Introduction

Coffee (arábica variety) is the second largest traded commodity worldwide, with about $100 billion in volume traded annually. Coffee futures are standardized, exchange-traded contracts in which the contract buyer agrees to take delivery, from the seller, a specific quantity of coffee (eg. 10 tonnes) at a predetermined price on a future delivery date. Coffee futures are traded on the New York Stock Exchange (NYSE) from 9:30 a.m. - 1:30 p.m. daily.

The goal of this project is to forecast daily closing prices of coffee futures using machine learning algorithms. This research is meaningful since there is no known research of this type to date that predicts *daily* coffee futures prices with a low error. This research could be used by anyone affected by coffee futures markets and prices.

Hypotheses

1. "Opening Price", "High Price", "Low Price" and "Closing Price" are related variables for coffee commodities.
2. Historical data of daily coffee futures commodities can be used to more accurately predict the future daily "Closing Price."

Related Work

Montague uses neural networks, random forest, linear ridge regression and gradient-boost decision tree to predict 27 commodity futures using 3800 days of high, open, low and closing prices with the highest model goodness of fit of 0.713. Abdullah uses back propagation neural networks and decision tree to predict crude oil prices with a minimum error of 0.035. Kim uses artificial neural networks to predict stock prices to predict several commodities using 2,348 days of data with the highest accuracy of 68.97%. Ticklavilca uses Bayesian regression to forecast commodity prices of corn and hogs with 21 years worth of monthly data and a minimum error of 4.8%. Shahwan use artificial neural networks from 535 daily prices to forecast future prices of hog and canola oil with a minimum error of 0.0219.

Data & Methods

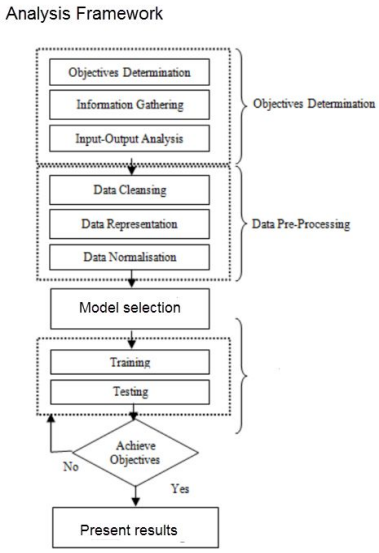
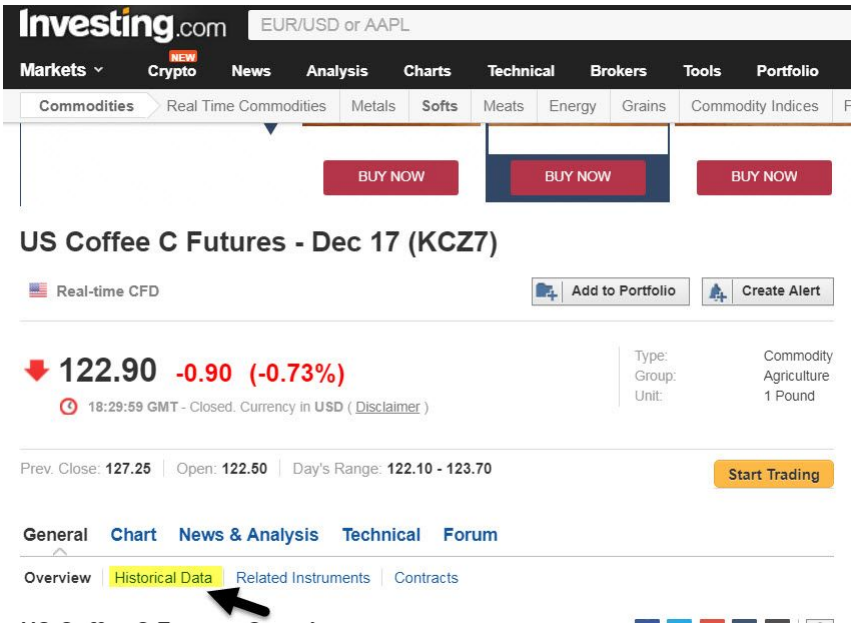


Figure 1 presents an overview of the data science analysis framework used in the analysis from beginning to end.

The objective of trying to predict daily coffee futures closing price using historical data and machine learning was established. Information was gathered including learning what coffee futures prices mean and possible factors that affect the prices. Since the data is publically available and not proprietary in any way, the data is not considered private. The reuse for this research does not violate any data ethics principles.

During the Input-Output Analysis phase, it was determined that coffee production variables such as temperature and rain would not be included as inputs for the models. Although these variables probably affect coffee futures prices, since no subject matter was available to model the exact quantitative relationship of these variables, they were excluded to prevent drawing any false conclusions. The final data set included four input features ("Opening Price", "High Price" and "Low Price") and one target output feature ("Closing Price"). Historical data from January 1, 2010 - November 15, 2017 of coffee futures was manually obtained from the "Historical Data" section of the investing.com website.

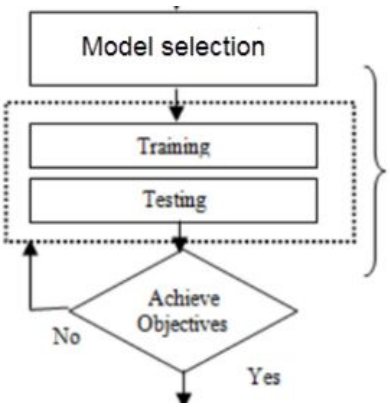


Closing Price, Open, High and Low Prices are measured throughout each trading day, Monday through Friday for coffee futures (units in 1000 60-kg bags of green cafe arabica beans). There are about 252 trading days on average per year in the New York Stock Exchange where coffee futures are traded. These futures are traded 390 minutes daily, from 0930-1330 EST.

The first part of the data acquisition step included creating a CSV file with all of the variables from the website. The data contains 2,805 daily observations for each of the four variables from January 1, 2010 - November 15, 2017. The prices are in US dollars and are measured to two decimal places. For the data cleansing step, the date column was reformatted and the column was renamed from "Price" to a more descriptive "Closing" to prevent confusion. Since the data was already in numerical format, there was no need to represent the data with numbers or normalize it as described in the analysis framework above.

Some exploratory data analysis including a correlation matrix was performed to see if and how the variables were correlated (see Appendix). This visualization confirms that the variables are highly positively correlated to each other, ranging from 0.998 to 0.9992. A highly positively correlated set of variables means that as "Open" price increases, so will the "High" and "Low" prices. The distribution of data was also plotted visually using a histogram and kernel density estimate. The variables had close enough values that the histograms and kernel density estimates were very similar to each other.

Machine Learning Model Selection



Initially, a linear regression, decision tree regression, and AdaBoost models were chosen since these models are transparent and easy to interpret. Furthermore, since the problem is a regression rather than classification or clustering problem, these models were a good choice for the analysis. A regular decision tree model would not work since there was continuous numerical data. After using these three models and seeing their performance, the Scikit-learn machine learning cheat sheet was used to choose one more model. Since there were more than 50 but less than 100,000 samples, there was no category to predict, the objective was to predict a quantity for the Closing Price and it was unknown if the features were important, a Ridge Regression model was chosen.

An 80/20 training / test data set was created from the pre-processed CSV file and each of the models was fitted. There were over 2,800 observations from this time period included in the analysis.

Results & Insights

For the first part of the results discussion, I look at the R^2 values for each model to see how well the model fits. The assumption is that, with a high R^2 value, the model is expected to predict well for data observed in the future.

Each of the models had a very high goodness of fit, R^2 value. Linear Regression had an R^2 value of 0.996, Decision Tree Regression = 0.799, Adaboost = 0.749 and Ridge Regression = 0.996. The best fitting model were Linear Regression and Ridge Regression. Now that each classifier showed a very goodness of fit, the predicted values were calculated.

The predicted values for Linear Regression, Decision Tree Regression, Adaboost and Ridge Regression were calculated and combined into one dataframe.

	Date	Closing	Open	High	Low	LR Predicted	DT Regr Predicted	AdaBoost Predicted	Ridge Predicted
0	15-Nov-17	126.80	126.85	127.53	126.28	125.946209	126.80	126.471500	125.946272
1	14-Nov-17	127.05	127.55	129.00	124.25	127.454415	127.05	126.474839	127.454415
2	13-Nov-17	127.60	127.50	128.15	126.65	126.553729	127.60	126.471500	126.553795
3	10-Nov-17	127.55	126.05	128.35	125.50	126.883475	127.55	126.471500	126.883403
4	9-Nov-17	126.40	125.55	127.40	125.00	125.912623	126.40	125.955581	125.912587
5	8-Nov-17	125.75	125.05	125.90	124.20	124.356366	125.75	123.800876	124.356412
6	7-Nov-17	124.70	125.50	125.80	123.10	124.213658	124.70	123.800876	124.213751
7	6-Nov-17	125.55	123.95	126.40	122.25	124.977213	125.55	123.800876	124.977126
8	3-Nov-17	123.95	127.15	127.40	126.25	125.783661	123.95	126.471500	125.783760
9	2-Nov-17	126.40	123.35	127.60	122.55	126.302820	126.40	125.596842	126.302583
10	1-Nov-17	122.95	124.70	126.20	121.20	124.703887	122.95	123.800876	124.703880

Next, prediction error was calculated for each algorithm with the following equation:

Prediction error = [(measured value - predicted value) / (measured value)] * 100.
'Closing' is the measured value in the combined data frame and is the daily Closing Price.

The mean percent prediction errors were: Linear Regression = 0.00009736, Decision Tree Regression = 0.00000542, AdaBoost = 0.00328, and Ridge Regression = 0.00009738. Since all of these classifiers had under 0.01% prediction error, these are results that could be implemented by customers trading coffee futures. The worse prediction error from this research is six times more accurate than existing research.

Algorithm Type	**Goodness of Fit**	**Prediction Error**
Linear Regression	0.996	0.00009736
Decision Tree Regression	0.799	0.00000542
Decision Tree with AdaBoost	0.749	0.00328
Ridge Regression	0.996	0.00009738

The hypothesis that “Opening Price”, “High Price”, “Low Price” are related variables to the “Closing Price” can be accepted. The hypothesis that historical data of “Opening Price”, “High Price” and “Low Price” can be used to predict the daily “Closing Price” was also verified. The objectives were achieved and the results were presented.

A sampling of how many times any of the “Opening”, “High”, “Low” or “Closing” prices were the same as one another on the same day. In the year 2010, this duplication of prices happened 47 out of 270 trading days, or about 17.4% of the time.

Future Work
Recommendations include implementing the linear regression, AdaBoost or ridge regression algorithms to predict daily coffee futures closing prices. Making this static process into a scalable process would allow businesses to predict daily closing Price of coffee futures in real-time. Both the Yahoo Finance and Google Finance APIs are no longer available.The Alpha Advantage API seems to be the only open source API for getting real-time stock prices.

If needed by the business, the granularity of the analysis could be increased to predict the price every minute rather than every day. Since futures prices theoretically can change each and every minute of a trading day, an analysis with all of these minute-by-minute changes and machine learning prediction could be done. This would potentially add an additional 98,280 (252 trading days * 390 minutes per trading day) price data points to the analysis.

References

Abdullah, Siti. Machine Learning Approach for Crude Oil Price Prediction. Doctor of Philosophy Thesis, 2013. University of Manchester. Accessed 25 October 2017.

Alpha Vantage API. https://www.alphavantage.co/documentation/. Accessed 29 November 2017.

Coffee C Futures. Intercontinental Exchange. https://www.theice.com/products/15/Coffee-C-Futures. Accessed 27 November 2017.

Economics of coffee. Wikipedia: https://en.wikipedia.org/wiki/Economics_of_coffee. Accessed 3 October 2017.

Kim, K. and W. Lee. Stock market prediction using artificial neural networks with optimal feature transformation. Neural Computing and Applications (2004), 13: 225-260.

Montage, D. Algorithmic Trading of Futures via Machine Learning. Stanford University CS229 Course material. http://cs229.stanford.edu/proj2014/David%20Montague,%20Algorithmic%20Trading%20of%20Futures%20via%20Machine%20Learning.pdf. Accessed 11 September 2017.

Prediction Accuracy. Stat897D, Applied Data Mining and Statistical Learning. Penn State course. https://onlinecourses.science.psu.edu/stat857/node/160. Accessed 29 November 2017.

Shahwan T., Odening M. (2007) Forecasting Agricultural Commodity Prices using Hybrid Neural Networks. In: Chen SH., Wang P.P., Kuo TW. (eds) Computational Intelligence in Economics and Finance. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-72821-4_3. Accessed 12 October 2017.

Seven Things You Must Know about Coffee Futures. https://tradingsim.com/blog/7-things-you-must-know-about-coffee-futures/. Accessed 12 September 2017.

Ticlavilca, A. M., Dillon M. Feuz and Mac McKee. 2010. “Forecasting Agricultural Commodity Prices Using Multivariate Bayesian Machine Learning Regression.” Proceedings of the NCCC-134 Conference on Applied Commodity Price Analysis, Forecasting, and Market Risk Management. St. Louis, MO. [http://www.farmdoc.illinois.edu/nccc134].

Trading Day. Wikipedia. https://en.wikipedia.org/wiki/Trading_day. Accessed 27 November 2017.

Appendix A - Tables and Figures

Source Data
Summary of Results

Table 1 *Summary Statistics*

	Closing	Open	High	Low
count	2805.000000	2805.000000	2805.000000	2805.000000
mean	169.241907	169.447355	171.432980	167.379430
std	48.736788	48.758244	49.428461	48.087293
min	101.500000	101.500000	103.750000	100.950000
25%	131.550000	131.800000	133.000000	130.050000
50%	153.700000	154.150000	156.400000	151.850000
75%	196.700000	197.400000	200.650000	193.700000
max	304.900000	305.300000	306.250000	304.000000

Figure 1 *Box Plot*

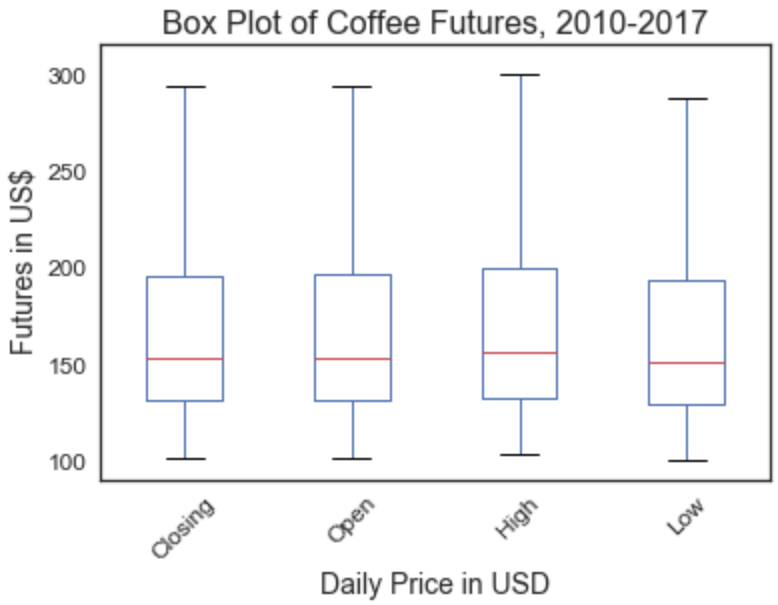


Figure 2 *Futures (all variables) Line Plot*

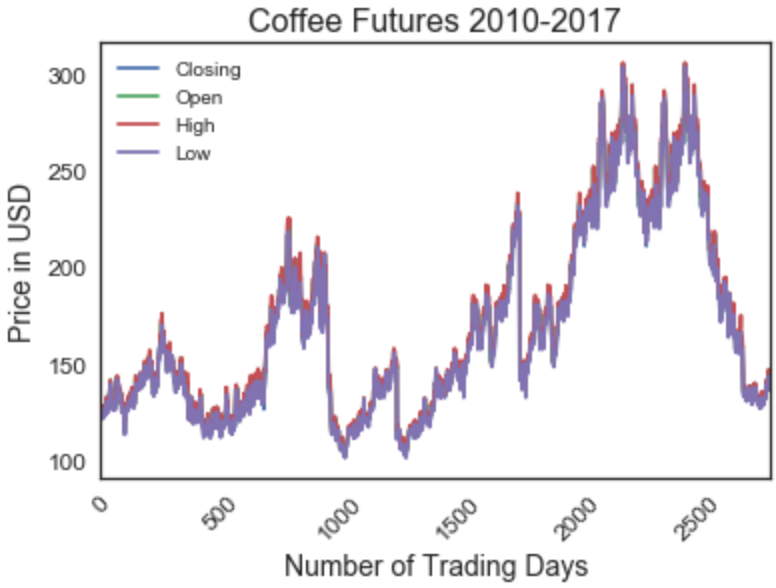


Figure 3 *Closing Price Line Plot*

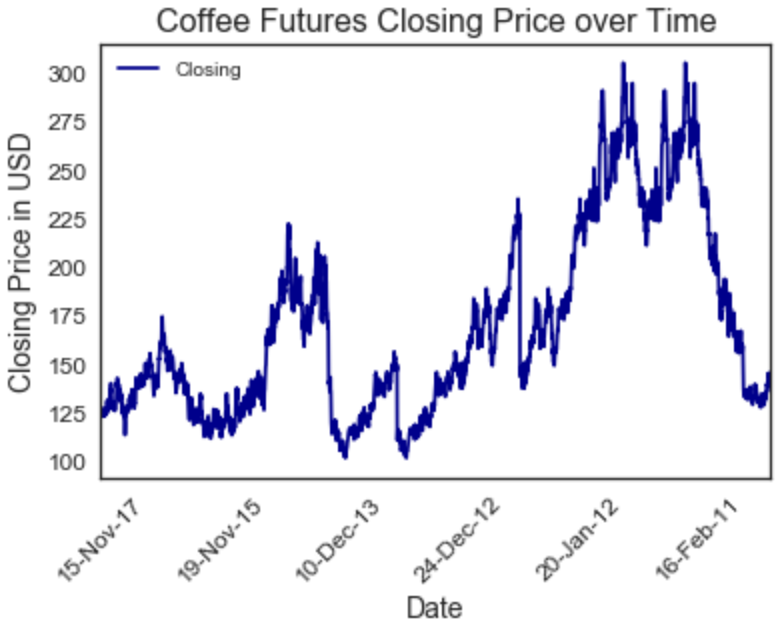


Figure 4 *Histogram / Kernel Density Estimate for Closing Price*

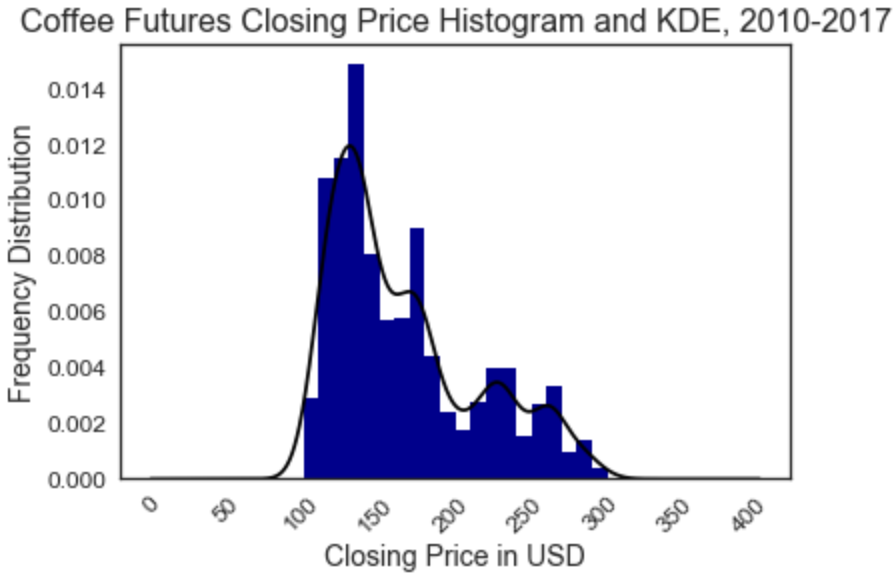


Figure 5 *Histogram / Kernel Density Estimate for Opening Price*

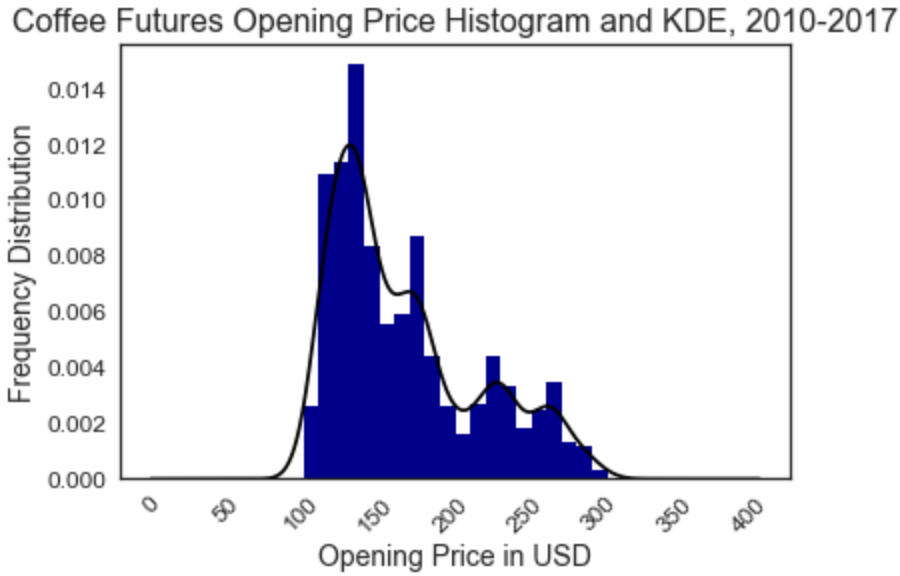


Figure 6 *Histogram / Kernel Density Estimate for High Price*

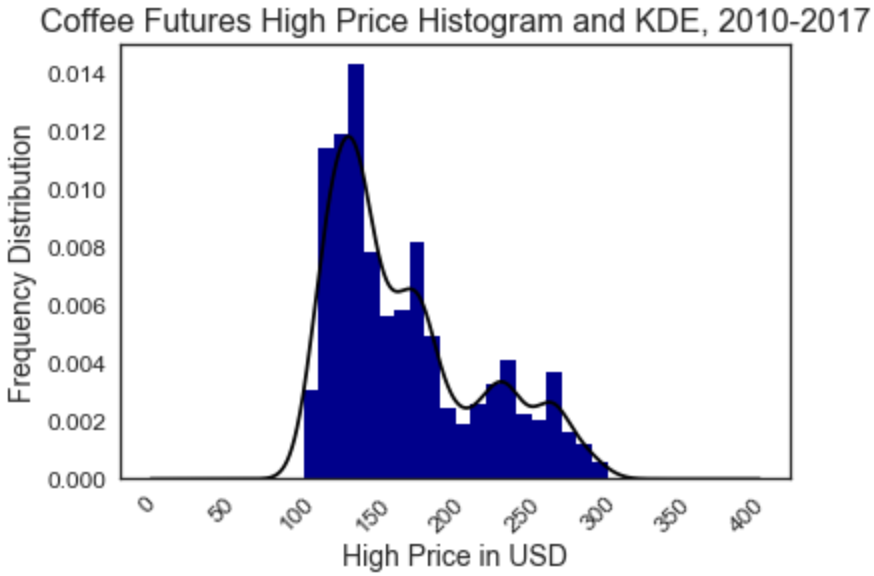


Figure 7 *Histogram / Kernel Density Estimate for Low Price*

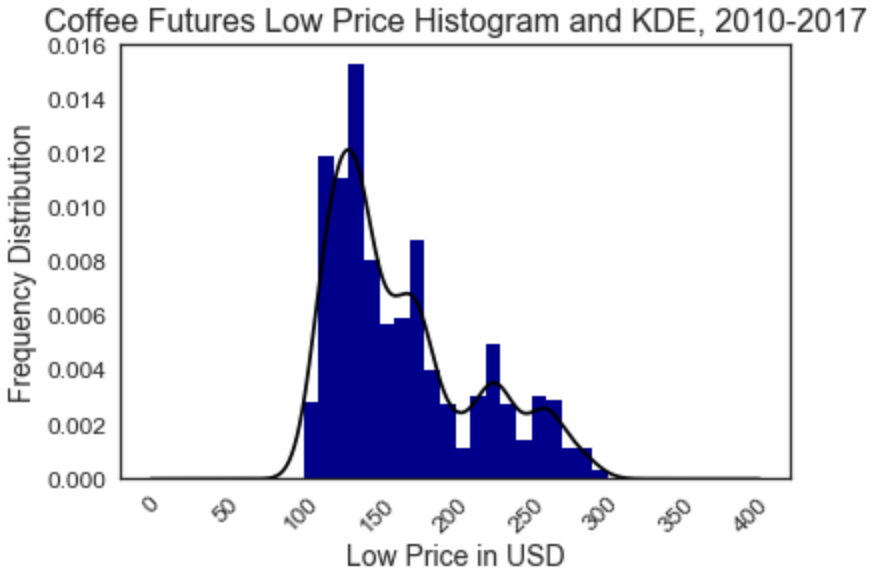


Figure 8 Correlation Matrix

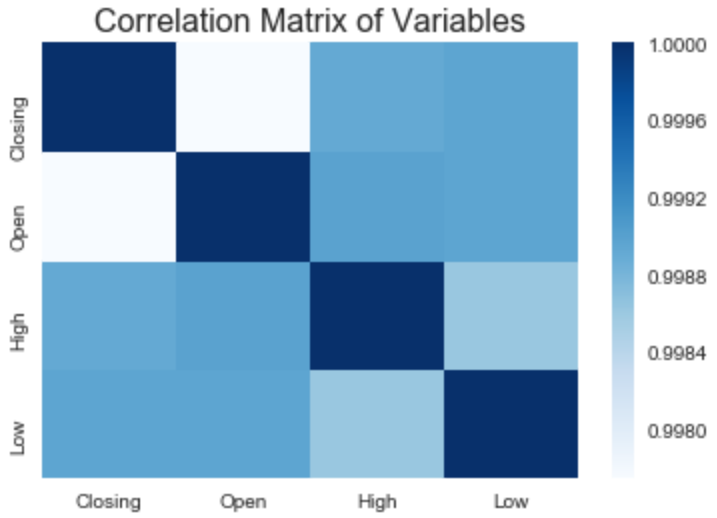


Figure 9 Linear Regression Plot

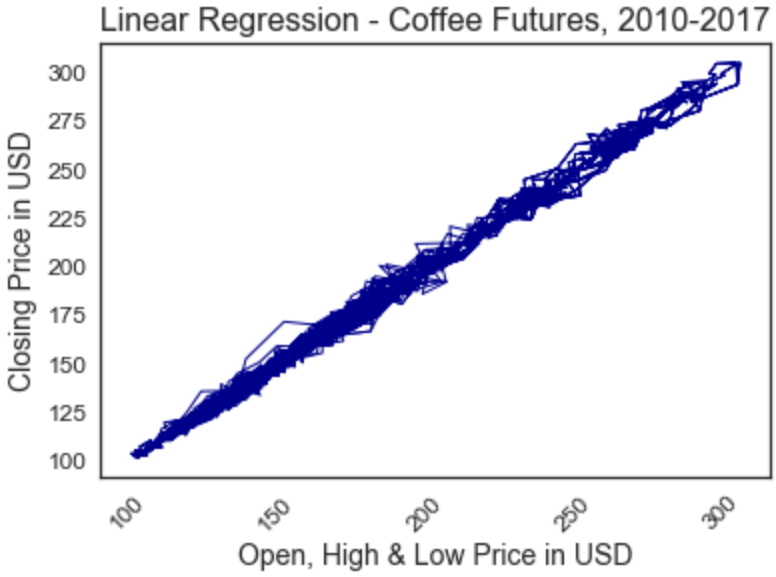


Figure 10 Decision Tree Regression Plot

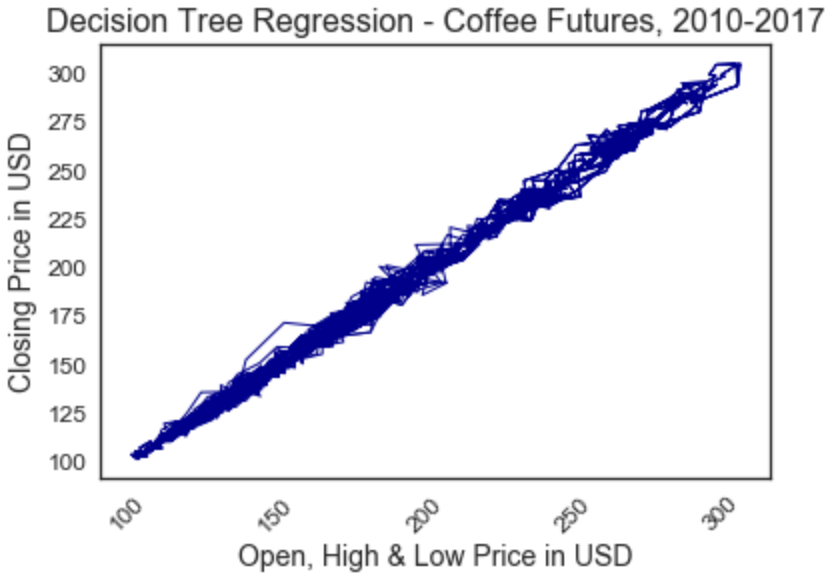


Figure 11 Decision Tree AdaBoost Plot

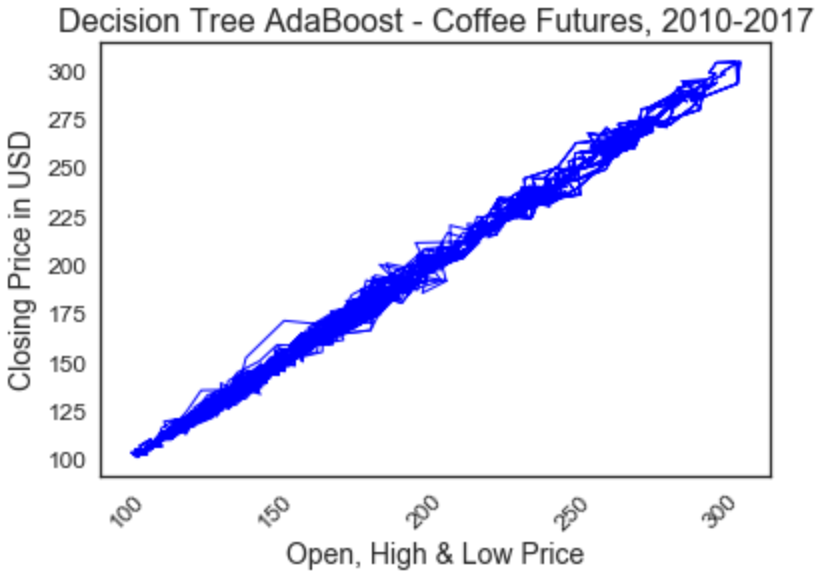


Figure 12 Ridge Regression Plot

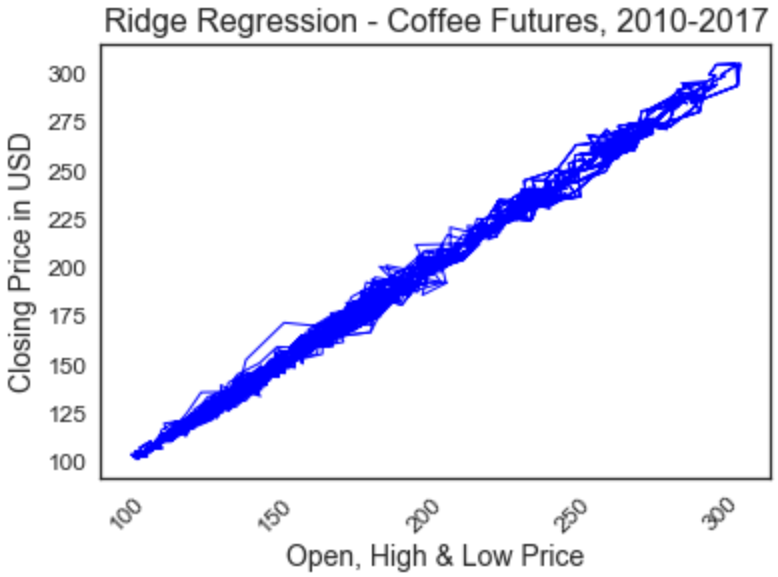


Figure 13 Algorithm Goodness of Fit

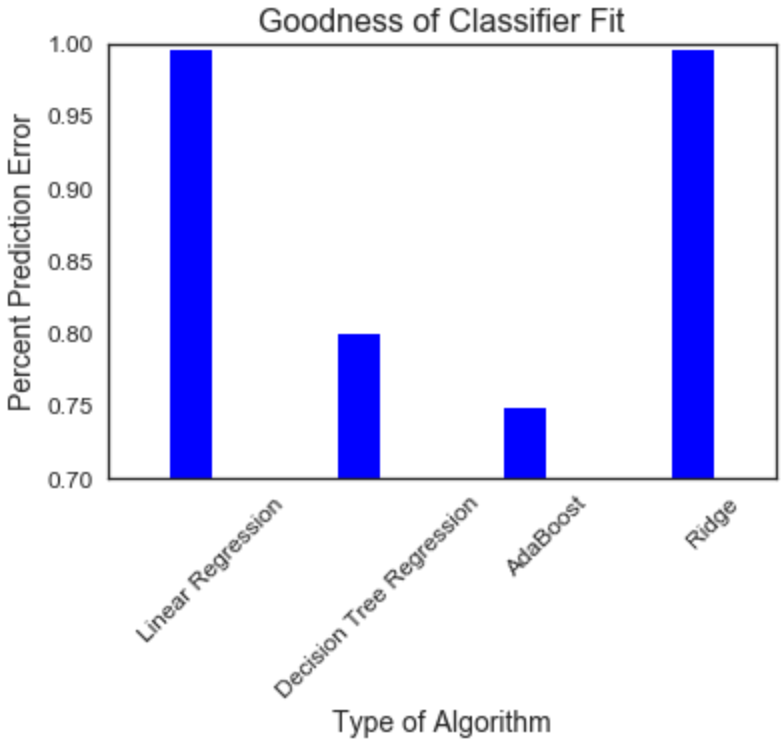


Figure 14 Classifier Percent Prediction Error

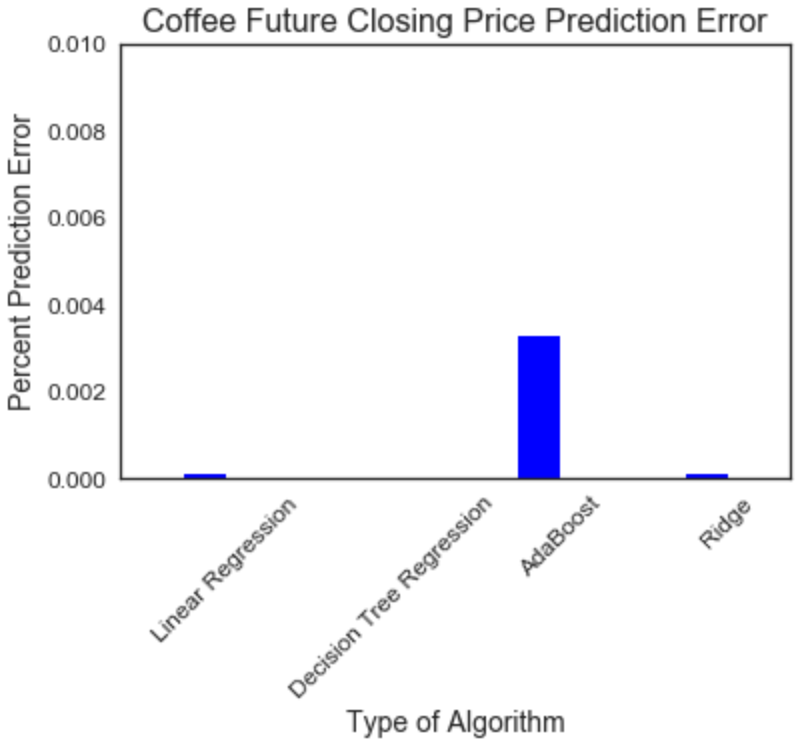


Table 2 Summary of Results

Algorithm Type	**Goodness of Fit**	**Prediction Error**
Linear Regression	0.996	0.00009736
Decision Tree Regression	0.799	0.00000542
AdaBoost	0.749	0.00328
Ridge Regression	0.996	0.00009738