Money Talks

Exploring the U.S Stock Market Pharmaceuticals Sector & Politics

---

Authors: Marcus Barnes, Hidaya Ismail, Faria Khandaker, Asel Kushkeyeva, Ngone Lo, & Shabrina J. Mardevi

Date: December 2, 2020.

Word count: [Final Word Count Here < 3500-4000 words excluding tables, charts, and references.]

Abstract

Are there patterns detectable using fundamental machine learning (ML) methods that hold for the US publicly traded pharmaceutical companies? When, generally, is the best time to invest in such companies? Does the presiding political party (Democrat versus Republican) impact pharmaceutical sector performance and the volume of types of drugs manufactured? Linear regression, KNN regression, Random Forest (RF), SVM, ensemble models, and Auto-AMIRA were applied against US stock market financial indicator data compiled by Carbone (2020), stock market daily sales data by Onyschchak (2020), and pharmaceutical sales data by Zdravkovic (2020), covering the years 2014 through 2018. The stock market was rather stable over this period. Linear regression performed best over this period, with 99.5% of the predicted values falling within the actual 5% of actual values. This model will help when determining the best investment time, but is not able to predict drastic price changes. For the drugs in our dataset, there were no significant changes in the volume of drugs sold between the Obama administration and the Trump administration year. Our results indicate that linear regression techniques provide decent insights into performance of pharmaceutical sector stocks, but only within shorter time frames during which sector price volatility is low. Other, more complex modelling analysis is needed to determine longer term price predictions.

Introduction

Yoo, Kim, and Jan (2005, pg. 835) point out that since the formation of stock markets, “many have attempted to predict the stock markets using various computational tools”. Different business sectors rely on different market factors. For example, insurance companies rely heavily on interest rates and gold company performance is highly dependent on the spot price of gold. Can more trandiational ML approaches still provide insight by looking at a subset of the market, in our case, the pharmaceutical sector? Our aim is to identify patterns using basic ML approaches which can inform an investing strategy that will outperform for the sector (and ideally the market as a whole). Patterns specific to the pharmaceutical sector of the market may not hold for other sectors and the market as a whole. Using such patterns as part of an investment strategy may provide market beating returns.

Specifically, in our study, we ask: are there patterns detectable using fundamental ML methods that hold for the US publicly traded pharmaceutical companies? When, generally, is the best time to invest in such companies? Does the presiding political party (Democrat versus Republican) impact pharmaceutical sector performance and the volume of types of drugs manufactured? Our work provides preliminary insights by running various fundamental ML methods against US stock market financial indicator data compiled by Carbone (2020), stock market daily sales data by Onyschchak (2020), and pharmaceutical sales data by Zdravkovic (2020).

Literature Review

Yoo, Kim, and Jan (2005) provide a survey of ML techniques and use of event information in stock market prediction, but with a focus on neural networks (ANN). Studies such as Patel

et al. (2015) and Yang et al. (2016) use neural networks and ensemble techniques to better understand stocks and markets values.

As the stock data is time-series, there are specific ML methods for time series data such as ARIMA (AutoRegressive Integrated Moving Average). The chapter on time series modeling by Ramasubramanian (2019) provides an induction to time series modeling in R including ARIMA. The paper by Adebiyi, Adewumi, and Ayo (2014) compares ARIMA to ANN and demonstrates the superiority of ANN models for stock price prediction. Both ARIMA and ANN are interesting techniques worth mentioning, but out of scope of the investigation in this report, though Auto-ARIMA will be touched on briefly.

By using regression, SVM, random forests, KNN, kmeans clustering, Naïve Bayes (NB) and other fundamental methods, we are in part exploring which factors may provide insight into investment performance. Arnott et al. in (Arnott, 2019) provides a critique of the factor investing approach, providing warnings about how the approach can break down. A book has been recently published by Coqueret and Guida (2020) specifically dealing with ML techniques for factor investing. Unfortunately, at the time of writing, the authors of this report were not able to secure a copy.

Dataset Description

The main dataset used for the two questions in this project is the *Pharma sales data: Six years data (2014-2019) on sales of drugs classified in 8 ATC categories* by Zdravkovic (2020). The data consists of 2,016 rows where each represents a date (time series) from 2014 to 2019, and 13 columns representing variables such as hour, day, and year (categorical), and daily stock sale values by medication types (numeric). The second dataset, only used in the first question, is the *Stock Market Dataset: Historical daily prices of all stocks and ETFs* by Onyschchak (2020). This dataset is in fact a repository of over 6,000 companies' daily stock prices and volume traded on NASDAQ up to April 2020, each in the separate .csv file. The .csv files of seven pharmaceutical companies were extracted and compiled, which resulted in a dataset of 8805 tuples and 8 variables such as company indicator, prices, and volumes (numeric) subsetted from 2014 to 2018. Finally, we used *200+ Financial Indicators of US Stocks (2014-2018*) by Carbone (2020), exclusively for question 2. The repository consists of 5 .csv files, where each represents a year from 2014 to 2018. The number of tuples representing companies in the datasets range from 3,800 to 5,000, with 224 columns financial indicators to represent sector, revenue, shares, and margins at a yearly unit. For the first question, both Zdravkovic (2020) and Onyschchak (2020) were subsetted to match the 2014 to 2018 study period and merged. Tables 1-3 below will provide a general statistical overview of characteristics of our data and datasets for this project. All datasets were retrieved from kaggle.com.

Table 1. Descriptive Statistics of Numeric Attributes of *Pharma Sales data: Six years data (2014-2019) on sales of drugs classified in 8 ATC*

Variable	1st-3rd Quartiles	Min.	Med.	Mean	Max.	Standard Deviation

Med4RheumArth	3-6.33	0	4.340	4.798	16.680	2.602594
Med4OstArth	2.33-5	0	3.425	3.684	13.340	1.960576
Aspirin	2.1-5.338	0	4	4.016	16	2.421335
Ibuprofen	19-37	0	26.20	28.76	161	13.92392
Med4Tension	5-13	0	9	9.602	54.833	5.917205
Med4Sleep	0-1	0	0	0.6395	9	1.155103
Meds4Asthma	1-7.2960	0	3	5.315	41	6.242541
Meds4Allergy	1-4	0	2	2.653	12	2.256522

Table 2. Descriptive Statistics of Numeric Attributes of *Stock Market Dataset: Historical daily prices of all stocks and ETFs*

Variable	**1st-3rd Quartiles**	**Min.**	**Med.**	**Mean**	**Max.**	**Standard Deviation**
Open	41.93-75.74	27.81	56.53	63.42	147.84	27.25942
High	42.22-76.10	27.97	57.03	63.87	148.99	27.44758
Low	41.66-75.30	27.51	56.05	62.96	147	27.06800
Close	41.94-75.73	27.70	56.51	63.43	148.14	27.26239
Adj_close	36.15-66.74	22.70	48.24	55.44	142.88	25.62536
Volume	2,386,200-10,227,700	228,600	5,646,000	8,459,660	284,468,100	9,877,061

Table 3. Descriptive Statistics of Numeric Attributes of *200+ Financial Indicators of US Stocks (2014-2018*) - Top Attributes

Variable	1st-3rd Quartiles	Min.	Med.	Mean	Max.	Standard Deviation
Long.term.investments	0 - 5.891e+07	0	0	2.931e+09	9.970e+11	34,899,214,793
Tax.Liabilities	0 - 5.392e+07	(1.784e+09)	4.020e+05	2.460e+11	4.938e+15	3.482694e+13
Financing.Cash.Flow	(7.700e+07)-5.686e+07	(8.880e+11)	0	(1.895e+08)	9.980e+11	15,440,104,591
Net.Cash.Marketcap	(0.436) - 0.109	(4943.553)	(0.063)	(0.958)	501.780	43.2072
priceEarningsToGrowthRatio	11.23-28.97	0	17.98	58.41	100419.29	954.0307
Net.Debt.to.EBITDA	(0.454)-3.163	(9616.000)	1.127	2.882	22776.529	236.6003
Graham.Number	0-29	0	12	738	8851966	69208.75
X5Y.Net.Income.Growth..per.Share.	0-0.105	(1)	0	0.057	4.476	0.2042184
Asset.Growth	(0.031)-0.177	(1)	0.048	1.134	5468.426	45.38108

Analytical Approaches

As mentioned in the *Introduction* section, we are using several ML methods to answer the following questions:

1. Given today's stock and pharmaceutical sales data, what is the best stock value to buy tomorrow?
2. a. Are there changes in the types of drugs sold in the years after Trump was elected?
2. b. Additionally, are there general changes in the overall financial standing of the pharmaceutical industry between 2014 and 2018?

In this section, we will lay out the analytical framework, as well as considerations in selecting models.

Research Question 1

"Given today's stock and pharmaceutical sales data, what is the best stock value to buy tomorrow?"

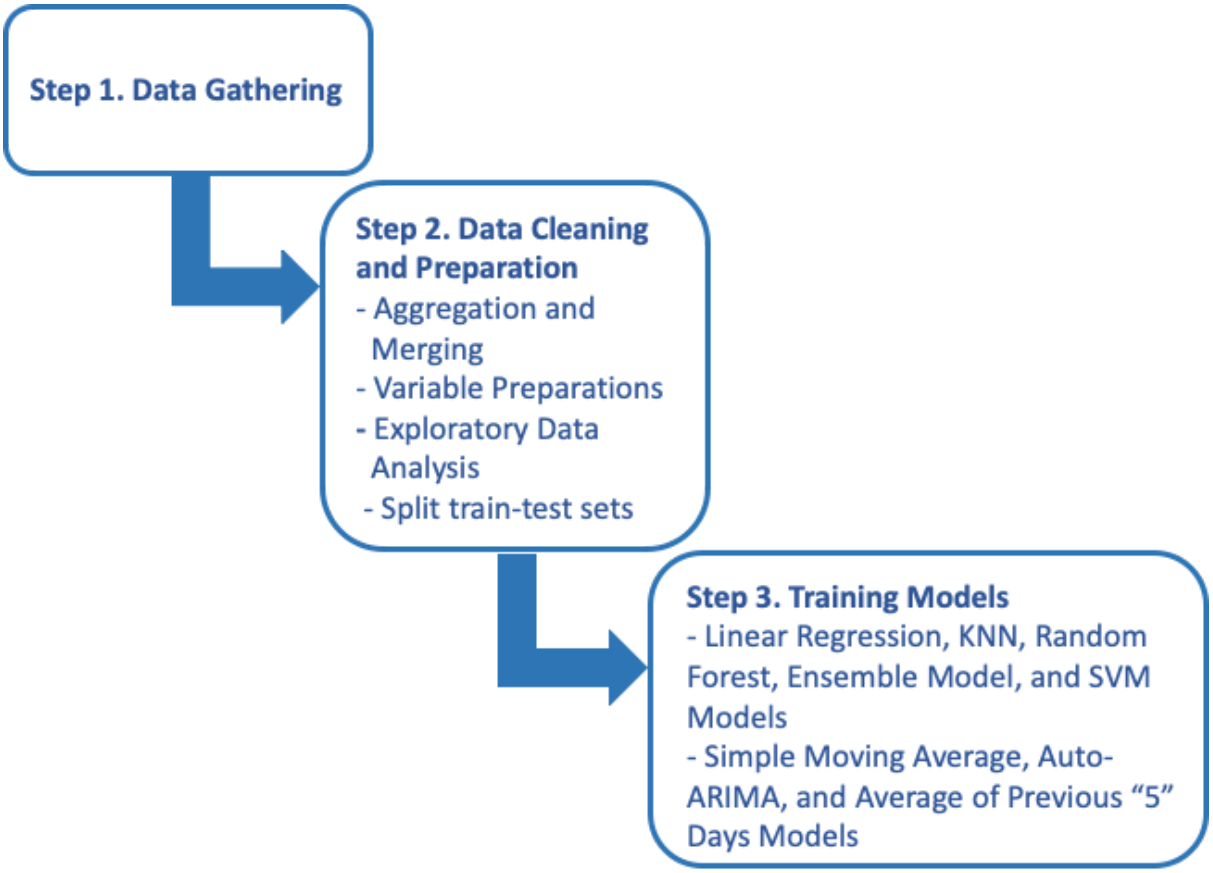


Figure 1. Block Diagram of Research Question 1 Approach

Step 1: Gathering Data

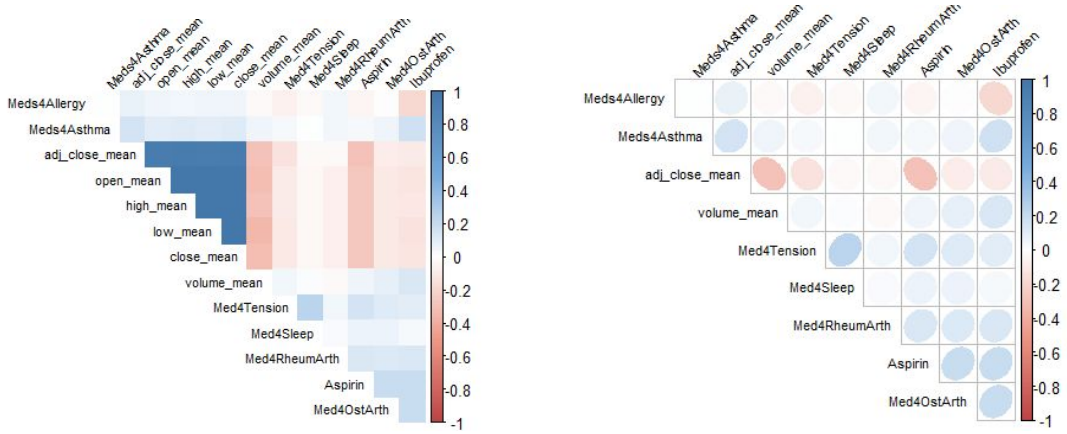
See dataset description section.

Step 2: Data Cleaning and Preparation

From the *Stock Market Dataset* (Onyschchak, 2020) repository, we extracted and compiled the *.csv* stock files of the following US largest pharmaceutical companies: Pfizer (PFE), Novartis (NVS), Merck & Company (MRK), GlaxoSmithKline (GSK), Johnson & Johnson (JNJ), AbbVie (ABBV), and Sanofi (SNY). Then, we subset the data to match the 2014 to 2018. Hence, we assume that their stock numbers should dictate and reflect the trend in the pharmaceutical market. For both merging and analytical purposes, we grouped the subsetted *Pharma Stock Market* dataset above by date, and summarised the averages of Open, High, Low, Close, Adjusted Close, and Volume values. Finally, the dataset is merged with the *Pharma sales data* by Zdravkovic (2020), by reformatting dates of both datasets into mutual forms.

For better interpretability, medicines codes in the *Pharma sales data* were changed into specific medicine groups and names. Several columns irrelevant to our analysis, such as Index, Year, Month, and Hour, as well as unmatched N/A row, were dropped. As we are predicting the low value of the next day, we created a new variable called '*low_price_next_day*' (Low Price of stock of next day) by shifting up the variable Low (low stock price) by one day. The last row because it has no value for the dependent variable low_price_next_day. This obtained dataset is final and consists of 1257 tuples and 17 variables: the dependent variable low_price_next_day, 15 independent variables and the variable date.

Next, **correlation analysis** was performed on the 14 numeric independent variables (IVs) in order to identify and remove highly correlated independent variables, with the cutoff set at 0.6. Please see **Figure 2** below for the correlation matrices comparison of the numerics IVs, pre and post highly-correlated IVs removal. The remaining IVs were then normalised.



Pre-removal: Past-removal:

Figure 2. Comparison of Correlation Matrices Before and After Removal of Highly-Correlated Independent Variables, respectively

Exploratory data analysis (EDA) of the *low_price_next_day* was then performed in order to observe pattern and distribution of the dependent variable. In this process, we found that the target variable is fairly symmetrical/normally distributed, with a skewness of 0.37, as shown below in **Figures 3 to 5**.

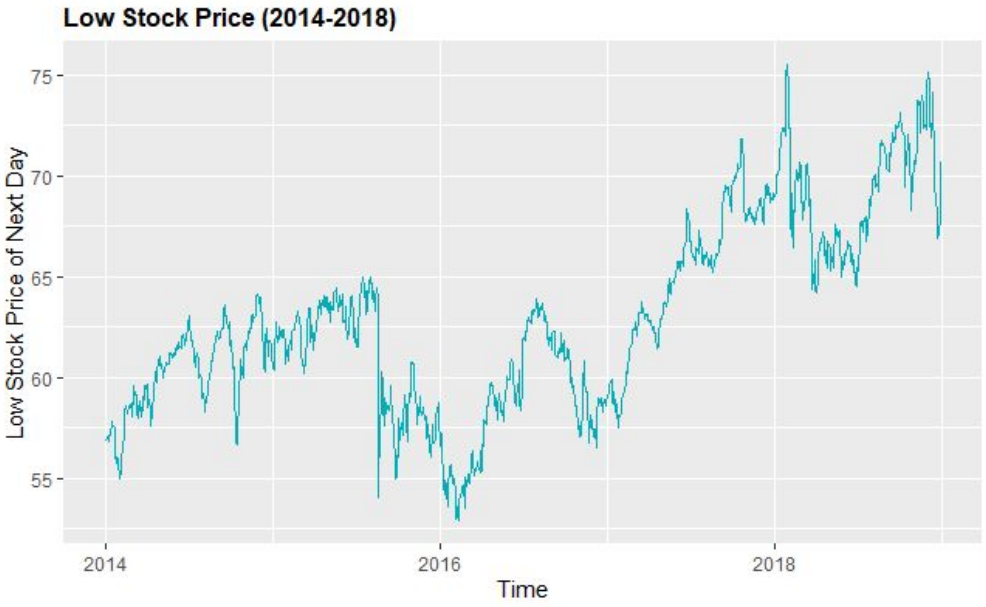


Figure 3. Chart of Low Stock Price (2014-2018)

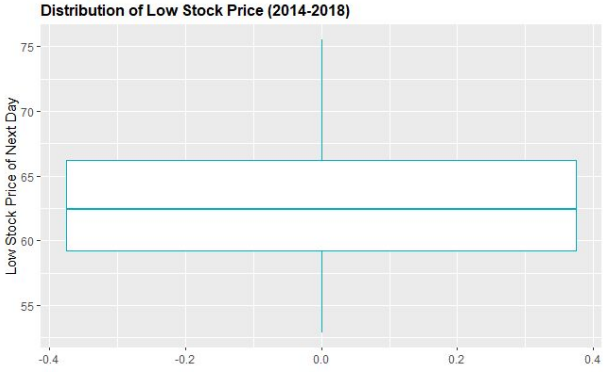


Figure 4. Distribution of Low Stock Price (2014-2018)

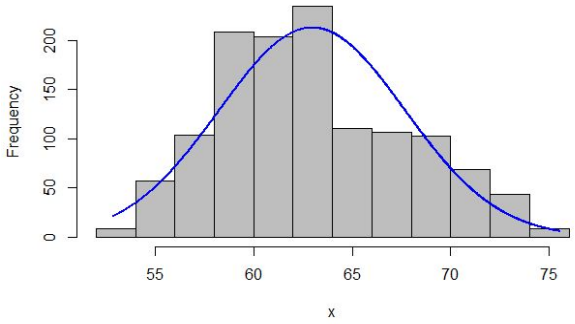


Figure 5. Normal Distribution Plot of Low Stock Price (2014-2018)

Feature selection was then performed using the backward elimination method in order to see which IVs are insignificant in helping our models achieve better results. Consequently, the non-significant IVs of *Med4RheumArth, Aspirin, Ibuprofen, Med4Tension, Meds4Asthma,*

Meds4Allergy, adj_close_mean, and *volume_mean,* as shown in **Image 1***,* were then dropped.

```
Call:
lm(formula = low_price_next_day ~ Med4RheumArth + Aspirin + Ibuprofen +
    Med4Tension + Med4Sleep + Meds4Asthma + adj_close_mean +
    volume_mean, data = data_n[, -11])

Residuals:
    Min      1Q  Median      3Q     Max
-5.7744 -1.1178 -0.0119  1.2123  3.5585

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    56.2114     0.1981 283.737  < 2e-16 ***
Med4RheumArth  -1.2701     0.2691  -4.720 2.62e-06 ***
Aspirin         0.6686     0.2939   2.275 0.023065 *
Ibuprofen      -2.0623     0.5022  -4.106 4.28e-05 ***
Med4Tension     1.4132     0.4018   3.517 0.000451 ***
Med4Sleep      -0.6049     0.3319  -1.822 0.068625 .
Meds4Asthma    -0.8475     0.2812  -3.014 0.002631 **
adj_close_mean 19.3454     0.2018  95.873  < 2e-16 ***
volume_mean    -4.0254     0.6194  -6.499 1.17e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.463 on 1248 degrees of freedom
Multiple R-squared:  0.9045,    Adjusted R-squared:  0.9039
F-statistic:  1477 on 8 and 1248 DF,  p-value: < 2.2e-16
```

Image 1. Feature Selection Results of Low Price (Next Day)

Step 3: Training the Models

We trained eight different regression models using eight different techniques: linear regression, KNN regression, RF, SVM, ensemble model, simple moving average method, Auto-ARIMA, and average of previous “n” days method with n arbitrarily set to 5 because the stock market operates only on weekdays except holidays. For the linear regression, KNN, RF, SVM, ensemble model, simple moving average, Auto-ARIMA models, the data was split in training-test sets by year with 2014-2017 used as training and 2018 as test.

Linear Regression, KNN, RF, Ensemble Model, and SVM Models

For the linear regression, KNN, RF, ensemble model, and SVM, models, the variable date was dropped from both the training and test sets after it served its splitting purpose. Cross validation with 5 folds and 3 repeats were used for all these five models. For the KNN regression, the K was determined to be 3; for RF, the best number of trees was found to be 425.

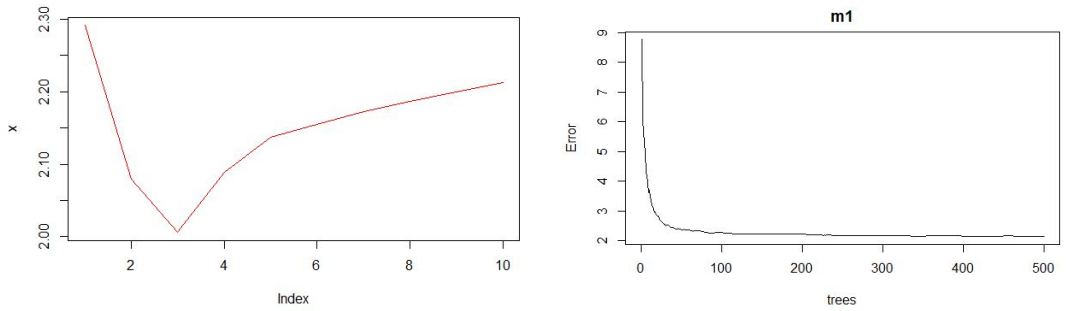


Image 2. Best K for KNN and Best ntrees for RF Models

For the linear regression model, All four assumptions of parametric linear regression (linearity, independence, normality and equal variance) were violated here. See Figure 9. However, we decided to go against a transformation of the dependent variable for interpretability and comparison (with other models) reasons.

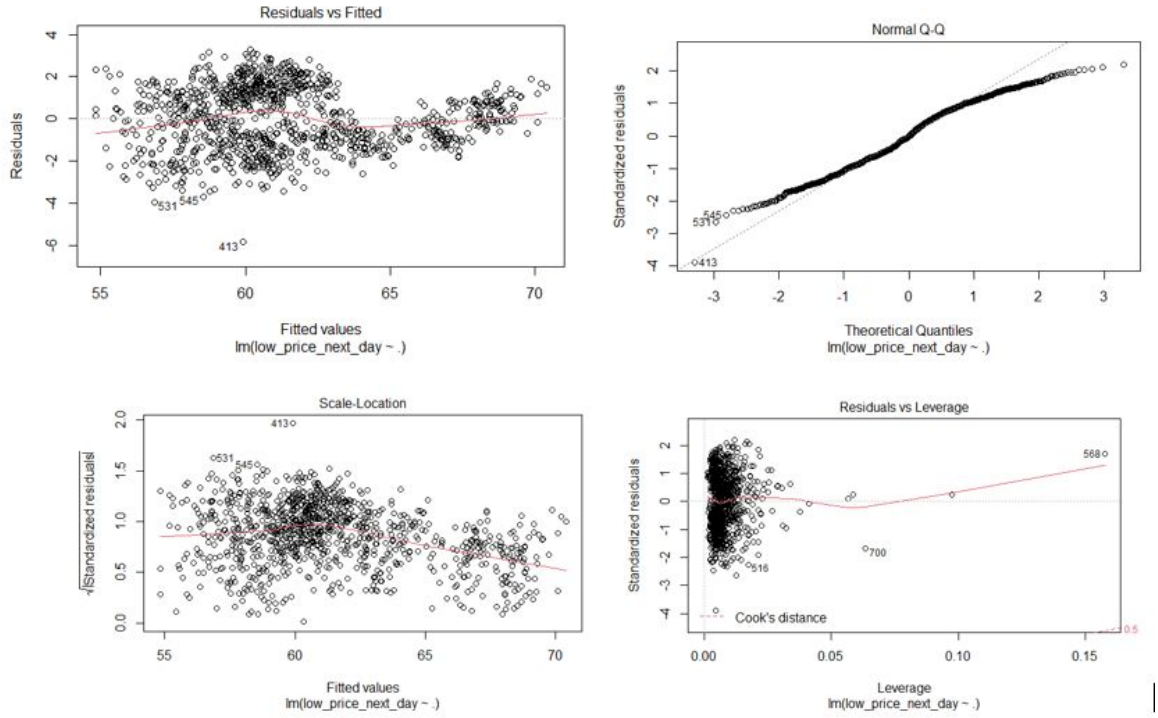


Image 3. Residual vs Fitted, Scale-Location, Residual vs Leverage and Normal Q-Q

As for the ensemble model, the best ensemble was determined to be random forest and linear regression. Please see **Image 4.**

```
A rf ensemble of 2 base models: rf, lm

Ensemble results:
Random Forest

3021 samples
   2 predictor

No pre-processing
Resampling: Cross-Validated (5 fold, repeated 3 times)
Summary of sample sizes: 2416, 2417, 2416, 2417, 2418, 2417, ...
Resampling results:

  RMSE      Rsquared   MAE
  1.385207  0.8580494  1.025302

Tuning parameter 'mtry' was held constant at a value of 2
```

Image 4. Results of RF ensemble model

Simple Moving Average, Auto-ARIMA, and Average of Previous "5" Days Models

For these methods, a time-series dataset was created from the final dataset with just the variables date and Low (stock price) from which the variable of interest low_price_next_day (Low Price of stock of next day) is derived.

RQ 1 :Results and Discussion

For all eight models, the main focus was on the following evaluation metrics (please see **Table 4** for evaluation summary):

1. RMSE (root mean square error)
2. Pred25%: percentage predicted values within 25% of actual values (less than 25% error)
3. Pred10%: percentage predicted values within 10% of actual values (less than 10% error)
4. Pred5%: percentage predicted values within 5% of actual values (less than 5% error)
5. Pred1%: percentage predicted values within 1% of actual values (less than 1% error)

Table 4: Heat-Map of Evaluation Metrics on Test Set for all Models for Q#1

Models	RMSE	Pred25%	Pred10%	Pred5%	Pred1%
Linear Regression	1.36	100.0%	100.0%	99.6%	34.8%
KNN Regression	2.01	100.0%	100.0%	94.4%	24.4%
Random Forest	1.68	100.0%	100.0%	97.2%	22.4%
SVM	1.38	100.0%	100.0%	98.8%	33.6%
Ensemble Model	1.70	100.0%	100.0%	97.6%	28.4%

Simple Moving Average	3.69	100.0%	97.2%	60.4%	16.4%
Auto-Arima	2.76	100.0%	100.0%	79.6%	13.6%
Average of Previous 5 Days	0.89	100.0%	99.9%	99.6%	60.2%

As shown in the model (See Table 4), linear regression has an RMSE score of 1.36 and 99.6% of predicted values fall within 5% of the actual values. On another note, the average of the previous 5 days method outperformed the ML techniques. The reported results of the average of the previous 5 days model is RMSE of 0.98 and 99.6% of prediction values falling within 5%. The average of the previous 5 days model may be performing optimally due to the stability of the stock market data inputted into the train and test dataset. In a case where the stock market performs chaotically, this technique would not work fittingly.

On another note, the worst-performing models were Auto-ARIMA and the simple moving average method. The Auto-ARIMA model builds on the moving average method. This explains why the simple moving average method yielded the worst performance. Both of these models can predict the low stock market value with an error tolerance of up to 10% with high accuracy as shown in **Table 4**. At a 5% error rate there the models performed moderately. This might influence us to believe that they were “good models". However, from an investor perspective, these error rates translate into the potential monetary loss of individuals and corporations. **Figures 6 and 7** are visualizations of the models.

Figure 6. Visualization of the Models Part 1

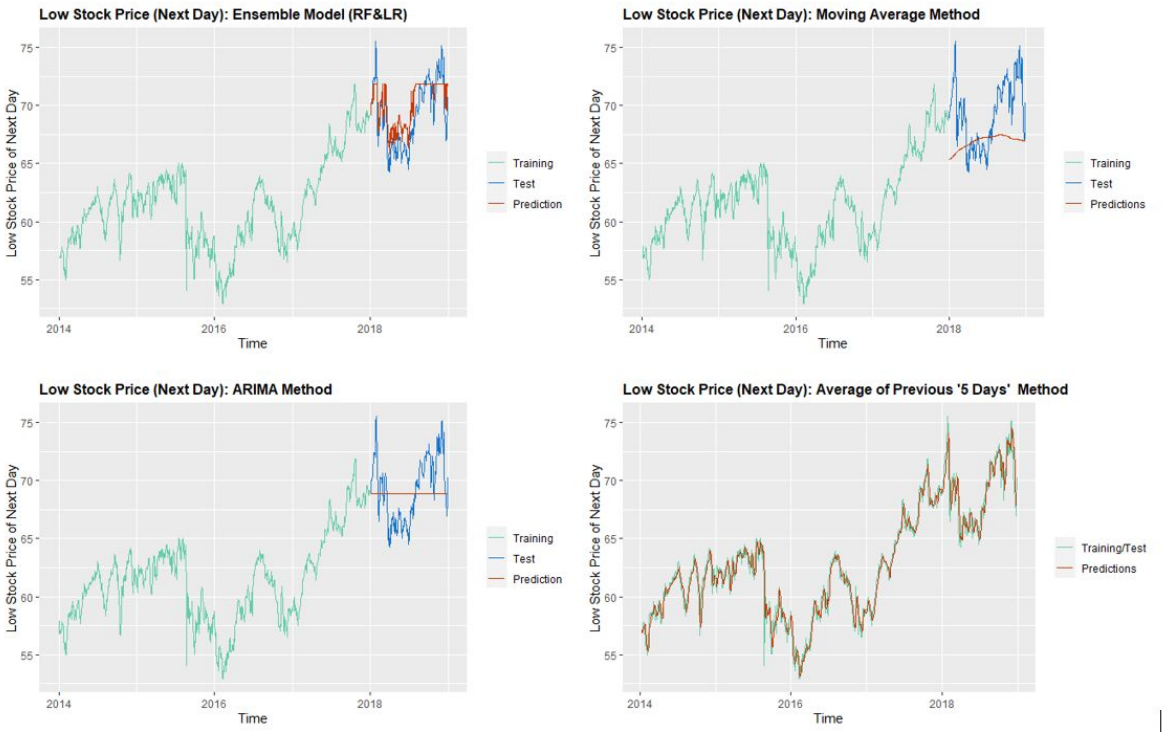
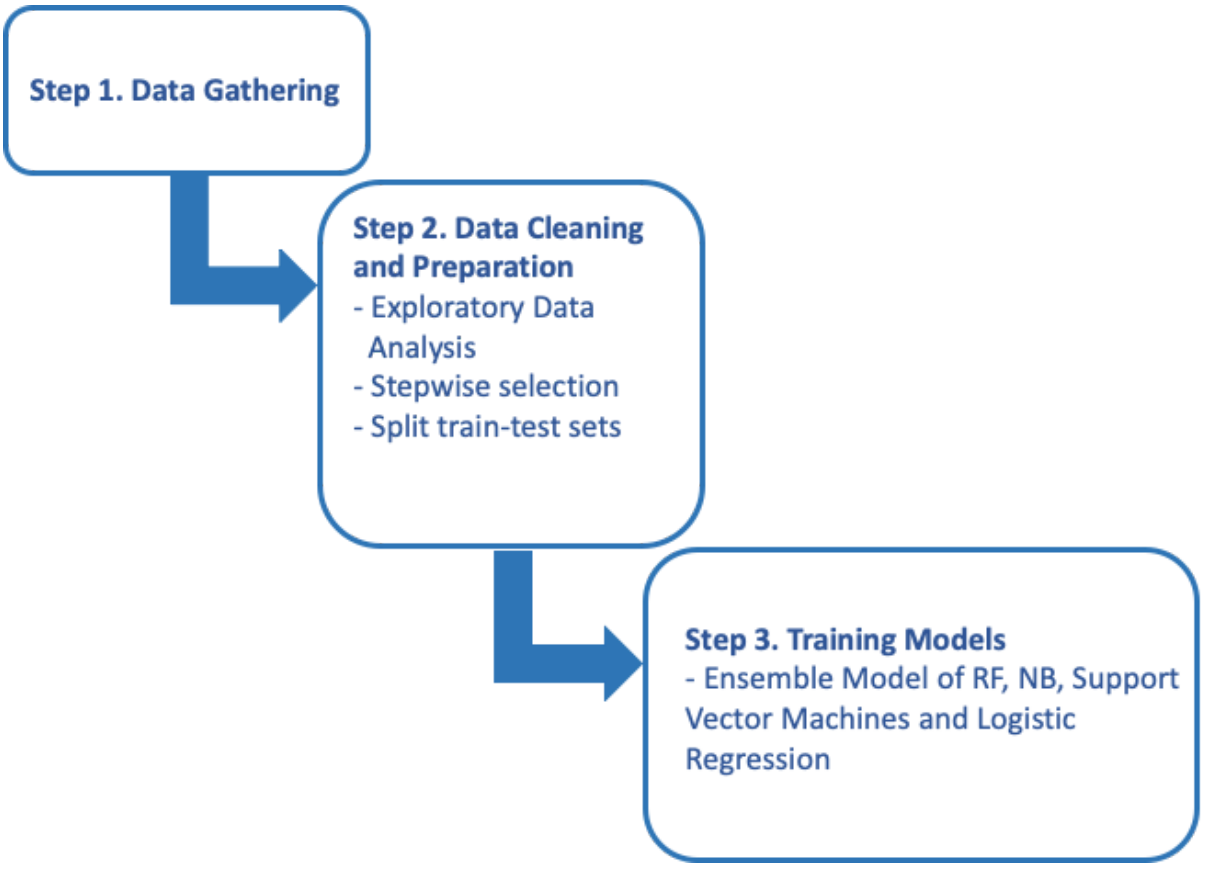


Figure 7. Visualization of the Models Part 2

Research Question(s) 2

Our second question has two subquestions.

1) Are there changes in the types of drugs sold in the years after Trump was elected?
2) Additionally, are there general changes in the overall financial standing of the pharmaceutical industry between 2014 and 2018?



Step 1: Gathering Data

See dataset description section.

Step 2: Data Cleaning and Preparation

For these questions we used 200+ Financial Indicators of US stocks 2014-2018 and Pharma Sales Data; Pharma Sales Data is a six years data 2014-2019 (Zdravkovic, 2020) on sales of drugs classified in 8 categories. We did not combine the datasets, and rather attempted to answer our questions using them separately. Pharma Sales data was normalized using log function as the data was right-skewed, and, for better interpretability, medicines codes were changed into specific medicine groups and names. Feature selection was done to find the best predictor of each drug. Although our interest was in how well "date" could be used to predict the volume of certain drugs, based on the few features available to use on this dataset, we suspected there could be a relationship between the dispensing of drugs. (for example if certain prescriptions were written together). **Stepwise selection** indicated to us that out of the eight drugs that were contained in the dataset five of them had 'date' a significant variable.

Financial Indicators data comprises five csv for each year. After standard data transformation to include year and merge the datasets, we found and removed correlated predictors; for missing values, we decided to impute mean values. Since the dataset did not have an explicit pharmaceutical sector, we found a list of major pharma companies and their stock tickers (TopForeignStocks.com Team, 2019 & 2020). The dataset we worked with had stocks from both The New York Stock Exchange and NASDAQ. We subsetted our dataset cross-referencing the list, normalized the values (x - min(x)) / (max(x) - min(x))).

For the stocks dataset, since there were over 200 features, we ran a correlation matrix which eliminated half of the features in the dataset. Then we used the Backward Elimination method to find the most important features to feed into our models. For the classification model, the binary class labels of “0” and “1” were transformed into “dec” and “inc” respectively.

Step 3: Training the Models

For subquestion 1 (Are there changes in the types of drugs sold in the years after Trump was elected?), We trained 5 linear regression models on data from 2014-2015 years with 70/30 train-test split. This was done to the model was run individually on five types of medicines with ‘date’ being the dependent variable. This was done to test whether the errors of the model were within an acceptable range. Since the results were satisfactory, we then took the entire 2014-2015 (Obama period) dataset as training and tested 2017/2018 dataset on the trained model. The results are discussed below.

To answer the subquestion 2 (are there general changes in the overall financial standing of the pharmaceutical industry between 2014 and 2018?) We used both classification and regression techniques. We decided to merge the 2014 and 2015 dataset as the years of Obama’s office, and 2017 and 2018 were combined as Trump’s office years. We refer to 2014/2015 merged data as Obama and 2017/2018 merged data as Trump. We did not use 2016 data as that was an election year and the data may have unexplained noise.

For classification we trained an ensemble model.

The US stocks dataset with more than 200 financial indicators contains a ‘class’ column to indicate whether stock prices rose during that year. Based on this class column an ensemble model of classifiers was built. After running correlation analysis and feature selection, 11 columns were found to be significant in identifying whether stock price increased during the year. The 11 columns which made up the subsetted dataset include: Long Term Investments, Tax Liabilities, Financing Cash Flow, Net Cash Marketcap, price earnings to growth ratio, Net debt to ebitda,Graham number, net-income growth per share, asset growth and class.

Using the predictors selected earlier through feature selection, we trained the Ensemble Model of RF, NB, Support Vector Machines and Logistic Regression to predict the ‘Class’ label using the Obama data with 70/30 train-test split. None of the models were highly correlated with each other but NB had lower accuracy than the other three so it was removed and another ensemble model was trained, which performed moderately. Data from the Trump Era were tested against the model.

Regression. First, we split obama data into train and test, 70:30 ratio and ran multiple

regression to predict price of a stock. Its Root Mean Square Error is 0.16, which shows that our estimations are good for our data as it was normalized and ranged between 0 and 1. Then, we trained the KNN model on the same Obama data, split into 70:30. The model performed worse than the multiple linear regression but not by a lot with RMSE of 0.21. Although KNN in general does not perform well in a high-dimensional data, its predictions were still acceptable. We did all the above to check the performance of our models. Now, we were ready to apply the same models to Trump data, only this time Obama data would be our train and Trump is our test set, which is about 50:50 split. The models performed well on our new test data as well. Going back to our question, this suggests that, according to our data and judging by pharma stock prices, the pharmaceutical industry's financial condition remained similar in two years of Trump's office.

RQ 2: Results and Discussion

Question 2, Subquestion 1: "Are there changes in the types of drugs sold in the years after Trump was elected?"

The linear regression performed fairly well for the pharma sales dataset. There were 10 models in total; two for each medicine for each of the time periods. The first 5 models were trained and tested on 2014-2015 data. The model performed fairly well for Rheumatoid Arthritis, Ibuprofen and Aspirin with 88% predicted values being within the 10% of the actual values. The model seems to be overfitting slightly 100% of the time the predicted values matched around 25% of the actual value

The metrics were not very high for Allergy and Sleep Medication. The metrics for sleep medicine were the worst of the 5 drugs. The model was predicting values within 10% of the actual values, only 34% of the time. However 66% of the, the predicted values fell within 25% of actual values.

The next 5 models were trained on 2014-2015 data and tested on 2017-2018.

[1] "2014-2015 model for Rheumatoid Arthritis Meds" [1] "RMSE: 0.207611965685788" [1] "PRED(10): 0.88" [1] "PRED(25): 1" [1] "Summary of Prediction" Min. 1st Qu. Median Mean 3rd Qu. Max. 3.237 3.287 3.379 3.427 3.575 3.673	[1] "2014-2015 model for Aspirin" [1] "RMSE: 0.23488553992087" [1] "PRED(10): 0.88" [1] "PRED(25): 1" [1] "Summary of Prediction" Min. 1st Qu. Median Mean 3rd Qu. Max. 3.428 3.431 3.437 3.439 3.448 3.454	[1] "2014-2015 model for Ibuprofen" [1] "RMSE: 0.294812540818012" [1] "PRED(10): 0.88" [1] "PRED(25): 1" [1] "Summary of Prediction" Min. 1st Qu. Median Mean 3rd Qu. Max. 5.034 5.093 5.198 5.253 5.423 5.536	[1] "2014-2015 model for Sleep Meds" [1] "RMSE: 0.534654841947156" [1] "PRED(10): 0.22" [1] "PRED(25): 0.66" [1] "Summary of Prediction" Min. 1st Qu. Median Mean 3rd Qu. Max. 1.065 1.212 1.506 1.434 1.643 1.720	[1] "2014-2015 model for Allergy Meds" [1] "RMSE: 0.485796867272014" [1] "PRED(10): 0.34" [1] "PRED(25): 0.88" [1] "Summary of Prediction" Min. 1st Qu. Median Mean 3rd Qu. Max. 2.624 2.649 2.694 2.718 2.791 2.840
[1] "TrumpTest: Rheumatoid Athritis Med" [1] "RMSE: 0.644350094321413" [1] "PRED(10): 0.19" [1] "PRED(25): 0.86" [1] "Summary of Prediction" Min. 1st Qu. Median Mean 3rd Qu. Max. 3.922 4.034 4.147 4.147 4.259 4.371	[1] "TrumpTest: Aspirin" [1] "RMSE: 0.383906036858621" [1] "PRED(10): 0.6" [1] "PRED(25): 0.91" [1] "Summary of Prediction" Min. 1st Qu. Median Mean 3rd Qu. Max. 3.350 3.361 3.372 3.372 3.383 3.393	[1] "TrumpTest: Ibuprofen" [1] "RMSE: 0.8095214122615 35" [1] "PRED(10): 0.26" [1] "PRED(25): 0.91" [1] "Summary of Prediction" Min. 1st Qu. Median Mean 3rd Qu. Max. 5.727 5.841 5.954 5.954 6.067 6.180	[1] "TrumpTest: Sleep Medicine" [1] "RMSE: 1.256508138667 16" [1] "PRED(10): NA" [1] "PRED(25): 0.02" [1] "Summary of Prediction" Min. 1st Qu. Median Mean 3rd Qu. Max. 0.04665 0.22182 0.39700 0.39700 0.57218 0.74735	[1] "TrumpTest: Allergy Medicine" [1] "RMSE: 0.554797612893333" [1] "PRED(10): 0.39" [1] "PRED(25): 0.81" [1] "Summary of Predictions" Min. 1st Qu. Median Mean 3rd Qu. Max. 2.955 2.998 3.041 3.041 3.085 3.128

Linear Regression Model Metrics: "TrumpTest" refers to model being tested on 2017-2018 dataset

Table 5. Train data for 2014-2015 and tested 2017-2018.

The models performed slightly worse but the previous pattern remained the same with Allergy medicine and Sleep medicine having worse predictions than the other two meds. Sleep medicine also had the worst results in this case with only two percent of the predictions falling within 25% of the actual values. Please see table 2 for all medicine metrics.

We can see from table 2 that the model performed well with more than 80% of the predicted values being within 25% of the actual values. The metrics are especially good because the test dataset is completely separate from the training data. Does this mean our

model is good at predicting when certain medicines will be sold since it is predicting medicine volume based on date? Not necessarily because if we look at the medicines that scored consistently well across models (Rheumatoid Arthritis, Aspirin and Ibuprofen) they are drugs that are always in need. Rheumatoid arthritis does not go away after a while, it stays with people their whole lives and so that drug will most likely be bought at a consistent frequency. The same logic can be applied to Ibuprofen and Aspirin as they are both pain medicine that people also buy on a consistent basis for a variety of reasons. The poorly scored drugs like allergy and sleep medicine are bought seasonally and don't follow a straightforward pattern hence the model trained on one kind of pattern infrequently wouldn't necessarily predict the volumes of the drugs accurately if the drugs sold in the test set followed completely different patterns.

Question 2 subquestion 2 : "Are there general changes in the overall financial standing of the pharmaceutical industry between 2014 and 2018?"

Overall, the Regression Model had better results than classification

```
Confusion Matrix and Statistics

          Reference
Prediction dec inc
       dec 137  67
       inc  21  15

               Accuracy : 0.6333
                 95% CI : (0.5689, 0.6944)
    No Information Rate : 0.6583
    P-Value [Acc > NIR] : 0.8123

                  Kappa : 0.0578

 Mcnemar's Test P-Value : 1.61e-06

            Sensitivity : 0.1829
            Specificity : 0.8671
         Pos Pred Value : 0.4167
         Neg Pred Value : 0.6716
              Precision : 0.4167
                 Recall : 0.1829
                     F1 : 0.2542
             Prevalence : 0.3417
         Detection Rate : 0.0625
   Detection Prevalence : 0.1500
      Balanced Accuracy : 0.5250

       'Positive' Class : inc
```

Classification metrics Obama 70/30 split

Image 5

Results for classification

The model performed moderately when trained and tested on Obama era data. The model returned an accuracy of 63% with precision of 42%, recall score of 18% and f1 of 25%. When the trained model was tested on Trump era data, the Accuracy(51%) and Precision(38%) fell but Recall (43%) and the F1(40%) scores improved slightly (although the scores were still very poor). See images 5 and 6.

Low precision means there is a high false positive rate. This means our model frequently predicted that stocks prices would have increased when in actuality they didn't. High recall refers to low false negative rates which means the model predicted decrease in price when in reality the price did increase. Overall the classification model performed poorly in predicting increase or decrease in stock prices.

In order to better interpret the features that were deemed significant by feature selection we would need some background into stocks and finance but even without the domain knowledge, it is clear that the features that were selected reflect a company's hold on their assets. Whether or not a company has high or low Long Term Investments, Price Earning to Growth Ratio, Net Debt to EBITDA, etc, all these factors reflect a company's financial standing based on their assets and these features are what the ensemble model uses to best classify whether stocks increased or decreased in a certain year. Increases and decreases in stock prices are not only related to how much wealth or liabilities a company has. For the sake of consistency in our classification analysis, we used the same features for the Trump data that we trained our Obama model. This may be one of the reasons why classification did so poorly as there may have been other

features in the trump data that better predicted stock prices, which were not considered in our analysis.

```
Confusion Matrix and Statistics

          Reference
Prediction dec inc
       dec 330 208
       inc 253 155

               Accuracy : 0.5127
                 95% CI : (0.4803, 0.545)
    No Information Rate : 0.6163
    P-Value [Acc > NIR] : 1.00000

                  Kappa : -0.0068

 Mcnemar's Test P-Value : 0.04043

            Sensitivity : 0.4270
            Specificity : 0.5660
         Pos Pred Value : 0.3799
         Neg Pred Value : 0.6134
              Precision : 0.3799
                 Recall : 0.4270
                     F1 : 0.4021
             Prevalence : 0.3837
         Detection Rate : 0.1638
   Detection Prevalence : 0.4313
      Balanced Accuracy : 0.4965

       'Positive' Class : inc
```

Classification metrics Obama trained model with Trump test data

Image 6

Results for Regression

For multiple linear regression, we used all non-correlated features to predict stock price as the model constrained only to selected features performed worse. The regression was trained and tested on 2014-2015 years (“Obama” data) which was split 70:30 ratio.

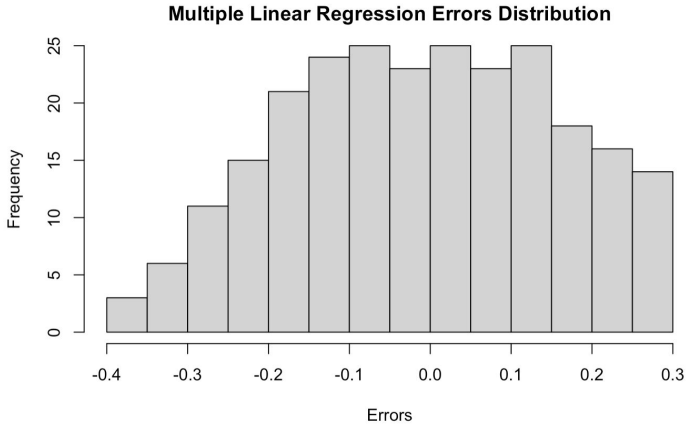


Image 7. Multiple Linear Regression Error Distribution

As shown in the histogram (Image 7), the errors range from -0.4 to 0.3. Root Mean Square Error of 0.1642 is also within the acceptable range for our data. We also found that 46% of the predicted data lay within 25% of the observed data.

Since the model produced satisfactory results, we trained it on the entire “Obama” data and tested on 2017-2018 years (“Trump” data). The results were very close to the previous one. Although RMSE increased by 0.005, there was a 5% increase in the predicted values within 25% of the trained data.

K-Nearest Neighbor (KNN)

Image 8. Breakdown of KNN

We performed a similar analysis with the KNN model first on “Obama” and then on “Trump” data. The model showed best results with K = 8. Figure X illustrates observed versus predicted values and predicted values versus errors. Since the values in the first plot follow a straight line and there are no clear patterns in the second plot, this suggests that the model predicted moderately well.

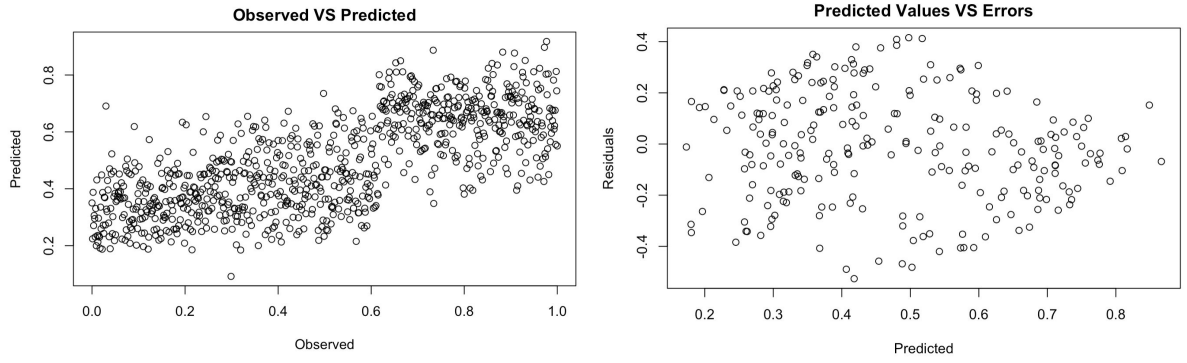


Image 9. KNN Observed vs predicted and Predicted values vs Errors

KNN model produced RMSE of 0.21 and 43 to 44% of the predicted values are within 25% of the actual values.

Conclusions

The results of our study indicates that linear regression was the best model to detect and predict trends based on our data set of US pharmaceutical stocks over the period 2014 to 2018. The drug sales patterns remained relatively the same between the two differing party presidential administrations. Looking closer at the drugs covered in our data set, this makes

sense since the drugs are primarily related to chronic conditions, so we would expect usages to remain relatively consistent throughout the years. As the overall market was stable over the years 2014-18, averaging stock price over the previous days worked as well or better than the ML techniques. Further research will be required to ensure that models perform well over longer time frames and over periods of higher price volatility. Promising areas for further explorations include neural network models as well as with more complex models specific to time-series data. That said, results of this report provide initial insights on ML approaches to creating an investment strategy that produces market beating returns.

References

Adebiyi A. A., & Adewumi, A. O., & Ayo, C. K.. (2014) Comparison of ARIMA and Artificial Neural Networks Models for Stock Price Prediction. *Journal of Applied Mathematics*, *2014*. https://doi.org/10.1155/2014/614342

Arnott, R. et al. (2019) Alice's Adventures in Factorland: Three Blunders That Plague Factor Investing. *The Journal of Portfolio Management, 45* (4) 18-36; doi: 10.3905/jpm.2019.45.4.018

Carbone, N. (2020). 200+ Financial Indicators of US Stocks (2014-2018) [Data set]. Kaggle. Retrieved from https://www.kaggle.com/cnic92/200-financial-indicators-of-us-stocks-20142018

Coqueret, G, & Guida, T. (2020). *Machine Learning for Factor Investing: R Version* (1st ed.). Chapman and Hall/CRC.

Onyschchak, O. (2020). Stock Market Dataset: Historical daily prices of all stocks and ETFs [Data set]. Kaggle. Retrieved from https://www.kaggle.com/jacksoncrow/stock-market-dataset?select=symbols_valid_meta.csv

Patel, J., et al. (2015) Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques. *Expert Systems with Applications, 42*(1), 259-268; doi.org/10.1016/j.eswa.2014.07.040.

Ramasubramanian K., & Singh A. (2019) Time Series Modeling. In: *Machine Learning Using R*. Berkeley, CA: Apress..

TopForeignStocks.com Team. (2020). Stock Symbol lists used by Faria/Asel/Marcus - cite in the data and or method sections. [Webpage and Excel file] Retrieved from https://topforeignstocks.com/stock-lists/the-complete-list-of-biotech-stocks-trading-on-nasdaq/

TopForeignStocks.com Team. (2019) The Complete List of Major Pharmaceutical Stocks on the NYSE. [Webpage and Excel file] Retrieved from https://topforeignstocks.com/stock-lists/the-complete-list-of-major-pharmaceutical-stocks-on-the-nyse/

Yang, J. et al. (2016) Ensemble Model for Stock Price Movement Trend Prediction on Different Investing Periods. *2016 12th International Conference on Computational Intelligence and Security (CIS)*, Wuxi. 358-361; doi: 10.1109/CIS.2016.0087

Yoo, P. D., & Kim, M. H., & Jan, T. (2005) Machine Learning Techniques and Use of Event Information for Stock Market Prediction: A Survey and Evaluation. *International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06), 835-841;* doi: 10.1109/CIMCA.2005.1631572

Zdravkovic, M. (2020). Pharma sales data: Six years data (2014-2019) on sales of drugs classified in 8 ATC categories [Data set]. Kaggle.
Retrieved from https://www.kaggle.com/milanzdravkovic/pharma-sales-data