MARRKETING ANALYTICS REPORT

Stock Price Prediction by Time Series and Sentiment Analysis

Created By - Lei Zhao, Aman Kumar, Shaobo Pan, Athira Praveen

A. Executive Summary

The problem statement is to predict stock price taken 10-years-worth of historical price data (2006 - 2016) from the Dow Jones and the news headlines from New York times articles (sentiment analysis). Coding on R, firstly apply Time Series Analysis (Holt-Winter exponential smoothing) to make prediction. Next, incorporate Random Forest and Sentiment Analysis to enhance forecast.

RMSE is chosen as the major performance criteria. It successfully decreases from 4000 to 1524 by adding seasonality component into Time Series Analysis. Using sentiment analysis scores as predictors for Random Forest, RMSE keeps going down from 1027 to 770 by deploying smoothing techniques.

B. Goals

The goals of this project include:

- To predict the 2 years stock market movement with the help of 8 years historical data.
- To reduce RMSE of the predicted values.
- To find if the sentiment analysis on the news actually effect the forecasting accuracy.

C. Data Preprocessing

1. Original Data

Date	Open	High	Low	Close	Volume	Adj Close
12/30/16	19833.1699	19852.55078	19718.6699	19762.5996	271910000	19762.59961
12/29/16	19835.4609	19878.41945	19788.9395	19819.7793	172040000	19819.7793
12/28/16	19964.3105	19981.10938	19827.3105	19833.6797	188350000	19833.67969
12/27/16	19943.4609	19980.24023	19939.8008	19945.0391	158540000	19945.03906
12/23/16	19908.6094	19934.15039	19899.0605	19933.8105	158260000	19933.81055
12/22/16	19922.6797	19933.83008	19882.1895	19918.8809	258290000	19918.88086
12/21/16	19968.9707	19986.56055	19941.9609	19941.9609	256640000	19941.96094
12/20/16	19920.5898	19987.63086	19920.4199	19974.6191	284080000	19974.61914
12/19/16	19836.6602	19917.7793	19832.9492	19883.0605	302310000	19883.06055
12/16/16	19909.0098	19923.16992	19821	19843.4102	573470000	19843.41016
12/15/16	19811.5	19951.28906	19811.5	19852.2402	357350000	19852.24023
12/14/16	19876.1309	19966.42969	19748.6699	19792.5293	408430000	19792.5293
12/13/16	19852.2109	19953.75	19846.4492	19911.2109	388420000	19911.21094
12/12/16	19770.1992	19824.58984	19747.7402	19796.4297	333660000	19796.42969
12/9/16	19631.3496	19757.74023	19623.1895	19756.8496	334470000	19756.84961
12/8/16	19559.9395	19664.9707	19527.8301	19614.8105	324570000	19614.81055
12/7/16	19241.9902	19558.41992	19229.8301	19549.6191	385200000	19549.61914
12/6/16	19219.9102	19255.89063	19184.7402	19251.7793	284960000	19251.7793
12/5/16	19244.3496	19274.84961	19186.7305	19216.2402	317800000	19216.24023
12/2/16	19161.25	19196.14063	19141.1797	19170.4199	84920000	19170.41992
12/1/16	19149.1992	19214.30078	19138.7891	19191.9297	108800000	19191.92969
11/30/16	19135.6406	19225.28906	19123.3809	19123.5801	164570000	19123.58008
11/29/16	19064.0703	19144.40039	19062.2207	19121.5996	81510000	19121.59961

Figure 1

Open : the total price of all stocks at the beginning of the day.

High : the highest price of all stocks daily

Low : the lowest price of all stocks daily

Close : the sum price of stocks when the stock market is closed

Volume : the number of stock purchased a day

Adj Close : the changed total price after the market closed

2. Web Scrapping

	articles
1/1/07	. What Sticks from '06. Somalia Orders Islamists to Turn in Weapons. Tehran Radio Lets Critic
1/2/07	. Heart Health: Vitamin Does Not Prevent Death by Heart Disease. Pilot Errs Over Bush Ranch.
1/3/07	. Google Answer to Filling Jobs Is an Algorithm. Germany: Daimler in Deal Over Suit. Israeli A
1/4/07	. Helping Make the Shift From Combat to Commerce. Addenda. Indications of a Slowdown in S
1/5/07	. Rise in Ethanol Raises Concerns About Corn as a Food. New Majority,Ãôs Choice: Should G.O
1/6/07	. A Status Quo Secretary General. Best Buy and Circuit City Report Brisk Sales for December. C
1/7/07	. THE COMMON APPLICATION; Typo.com. Jumbo Bonuses: Dial Your Envy Down a Notch. Can I
1/8/07	. VW
1/9/07	. The Claim: Hot Leftovers Should Cool at Room Temperature. I,Ãôll Have Fries and a Glass of
1/10/07	. Love Among the Ruins. Dell Says Plant a Tree, Help the Environment. Mayor Vows to Crack D
1/11/07	. The Computer With a TV, and a Family's Virtual Bulletin Board. Getting Graphic With Vista .
1/12/07	. Make Them Fight All of Us. Hire by the Contract Now, Risk a Big Regret Later. Massachusett
1/13/07	. Blair Urges Britain to Pursue an Aggressive Foreign Policy. A.M.D., in Price War With Intel, W
1/14/07	. Smoke Damage. Mr. Spitzer,Ãôs Task on Court Reform. A Tepid Winter Warms Some Wallet
1/15/07	. The Mentally Ill, Behind Bars. BP,Ãôs Chief to Join Apax, a Private Firm. The Light-Touch Tax
1/16/07	. King Day in Atlanta, ,Ãòthe One Without Mrs. King,Ãô. Israeli Leader Faces Inquiry Over Bank
1/17/07	. Racial Hate Feeds a Gang War,Ãôs Senseless Killing. Islamist Fighters Captured Fleeing Som
1/18/07	. Taliban Detainee Says Rebel Chief Hides in Pakistan. Airbus Predicts Operating Loss for 2006
1/19/07	. Data Breach Could Affect Millions of TJX Shoppers. Foreign Sales Push Up Harley-Davidson,Ã
1/20/07	. Archives of Spin. H.P. Chief Defends Timing of Stock Sale. Czech Republic: Vote Ends Stalema
1/21/07	. Connecticut,Ãôs Diaspora. Son of Dogs Playing Poker. Lying Like It,Ãôs 2003. Turkish Police A

To analyze articles from The New York Times, we have used the R package -“SentimentAnalysis”

3. Preprocessed Data

Returns a dictionary with 4 metrics, which are as follows:

1.Positive – percentage of positive words in the corpus

2.Negative – percentage of negative words in the corpus

3.Neutral – percentage of neutral words in the corpus

4.Compound – overall sentiment of the article which is then normalized to make it a value between 0 & 1

Combined the results from the sentiment analysis and the stock prices to make our final data set which looked like follows:

	X	prices	compound	neg	neu	pos
1	2007-01-01	12469	-0.9814	0.159	0.749	0.093
2	2007-01-02	12472	-0.8179	0.114	0.787	0.099
3	2007-01-03	12474	-0.9993	0.198	0.737	0.065
4	2007-01-04	12480	-0.9982	0.131	0.806	0.062
5	2007-01-05	12398	-0.9901	0.124	0.794	0.082
6	2007-01-06	12406	-0.9650	0.134	0.771	0.094
7	2007-01-07	12414	-0.9975	0.193	0.739	0.069
8	2007-01-08	12423	-0.9601	0.110	0.793	0.097
9	2007-01-09	12416	-0.9953	0.103	0.848	0.049
10	2007-01-10	12442	-0.9534	0.134	0.743	0.123
11	2007-01-11	12514	-0.9980	0.128	0.814	0.057
12	2007-01-12	12556	-0.9986	0.158	0.784	0.059
13	2007-01-13	12562	-0.9893	0.146	0.794	0.059
14	2007-01-14	12569	-0.9900	0.178	0.711	0.111
15	2007-01-15	12575	-0.2636	0.068	0.830	0.082
16	2007-01-16	12582	-0.9798	0.110	0.804	0.086
17	2007-01-17	12577	-0.9986	0.176	0.743	0.081

Figure 3

The calculation of the rate of neg/neu/pos is:

Number(neg/neu/pos words in the article)÷ Number(total words in the aticle)

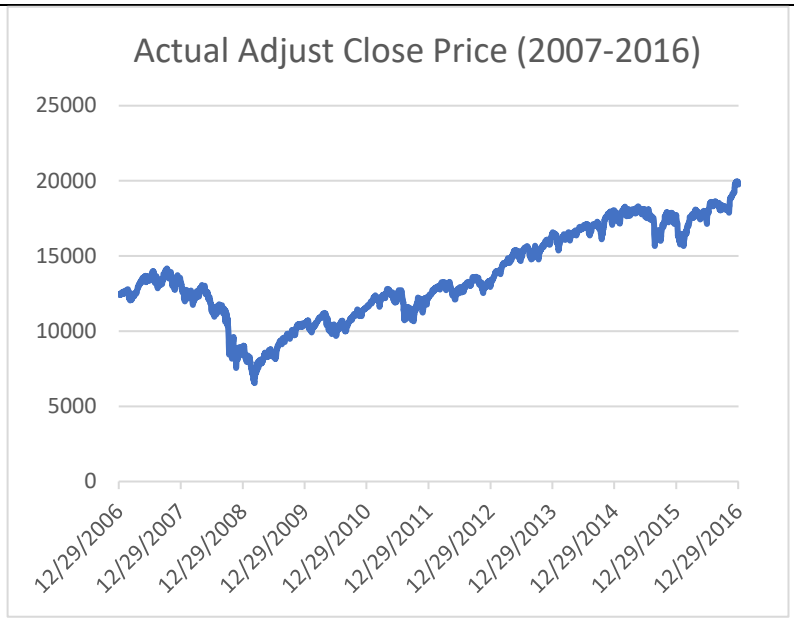
The calculation of the rate of compound is the sum of all of the lexicon ratings which have been standardized to range between -1 and 1.

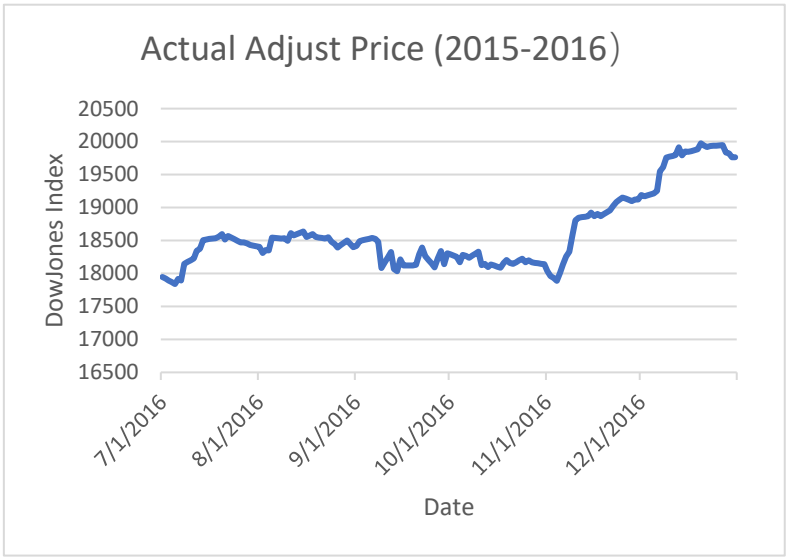
D. Analysis

1. Time Series Analysis

Why Time Series ?

- In reality, stock price fluctuates daily, which lead to profit changes customers pay most attention to
- Data is stock prices of each day from 2007 to 2016
- Trend and seasonality affects adjust close price. Overall price decreased from 2006 to 2008, and it goes up till 2016. Drilling down price from 2015 to 2016 (test data period), it increased at start of every month and dropped down at the end. This pattern continues around a season.





1) Data Split, Time Series Object

```
data = read.csv('DJIA indices data.csv',header = TRUE)
train_data = subset.data.frame(data, Date<="2014-12-31" & Date>="2007-01-01")
test_data = subset.data.frame(data,Date>="2015-01-01")

# set time(daily) range for train data
data$Date = as.Date(data$Date)
range = seq(as.Date("2007-01-01"), as.Date("2014-12-31"), by = "day")

# create a time series object for price (train data)
myts = ts(train_data$Adj.Close,
          start = c(2007,as.numeric(format(range[1], "%j"))),
          end = c(2014,as.numeric(format(range[1], "%j"))),
          frequency = 365)
```

2) Holt Model – With Trend

```
# double exponential - models level and trend
fit1 <- HoltWinters(myts, gamma=FALSE)

# predict price in 2015,2016
pred1 = forecast(fit1, h=length(test_data$Adj.Close))
p1 = data.frame(pred1)

# prediction graph
plot(pred1, main = 'Holt Forecast With Trend (2007-2014)')

# add constant value into prediction
# how to get constant?
# for each year, use the first 10 month data as train data, the last 2 month as test data
# get average error of the month, average them
p1$Point.Forecast = p1$Point.Forecast + 4528.33

# Actual vs Predicted pred1
plotting_data_frame = data.frame(date = test_data$Date,
                                 actual = test_data$Adj.Close, predicted = p1$Point.Forecast)
ggplot(plotting_data_frame, aes(date))+
       geom_line(aes(y = predicted,colour = "Predicted"))+
       geom_line(aes(y = actual,colour = "Actual")) + ggtitle("With Trend (2015-2016)")

rmse1 = rmse(test_data$Adj.Close,p1$Point.Forecast)
rmse1 #4001.85
```

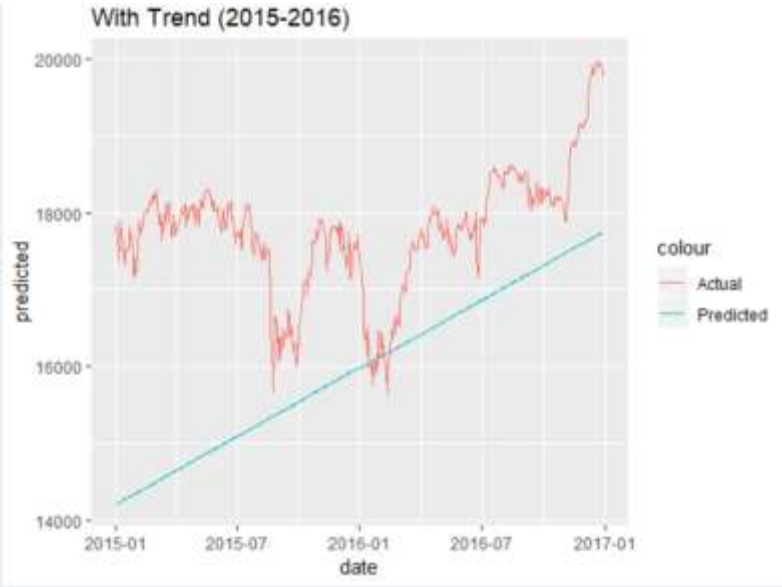


Figure 6

Using Holt exponential smoothing, the predicted price is increasing constantly since this model only contains trend component, which is not applicable to the reality. From line chart, predicted line is far away from actual line. And it does not have similar pattern as actual one.

3) Holt-Winter Model : 2477 of RMSE Reduction

```
# triple exponential - models level, trend, and seasonal components
fit2 <- HoltWinters(myts)

# predict price in 2015, 2016
pred2 = forecast(fit2, h=length(test_data$Adj.Close))
p2 = data.frame(pred2)
p2$Point.Forecast = p2$Point.Forecast+4528.33

# prediction graph
plot(pred2, main = 'Holt Forecast With Trend & Seasonality (2007-2014)')

# Avtual vs Predicted
plotting_data_frame = data.frame(date = test_data$Date,
                                 actual = test_data$Adj.Close, predicted = p2$Point.Forecast)
ggplot(plotting_data_frame, aes(date))+
  geom_line(aes(y = predicted,colour = "Predicted"))+
  geom_line(aes(y = actual,colour = "Actual"))+ ggtitle("With Trend & Seasonality (2015-2016)")

rmse2 = rmse(p2$Point.Forecast,test_data$Adj.Close)
rmse2  #1524.355
```

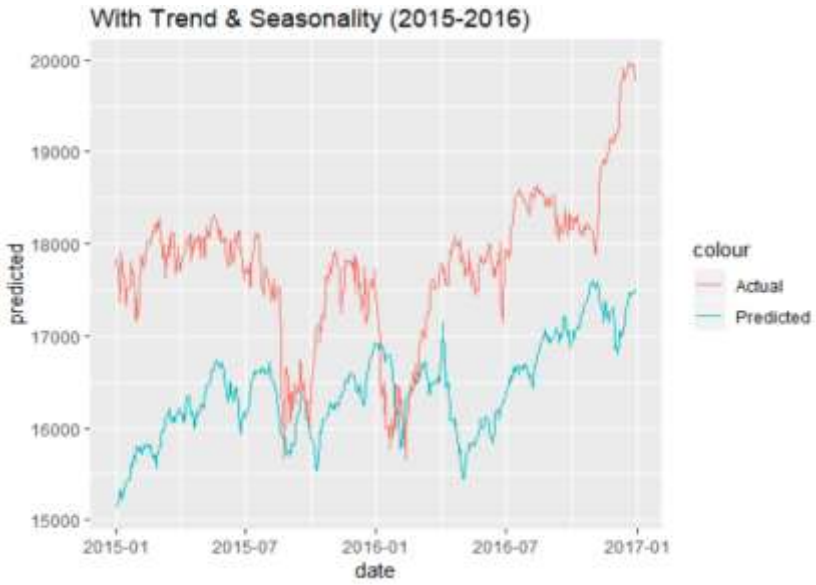


Figure 7

Replacing Holt model by Holt-Winter exponential smoothing, RMSE is reduced from 4000.5 to 1524.23. Also, predicted line is closer to actual line. And their changing pattern is becoming more similar. Besides, predicted line itself could catch small changes seasonally, which means more prediction precision.

2. Sentiment Analysis

1) Without Smoothing:

When the sentiment analysis results that we obtained were used to predict the Dow Jones index, it was evident that the sentiment plays an important role in the stock market movement as our RMSE reduced to 1027.42. from 1524.23 and the average predicted price and the average actual price are in the same range.

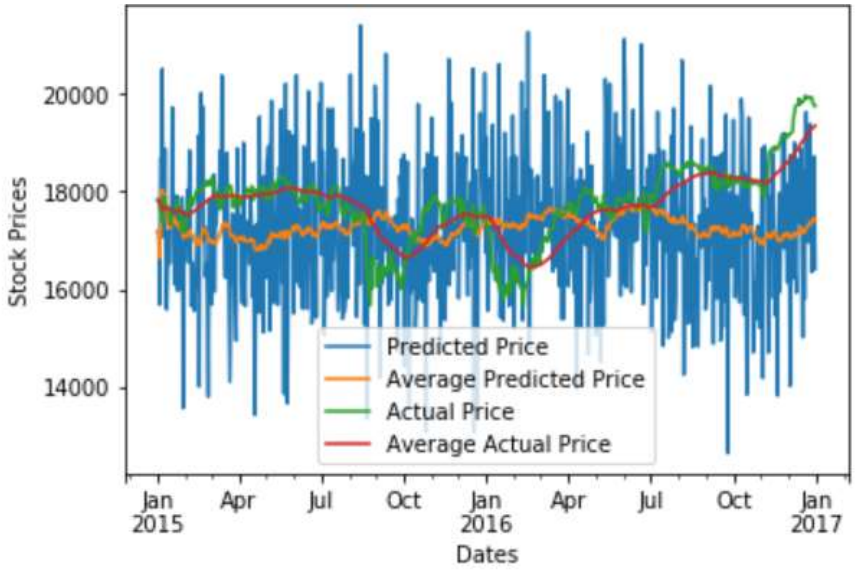


Figure 8

2) With Smoothing:

We further tried to improve the results by using the 7-day moving average to determine the stock prices. After applying the moving average smoothing we could see that the average predicted price line follows the trends of the average actual price line. This could be verified by looking at the RMSE, for this case RMSE was reduced to 770.83, which means that the error in prediction are reduced significantly.

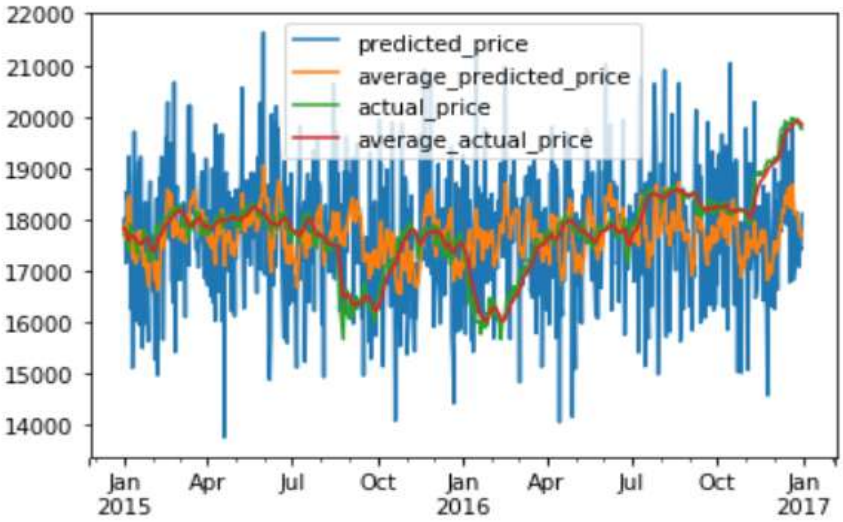


Figure 9

E. Conclusion & Marketing Implications

1. Holt-Winter Prediction Fits Stock Price Well

Stock price is affected by trend and seasonality. When doing Time Series Analysis, Holt-Winter model is a good candidate since it can catch such kind of change precisely. But it is not precise enough, we need to incorporate other techniques

2. Sentiment Analysis Enhances Time Series Analysis

Proved by the performance of our sentiment analysis, this techniques could enhance the prediction of stock price. If only scrapping related headlines, the sentiment analysis will give better prediction

3. Other Important Predictors

In Stock Market Accuracy can be further improved by taking important factors in stock market into tagging schema. For example, use sales revenues after close time as a predictor because that directly causes adjust price changes

4. Sentiment Analysis With Other Techniques

Sentiment analysis can provide compound scores from negative, positive and neural scores. Then this overall sentiment measurement can be used as strong predictors for other techniques, like Decision Trees and Logistic Regression.