Algorithmic Trading of Coffee Futures with Machine Learning

Laura H. Kahn
School of Informatics and Computing
Indiana University
Bloomington, IN USA
lkahn@indiana.edu

***Abstract*—Data science techniques are applied to financial commodities prices. Machine learning algorithms are used to predict daily closing price of coffee futures with a maximum percent prediction error of 0.00328%.**

Keywords—commodity futures; machine learning; coffee trading

I. INTRODUCTION

Coffee (arábica variety) is the second largest traded commodity worldwide, with about $100 billion in volume traded annually [4, 10]. Coffee futures are standardized, exchange-traded contracts in which the contract buyer agrees to take delivery, from the seller, a specific quantity of coffee at a predetermined price on a future delivery date. Coffee futures are traded on average 252 days on the New York Stock Exchange from 9:30 a.m. - 1:30 p.m. daily [3, 12].

The research will attempt 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 prediction error using data science principles. This research could be used to increase profits for anyone trading coffee futures.

The *hypotheses* are that*:*

a. Open, high and low prices are related to closing price,
b. Historical data of daily coffee futures commodities can be used to more accurately predict future daily closing price.

A. 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 [6]. Abdullah uses back propagation neural networks and decision tree to predict crude oil prices with a minimum error of 0.035 [1]. 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% [5]. 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% [11]. 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 [8].

II. DATA

A. Acquisition

The final data set includes four input features (“Opening Price”, “High Price” and “Low Price”) and one target output feature (“Closing Price”). Daily historical price data of Closing, Open, High and Low from January 1, 2010 - November 15, 2017 of coffee futures was obtained from the “Historical Data” section of the investing.com website in US dollars [13].

B. Pre-processing

Creating a CSV file with all of the variables from the website was the first step in the data acquisition process. The data contains 2,805 daily observations for each of the four variables from January 1, 2010 - November 15, 2017 measured to two decimal places. Minimal data cleaning was needed including formatting the date column and renaming the “Price” column to a more descriptive “Closing.” Since the data was already in numerical format, there was no need to represent the data with numbers or normalize it.

C. Exploratory Analysis

Some exploratory data analysis included creating a correlation matrix 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, etc.. 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 almost exactly the same.

III. Methods

Figure 1 presents an overview of the data science analysis framework method used from start to finish.

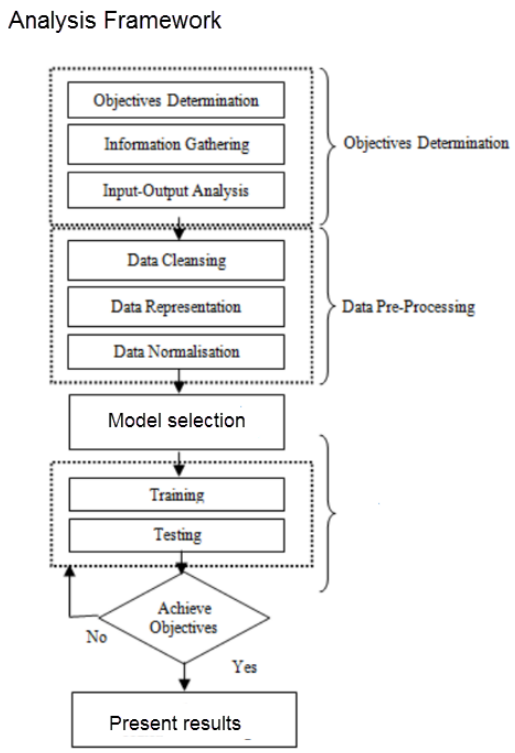

Fig. 1. Data Analysis Framework

The objective as mentioned in the Introduction section was to predict daily coffee futures closing price using historical data and machine learning. Information was gathered including learning what coffee futures prices mean and which 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 may 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

Initially, a Linear regression, Decision tree regression, and Decision tree regression with 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 [9]. 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 added to the algorithms chosen.

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

IV. Results

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.

For the first part of the results discussion, I look at the R^2 values (a.k.a. - goodness of fit 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, R2 value. Linear Regression had an R^2 value of 0.996, Decision Tree Regression = 0.799, Decision Tree with 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 each regression algorithm were calculated and combined into one dataframe as shown in Table 1.

	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	126.718889	125.946272
1	14-Nov-17	127.05	127.55	129.00	124.25	127.454415	127.05	127.254545	127.454415
2	13-Nov-17	127.60	127.50	128.15	126.65	126.553729	127.60	126.718889	126.553795
3	10-Nov-17	127.55	126.05	128.35	125.50	126.883475	127.55	126.209661	126.883403
4	9-Nov-17	126.40	125.55	127.40	125.00	125.912623	126.40	125.969611	125.912587
5	8-Nov-17	125.75	125.05	125.90	124.20	124.356366	125.75	124.382836	124.356412
6	7-Nov-17	124.70	125.50	125.80	123.10	124.213658	124.70	124.514756	124.213751
7	6-Nov-17	125.55	123.95	126.40	122.25	124.977213	125.55	124.310816	124.977126
8	3-Nov-17	123.95	127.15	127.40	126.25	125.783661	123.95	126.718889	125.783760
9	2-Nov-17	126.40	123.35	127.60	122.55	126.302820	126.40	124.716800	126.302583
10	1-Nov-17	122.95	124.70	126.20	121.20	124.703887	122.95	124.514756	124.703880

TABLE I. PREDICTED CLOSING PRICES

Next, prediction error (%) was calculated for each algorithm with the following equation [7]:

$$\text{Prediction error} = \frac{\text{(measured value - predicted value)}}{\text{measured value}} * 100$$

where '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, Decision Tree Regression with AdaBoost = 0.00328, and Ridge Regression = 0.00009738.

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

TABLE II. GOODNESS OF FIT AND PREDICTION ERROR

V. DISCUSSION

Since all of the regression algorithms had less than 0.004% prediction error, these are results that could be implemented by anyone trading coffee futures. The worse prediction error (0.00328) from this research is more than six times more accurate than existing research (0.0219). Since the models are such good fits and the data is highly correlated, this would explain a very low prediction error percent. The algorithms sufficiently meet the objective of predicting the daily closing price of coffee futures with open, high and low price inputs.

The hypothesis that “Opening Price”, “High Price”, “Low Price” are related to the “Closing Price” can be accepted. The second hypothesis that historical price data can be used to predict the daily “Closing Price” can also be accepted. The research objectives were achieved.

VI. CONCLUSION

Recommendations include implementing one or more of these 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 near real-time. The Alpha Advantage API seems to be the only supported open source API for getting real-time stock prices [2].

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

ACKNOWLEDGMENT

The author would like to thank Dr. Joanne Luciano and JT Wolohan for assistance guiding the scope of the research.

REFERENCES

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

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

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

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

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

[6] 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.

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

[8] 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.

[9] Choosing the Right Estimator. Scikit learn. http://scikit-learn.org/stable/tutorial/machine_learning_map/index.html. Accessed 1 December 2017.

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

[11] 11. 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. Accessed 2 December 2017.

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

[13] US Coffee C Futures (KCH8). https://www.investing.com/commodities/us-coffee-c. Accessed 12 October 2017.

APPENDIX

Source Data

	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

TABLE III. SUMMARY STATISTICS

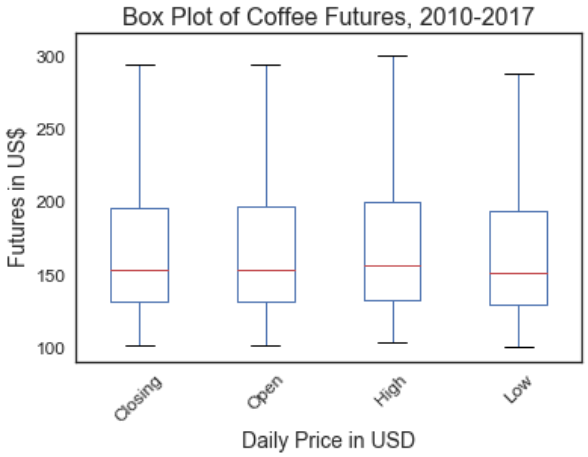

Figure 3. Box Plot of Coffee *Futures, 2010-2017*

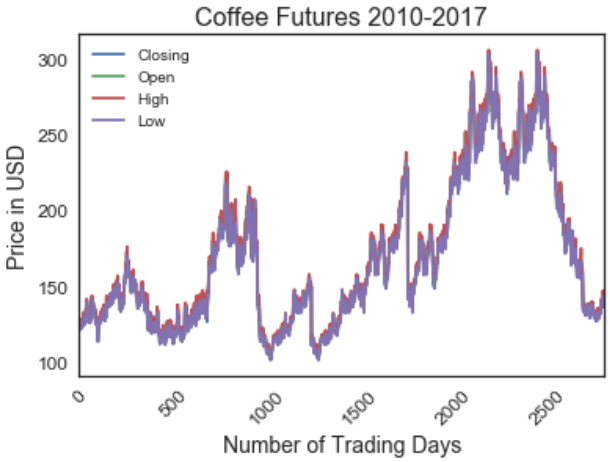

Figure 4. Line Plot of Coffee Futures, 2010-2017

Summary of Results

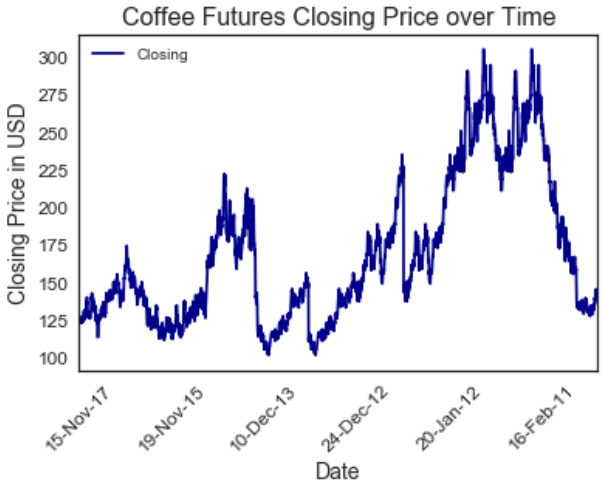

Figure 5. Line Plot of Closing Price over Time

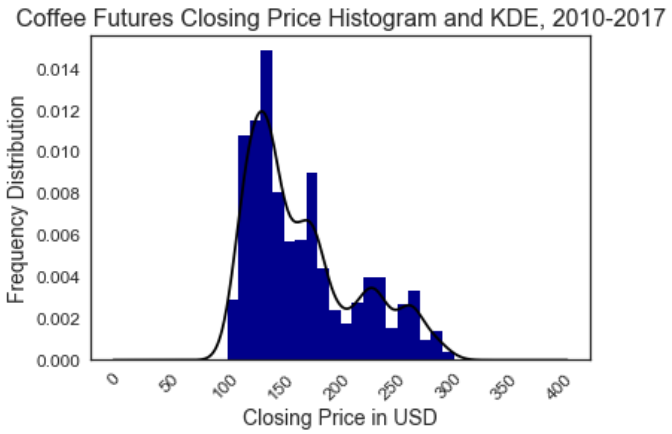

Figure 6. Histogram and Kernel Density Estimate of Closing

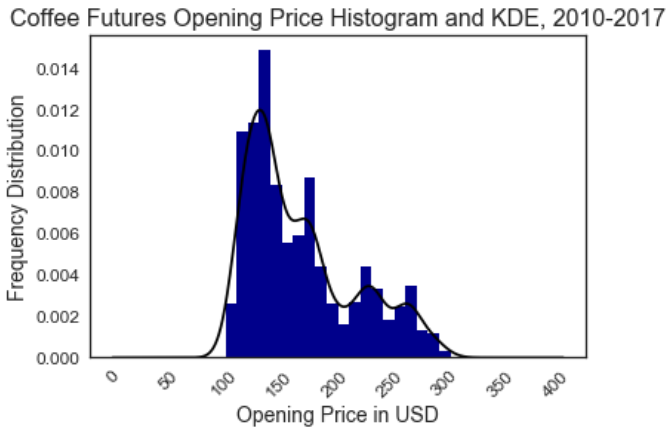

Figure 7. Histogram and Kernel Density Estimate of Open

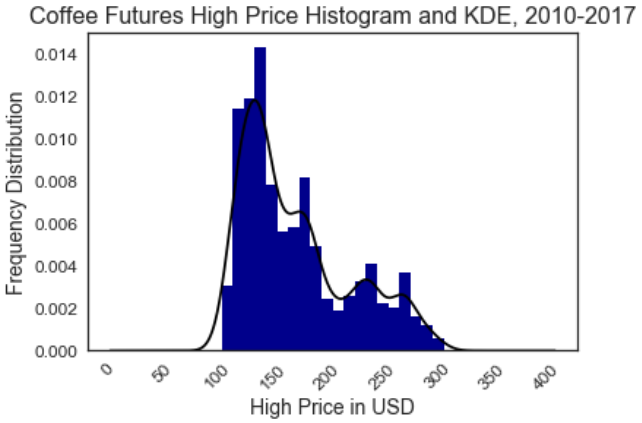

Figure 8. Histogram and Kernel Density Estimate of High

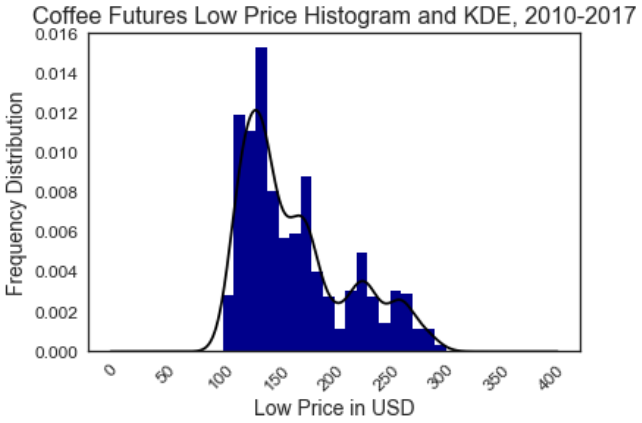

Figure 9. Histogram and Kernel Density Estimate of Low

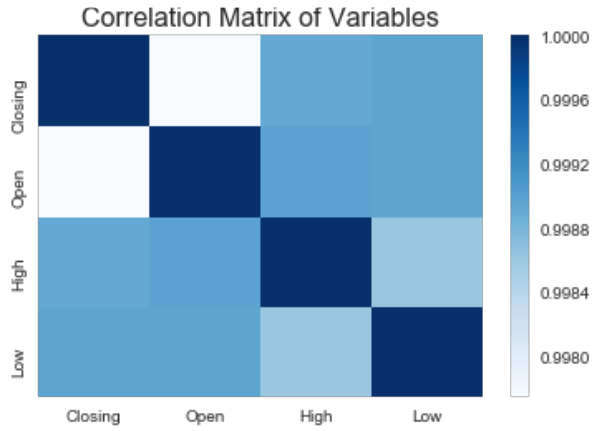

Figure 10. Correlation Matrix

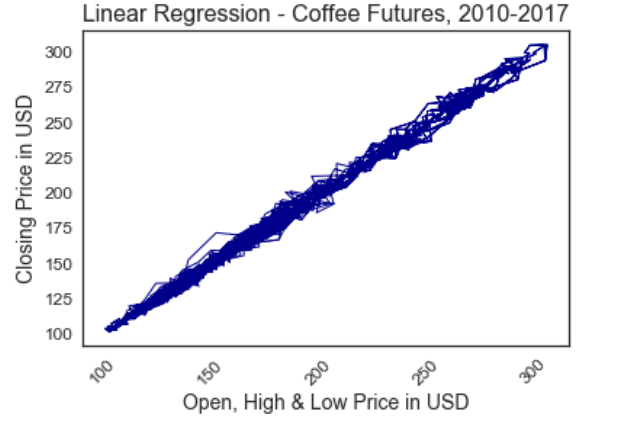

Figure 11. Linear Regression Plot

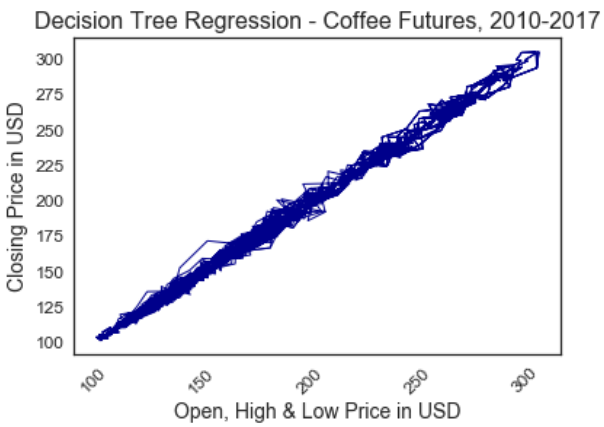

Figure 12. Decision Tree Regression Plot

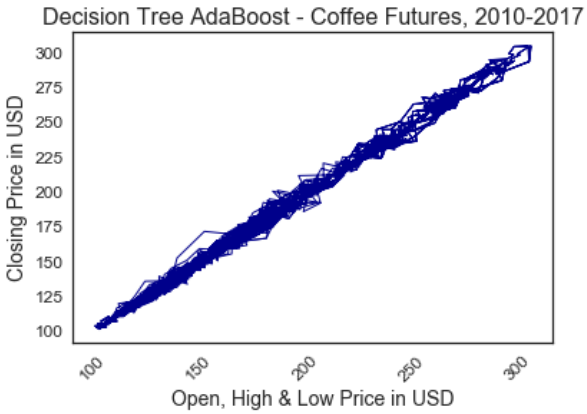

Figure 13. Decision Tree Regression with Adaboost Plot

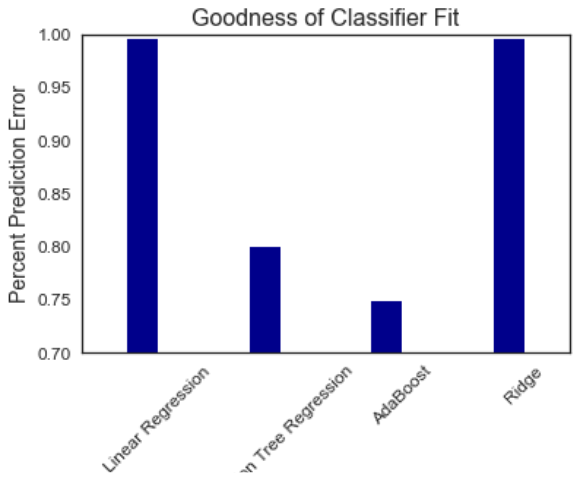

Figure 14. Classifier Performance - Goodness of Fit

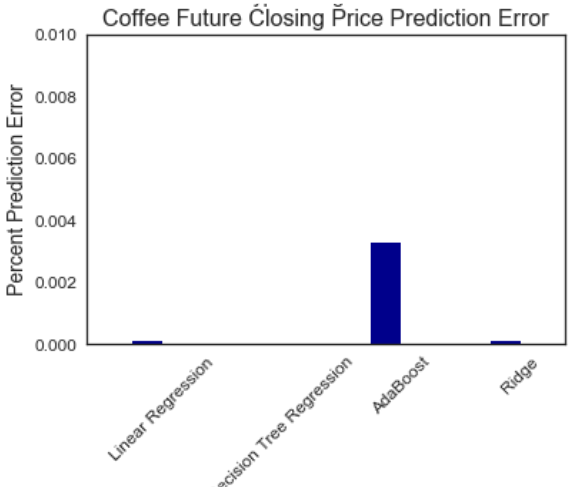

Figure 14. Classifier Performance - Prediction Error (%)