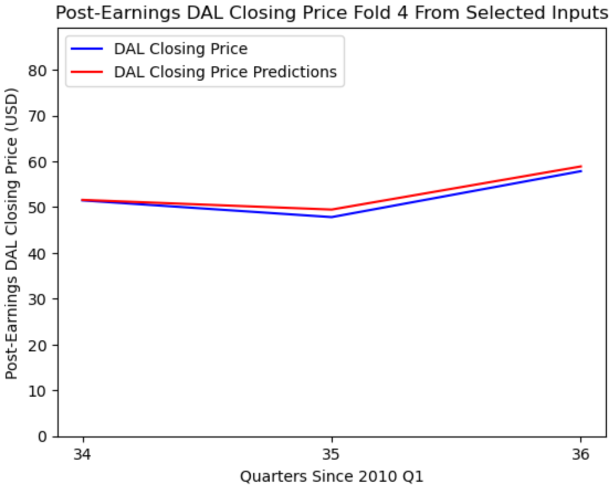

Figure 15: Selected Input Predictions for 2018 Q3 - 2019 Q1

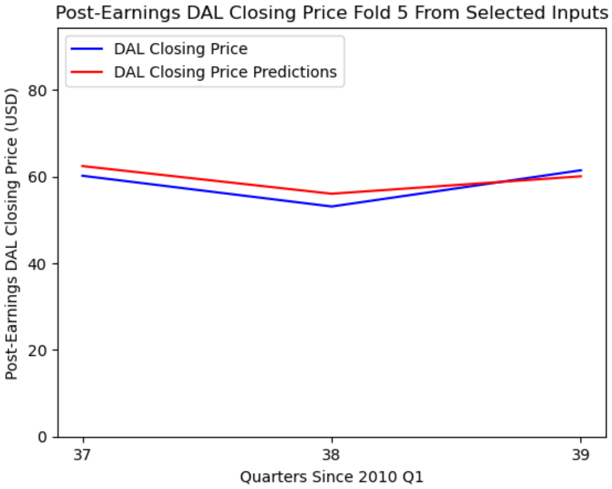

Figure 16: Selected Input Predictions for 2019 Q2 - 2019 Q4

4 External Data Source Citations

The following are the links to the data sources in order of the itemized list in Section 2.3

1. https://www.transtats.bts.gov/Data_Elements.aspx?Qn6n=F
2. https://www.transtats.bts.gov/Data_Elements.aspx?Qn6n=F
3. https://www.transtats.bts.gov/Data_Elements.aspx?Qn6n=F
4. https://data.bls.gov/pdq/SurveyOutputServlet
5. https://trends.google.com/trends/explore?date=2010-01-01
6. https://fred.stlouisfed.org/series/MJFUELUSGULF
7. https://fred.stlouisfed.org/series/CEU4000000001#0
8. https://fred.stlouisfed.org/series/CEU4000000003
9. https://fred.stlouisfed.org/series/PCU481111481111
10. https://finance.yahoo.com/quote/AXP
11. https://ir.delta.com/financials/default.aspx#sec
12. https://finance.yahoo.com/quote/DAL
13. https://finance.yahoo.com/quote/DAL

Data sources from 2017 discussed in the exploratory data analysis section were provided by Citadel.