Stock Price Prediction Based on Daily News Headlines

Siyi Xu

Project Goal

- Can news headlines improve the prediction of percentage change in stock price

  percentage change in stock price = (adj.close - open) * 100 /open

- Can news headlines be used to predict stock price directional change

  Upward and downward change

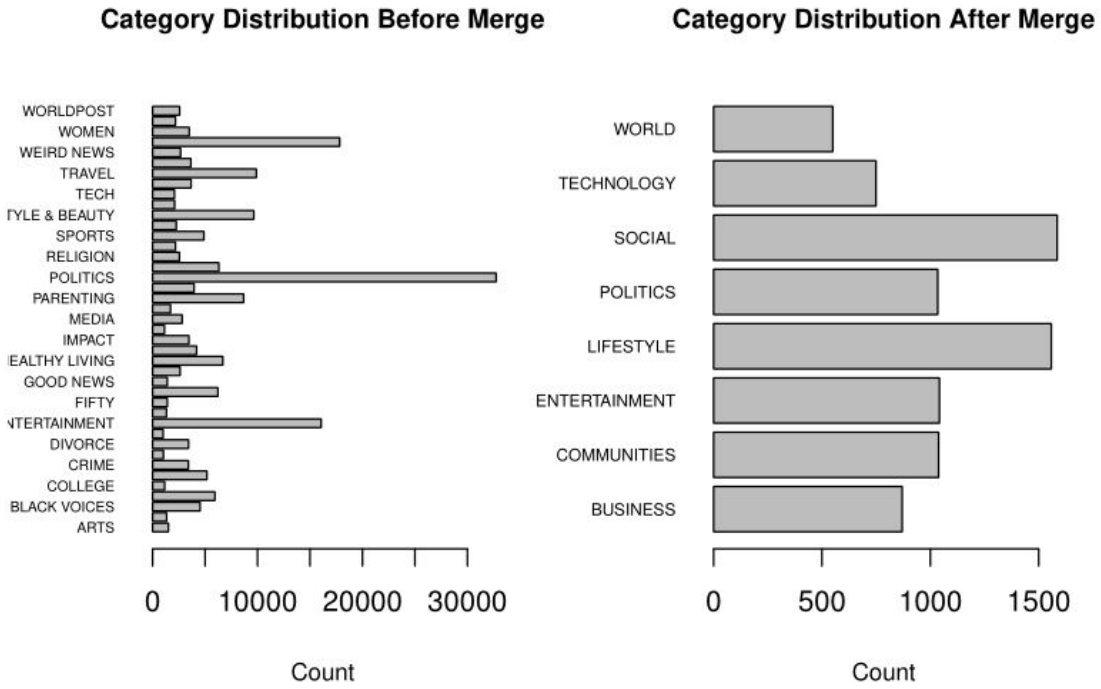
Data

- Two Data Sources
- Dow Jones Industrial Average(DJIA) from Yahoo
  - daily DJIA's open price, high price, low price, close price, adjusted close price, and volumes
  - from 2012-01-28 to 2018-05-26
- 200,000 news headlines dataset from Kaggle
  - from the year 2012 to 2018 obtained
  - Scraped from HuffPost, an American news aggregator and blog
  - With category tags (29 in total)

Data Cleaning

- Merge the Category into 8 based on IPTC NewsCode Taxonomies
  - politics, science & technology, art & entertainment & sport, business, communities, world, lifestyle, and social issues
- Sentiment Analysis
  - Sentimentr Package developed by Tyler Rinker
  - Balanced accuracy and speed
- Algorithm:
  - the words in each sentence searched and compared to a dictionary of polarized words
  - Each polarized word is weighted by the number of the valence shifters
  - Negators (not, can't) flip the sign of a polarized word
  - Amplifiers (absolutely, certainly) increase the weight
  - Negative scores = negative sentiment; positive scores = positive sentiment. 0 is neutral.
- Average the sentiment in each category for every day

Imbalanced Category



Variable Construction

- 0 is imputed for missing value to indicate neturality

Table 1: Description of dataset variables

	Type	Range	Missing
Date	Datetime	Date = 2012-01-28 to 2018-05-26	NA
Price Change in %	Numerical	Minimum: -4.18 ; Maximum:3.89	NA
Label	Categorical	Minimum: 0 ; Maximum:1	NA
BUSINESS	Numerical	Minimum: -2.00; Maximum: 1.05	166
COMMUNITIES	Numerical	Minimum: -0.27; Maximum:0.40	NA
ENTERTAINMENT	Numerical	Minimum: -0.17; Maximum: 0.28	NA
LIFESTYLE	Numerical	Minimum: -0.43; Maximum: 0.58	34
POLITICS	Numerical	Minimum: -0.40; Maximum: 0.97	NA
SOCIAL	Numerical	Minimum: -0.90; Maximum: 0.64	7
TECHNOLOGY & SCIENCE	Numerical	Minimum: -0.87; Maximum: 0.85	234
WORLD	Numerical	Minimum: -0.77; Maximum: 0.44	286

Goal 1: the Predictive Power of News Headlines in Percentage Change

- The ARIMA model (autoregressive integrated moving average)
  - univariate time series forecasting model
  - based on its own past values and the lagged forecast errors
- The ARIMAX model = ARIMA + Covariates

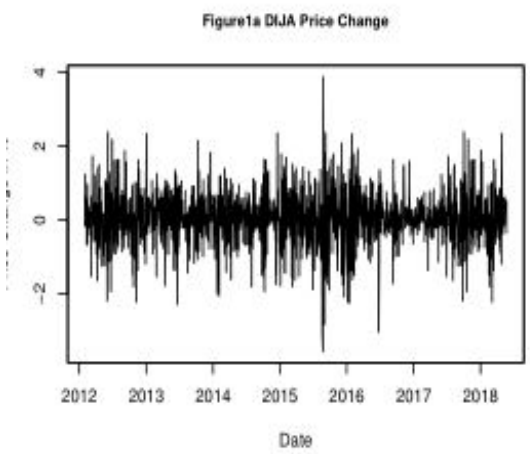
The ARIMA Model is,

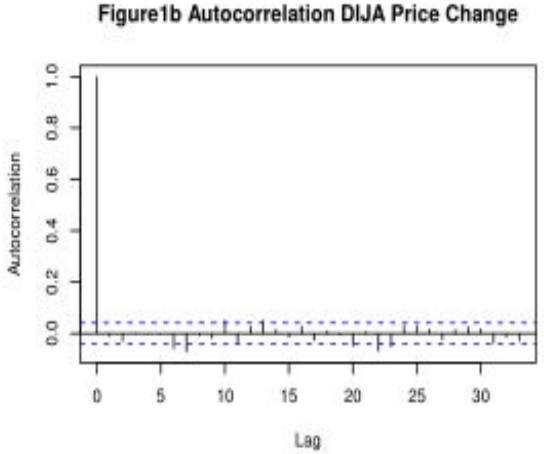
$$Y_t = \alpha + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \phi_p Y_{t-p} - \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} - \theta_q \epsilon_{t-q} + \epsilon_t$$

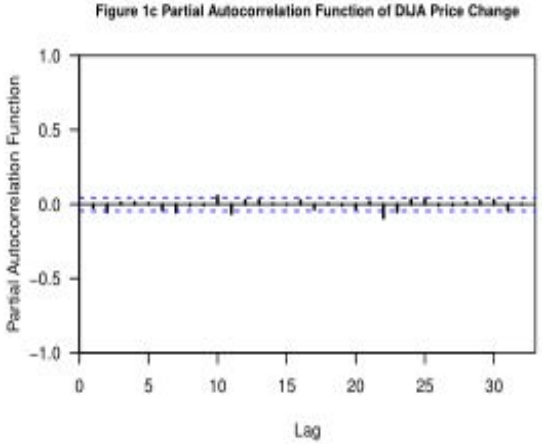
The ARIMAX Model is,

$$\begin{aligned} Y_t =& \alpha + \beta_1 \text{BUSINESS}_t + \beta_2 \text{COMMUNITIES}_t + \beta_3 \text{ENTERTAINMENT}_t + \\ & \beta_4 \text{LIFESTYLE}_t + \beta_5 \text{POLITICS}_t + \beta_6 \text{SOCIA}_t + \beta_7 \text{TECHNOLOGY}_t + \beta_8 \text{WORLD}_t + \\ & \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \phi_p Y_{t-p} - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \theta_q \epsilon_{t-q} + \epsilon_t \end{aligned}$$

Model Selection







Model Diagnostic

Figure3a Residuals Plot for ARIMA(2,0,0)

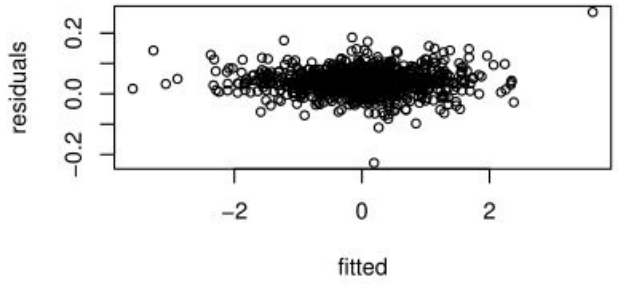


Figure3b Residuals Plot for ARIMAX(2,0,0)

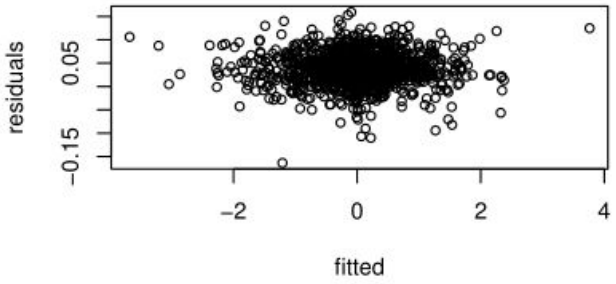


Figure3c Autocorrelation of Residuals of ARIM

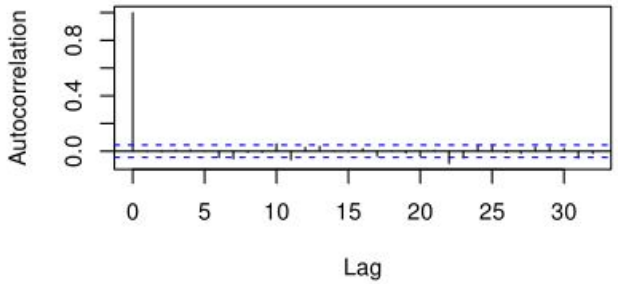
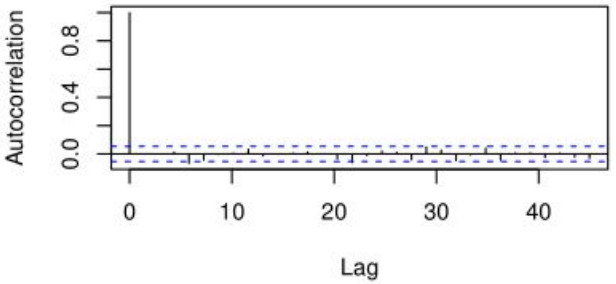


Figure3d Autocorrelation of Residuals of ARIMA



Model Result

The final ARIMA(2,0,0) Model is

$$Y_t = 0.0378 - 0.0257Y_{t-1} - 0.0257Y_{t-2} + \epsilon_t$$

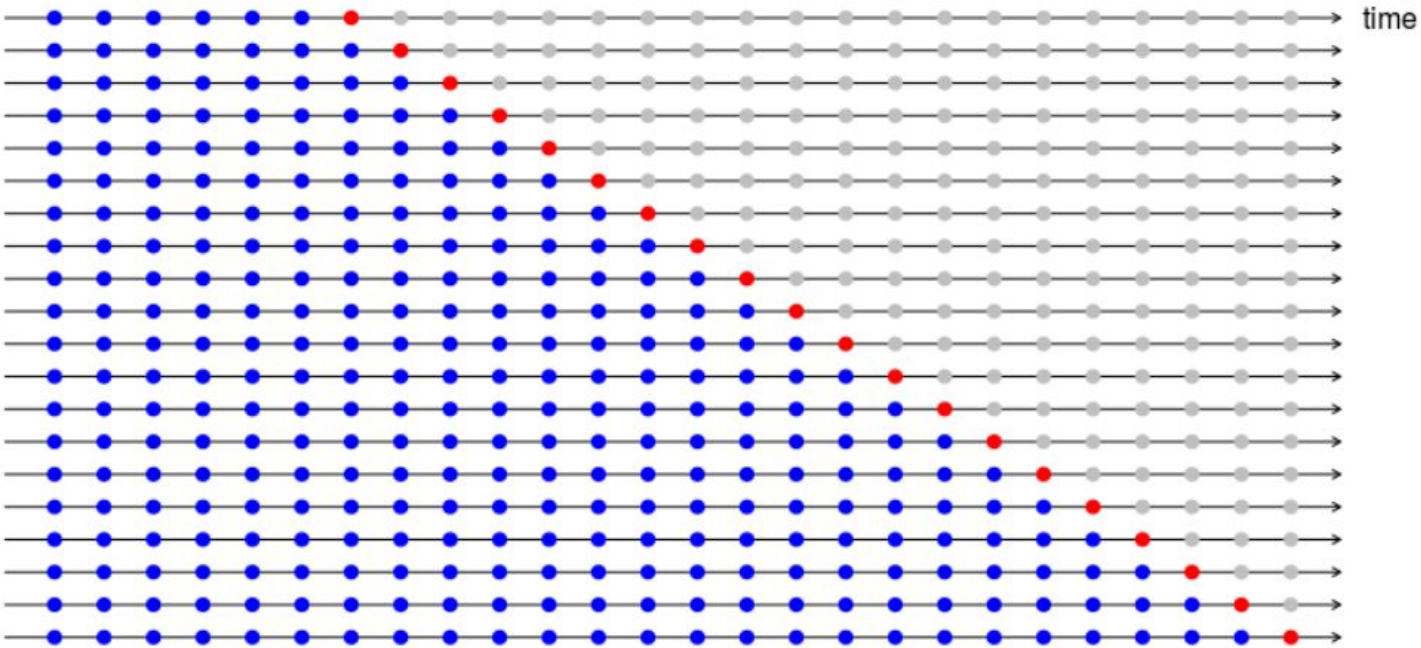
$$\begin{aligned} Y_t =& 0.0107 - 0.0264\text{BUSINESS}_t - 0.1368\text{COMMUNITIES}_t + 0.1835\text{ENTERTAINMENT}_t \\ &+ 0.1809\text{LIFESTYLE}_t - 0.1937\text{POLITICS}_t - 0.0314\text{SOCIA}_t \\ &+ 0.0861\text{TECHNOLOGY}_t - 0.0587\text{WORLD}_t - 0.0232Y_{t-1} - 0.0075Y_{t-2} + \epsilon_t \end{aligned}$$

- Time series Cross Validation is used

	Training RMSE	Cross Validation RMSE	Training AIC
ARIMA	0.7315844	0.737911	2972.73
ARIMAX	0.7321262	0.8006417	2988.91

Table1:The training RMSE, cross validation RMES,training AIC of ARIMA(2,0,0) an ARIMAX(2,0,0)

Time Series Cross Validation



Sensitivity Analysis on Lag Effects of News

- the RMSE do not change significantly for different lag effects
- Lag 3 have lower training RMSE but higher AIC

	Training RMSE	Training AIC
ARIMAX(2,0,0) with Lag0	0.7321262	2988.91
ARIMAX(2,0,0) with Lag1	0.733293	3554.19
ARIMAX(2,0,0) with Lag0,1	0.7325519	3566.97
ARIMAX(2,0,0) with Lag2	0.732716	3551.69
ARIMAX(2,0,0) with Lag3	0.7309783	3544.13

Goal 2:the Predictive Power of News Headlines in Directional Change

- Logistic Regression

$$log(\frac{p_i}{1-p_i}) = 0.07701 - 0.13621\text{BUSINESS}_i - 0.59542\text{COMMUNITIES}_i - 0.19503\text{ENTERTAINMENT}_i + 0.71519\text{LIFESTYLE}_i - 0.52136\text{POLITICS}_i - 0.25442\text{SOCIA}_i + 0.34788\text{TECHNOLOGY}_i + 0.11235\text{WORLD}_i + \epsilon_i$$

- Results
  - 0.5 is set as the decision threshold
  - Modest predictive power
  - Specificity, the proportion of actual downward directional change that are correctly identified, 0.58.
  - Avoid risk vs. Increase Income

	Accuracy	95% CI of Accuracy	Sensitivity	Specificity
Logistic Regression	0.5573	(0.4938, 0.6195)	0.41463	0.58491

Conclusion

- Conclusion
  - News headlines have no predictive power in estimating the values of stock prices change on the same day.
  - Predict the directional change of stock prices with a moderate success rate.
- Limitation:
  - Bias in news (source & financial news)
  - Bias in sentimental analysis
  - Predict on specific stock instead of industrial average