FINANCIAL INDICATORS OF US STOCKS

TEAM 8

DATASET OVERVIEW

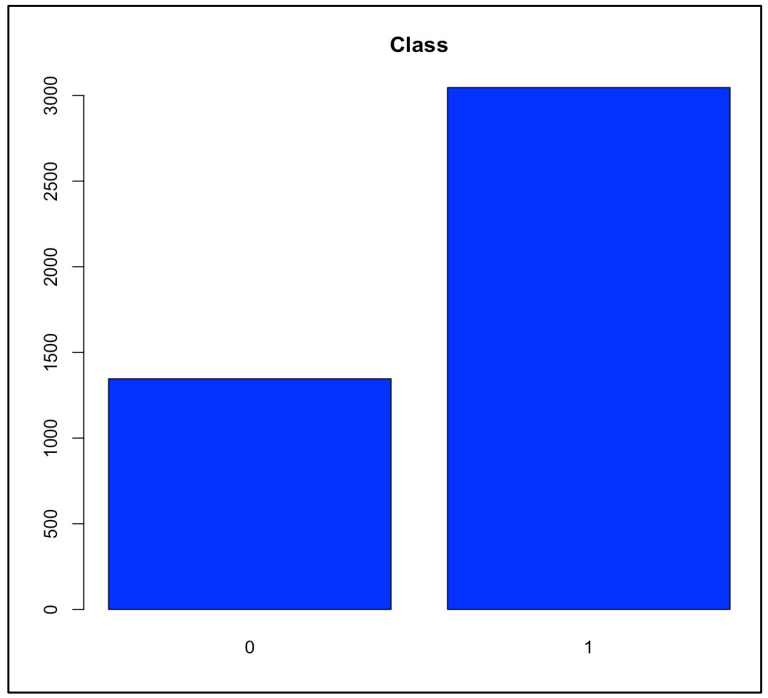
- We selected a dataset from Kaggle, which includes publicly available financial data of 4,392 public companies, from 2014-2018.

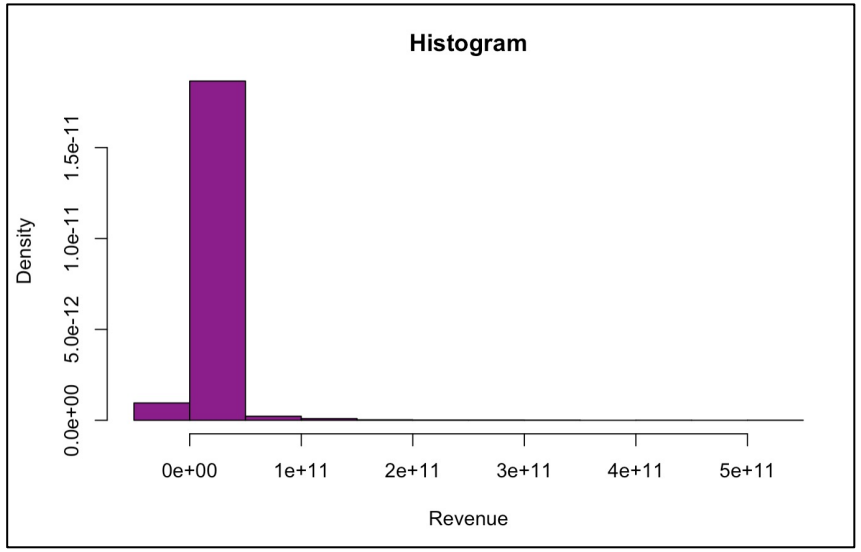
- The dataset includes 224 different variables, across 5 files (one for each year of data: 2014, 2015, 2016, 2017, 2018), totaling ~ 450K data points per year and 2.4MM total data points.

- 222/224 of the variables are financial variables (numeric) with the only other 2 variables being the stock ticker and company sector.

PROJECT GOAL AND EXPLORATORY ANALYSIS

- **Goal:** Predict whether a stock belongs in Class 1 by looking at the financial information released in the 10-K filings
  - Class 1 stocks are stocks that one should buy at the start of the next year (2019) and sell at the end of year (2019).
- **Data Exploration:**
  - A large amount of the rows have NA included in the data repository.
    - If we were to exclude all rows containing NA, there would be 0 rows left to analyze.





ANALYTIC METHODOLOGIES & ETHICAL CONSIDERATIONS

Models Utilized:

- Logistic Regression
- Random Forest
- Gradient Boosting Machine
- Classification and Regressions Trees (CART)
- K-means Clustering
- Hierarchical Clustering
- Time Series Analysis

Ethical Considerations:

- Many financial firms are using ML, AI and big data analytics to gain market advantages trading securities.
- ML programs whose objective is maximizing returns doesn't consider any other impacts, whether environmental, political or socio-economic, that a company undertakes to achieve their financial returns. Prioritizing investments in companies with profit as the end goal doesn't necessarily equate to improving the ethics or health of the larger global economy.
- Models are trained on past information; market volatility presents inherent risks that a ML model may miss, such as broader macro-economic trends or economic shocks.
- The addition of ML programs and large-scale data automation reduces the need of fund managers per client, replacing individual/personal workforce with automated technology and AI.

LOGISTIC REGRESSION

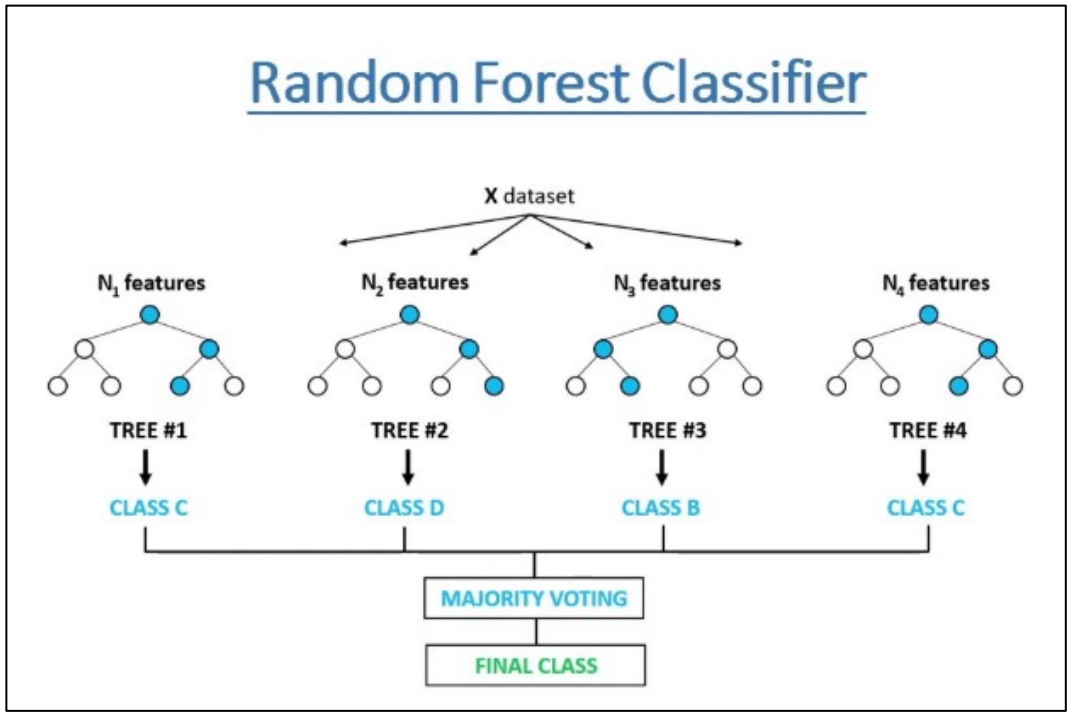
- Following predictor variables were found to be significant (for a 0.05 significance level):
  - PretaxProfitMargin
  - priceSalesRatio
  - Cost.of.Revenue
  - Other.Liabilities
  - Depreciation...Amortization
- Prediction accuracy of 0.78 compared to the accuracy of the baseline model, which was 0.31

Confusion Matrix

	FALSE	TRUE
0	3	130
1	5	473

RANDOM FOREST

- Random Forest model with a goal to predicted "Class" a binary variable
- Na.roughfix was used to input average mean for NA values
- Tested with an array of different model parameters
  - Number of trees
  - Node Size
- Strongest prediction level was 78%, compared to base model's 69%



Confusion Matrix

	FinancePredictForest	
	0	1
0	19	114
1	16	462

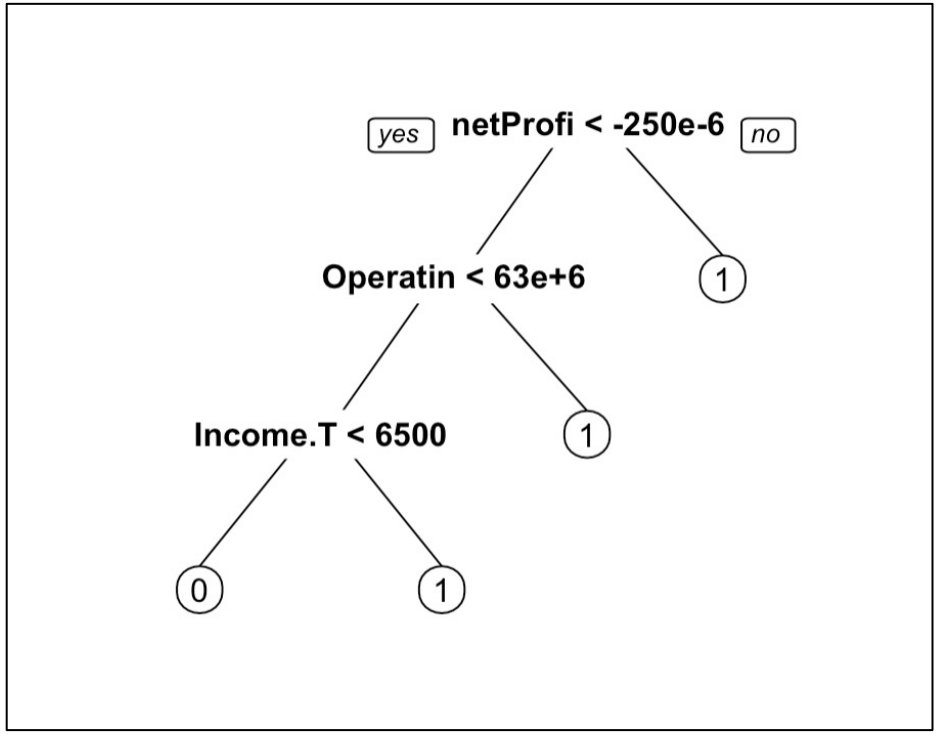
GRADIENT BOOSTING MACHINE

- Model with a goal of predicted "Class" binary variable
- Ran code to find the best iteration of the model on training data
- Best result of the model was able to produce an accuracy of 73% on the testing dataset
- Base model has an accuracy of 69%

	Predicted Response		
Actual Response	0	1	Row Total
0	152	252	404
	0.376	0.624	0.307
1	98	816	914
	0.107	0.893	0.693
Column Total	250	1068	1318

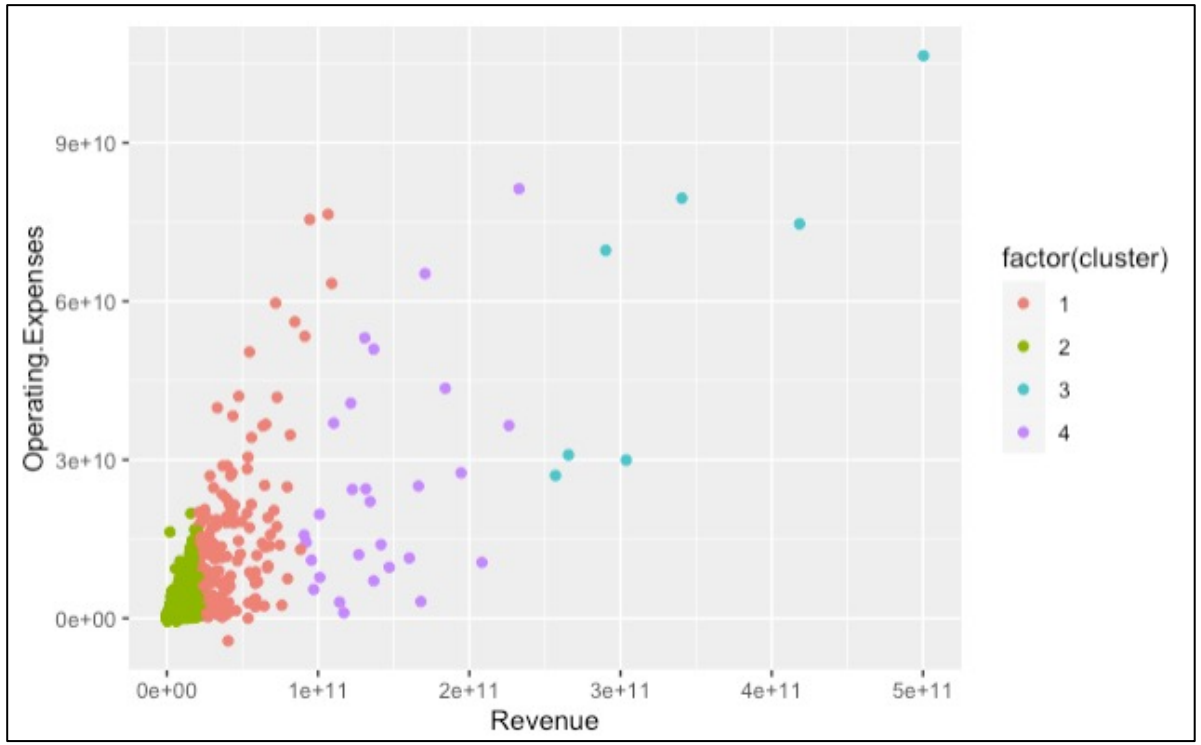
CLASSIFICATION AND REGRESSION TREE (CART)

- The model built could be useful to detect the stock's class by identifying the important financial information in the 10-K filings
- Prediction accuracy of 0.78 compared to baseline accuracy of 0.31
- Variables that appear to be the most significant is net profit margin and operating income. These two variables have the highest values for variable importance, and the higher the value, the more the variable contributes to improving the model.



K-MEANS CLUSTERING

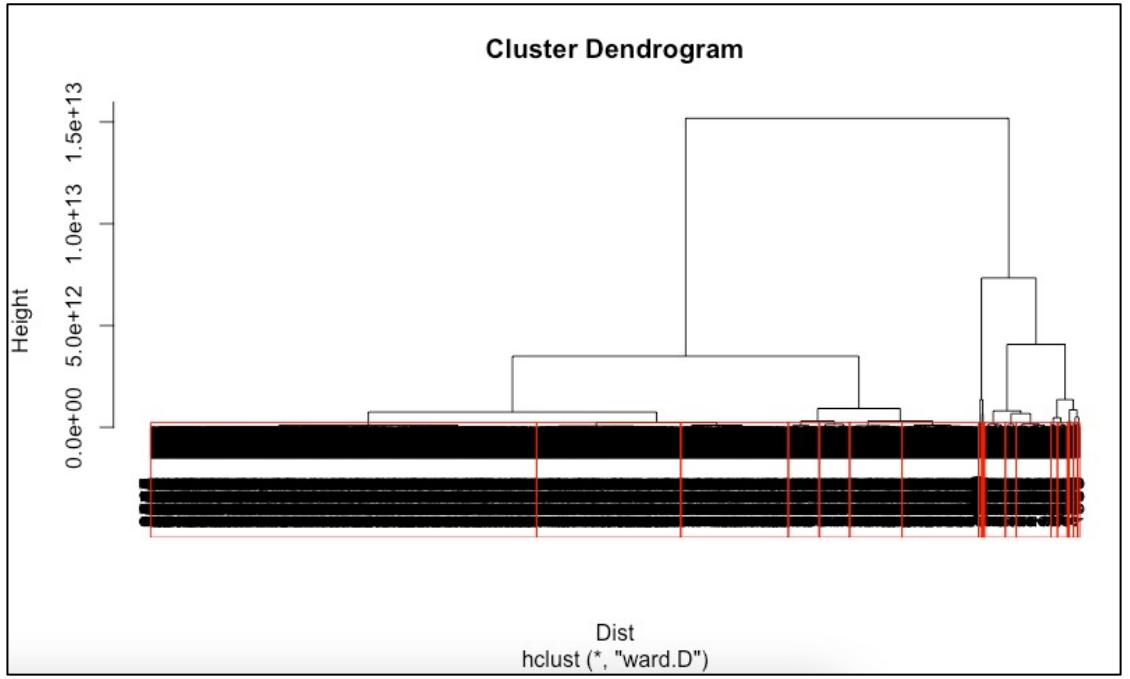
- To see whether each cluster partitioned by K-means method is distinct and perform logit regression on clusters, examining if accuracy will increase.
- We set clusters as 4. Revenue and Operating Expenses might be one of the distinct factor between clusters. Number of stocks in clusters are disproportionate that most of the stocks are in cluster 2.
- After running logit regression on each cluster, we found that prediction and baseline accuracy based on clusters are even lower than those based on the whole dataset. Take cluster 2 for example, prediction accuracy is 0.768 and the baseline accuracy is 0.319
- K-means would not be the preferred model for this dataset as the outcome of clusters are not significantly consistent to that of Class.



Cluster 1	Cluster 2	Cluster 3	Cluster 4
166	3801	7	28

HIERARCHICAL CLUSTERING

- From cluster dendrogram, we selected 20 clusters to shorten the distances between points
- A large proportion of the data is found in Cluster 15, which has the lowest average revenue among all the clusters



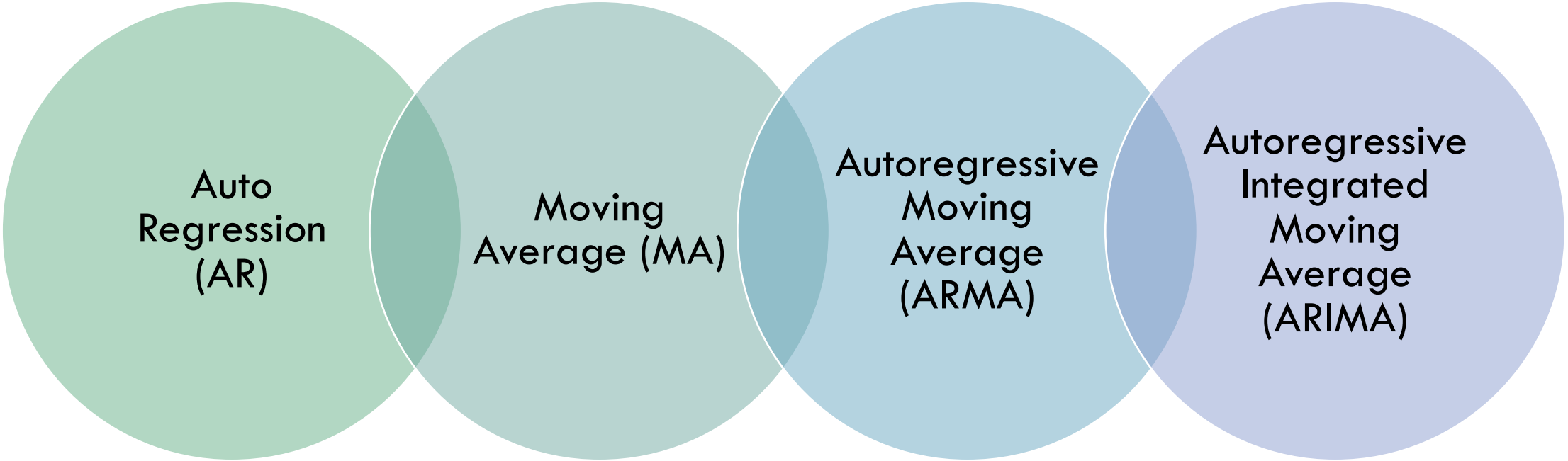
Average revenue for each cluster:

1	2	3	4	5	6	7	8	9
114986202766	12727756330	62867251180	35527228839	140892571075	239265946093	6715946565	63496612031	7000428778
10	11	12	13	14	15	16	17	18
1290452379	41003133921	2366613776	17732101860	22032236184	99850494	3945333046	590649266	93234772537
19	20							
459374032562	322172039074							

- The difference between rest of the clusters is not distinct. Hence, hierarchical clustering would not be a very intuitive method for this dataset

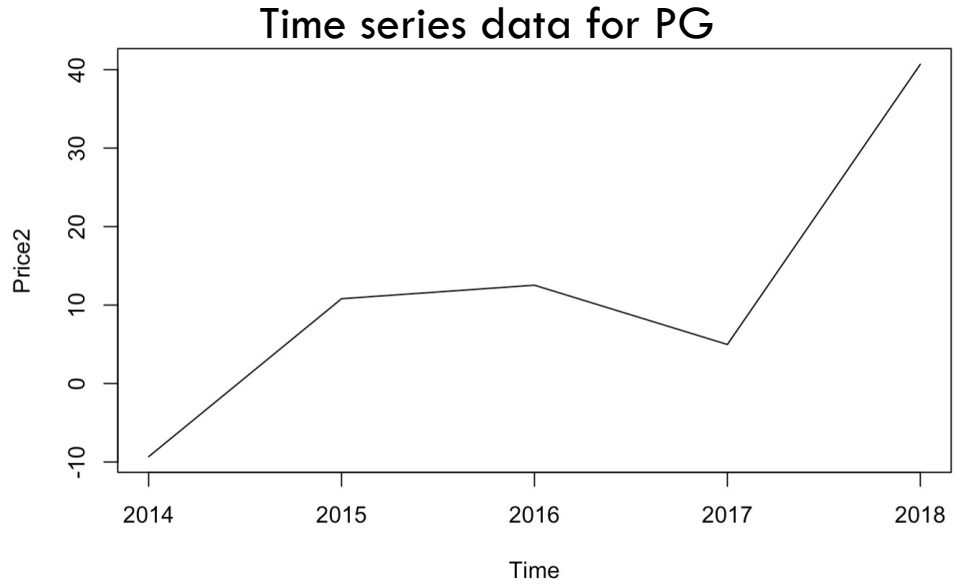
Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
9	149	18	43	14	7	131	18	133	463
Cluster 11	Cluster 12	Cluster 13	Cluster 14	Cluster 15	Cluster 16	Cluster 17	Cluster 18	Cluster 19	Cluster 20
28	330	47	90	1663	226	621	8	2	2

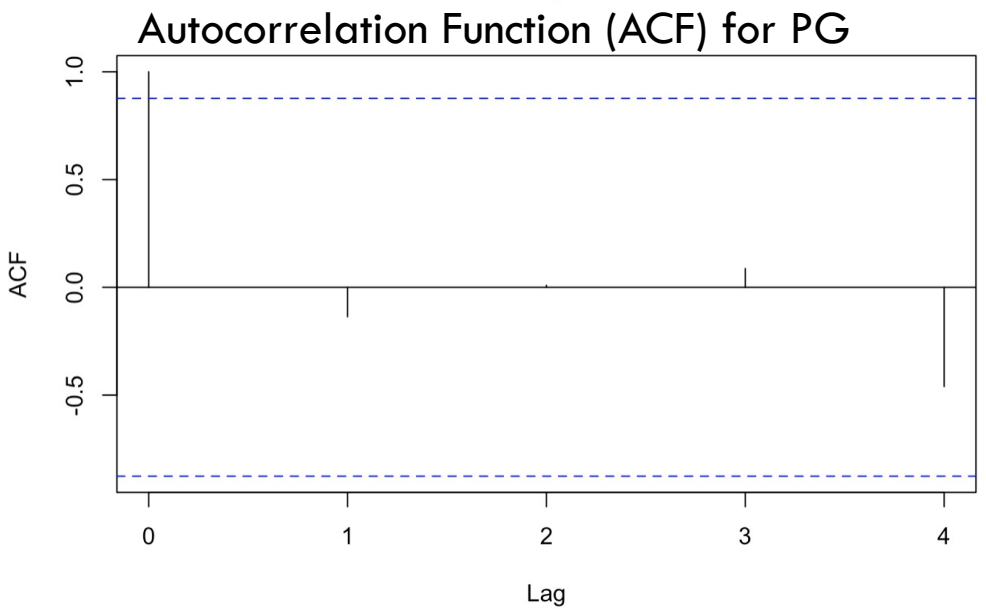
TIME SERIES ANALYSIS- MODELS



TIME SERIES ANALYSIS

- Data Characteristics:
  - Univariate → Price Variance (2014-2018)
  - Independent Variable → Time
  - Stationary – could be (not seasonal but can't verify trend)
  - Not Seasonal
  - Single Stock – PG
- Autocorrelation function (ACF):
  - Measure of correlation between observations
  - Degree of similarity between time series & lagged version of itself over successive time intervals
  - Blue dotted lines – point of statistical significance
  - Residuals – no significant autocorrelations



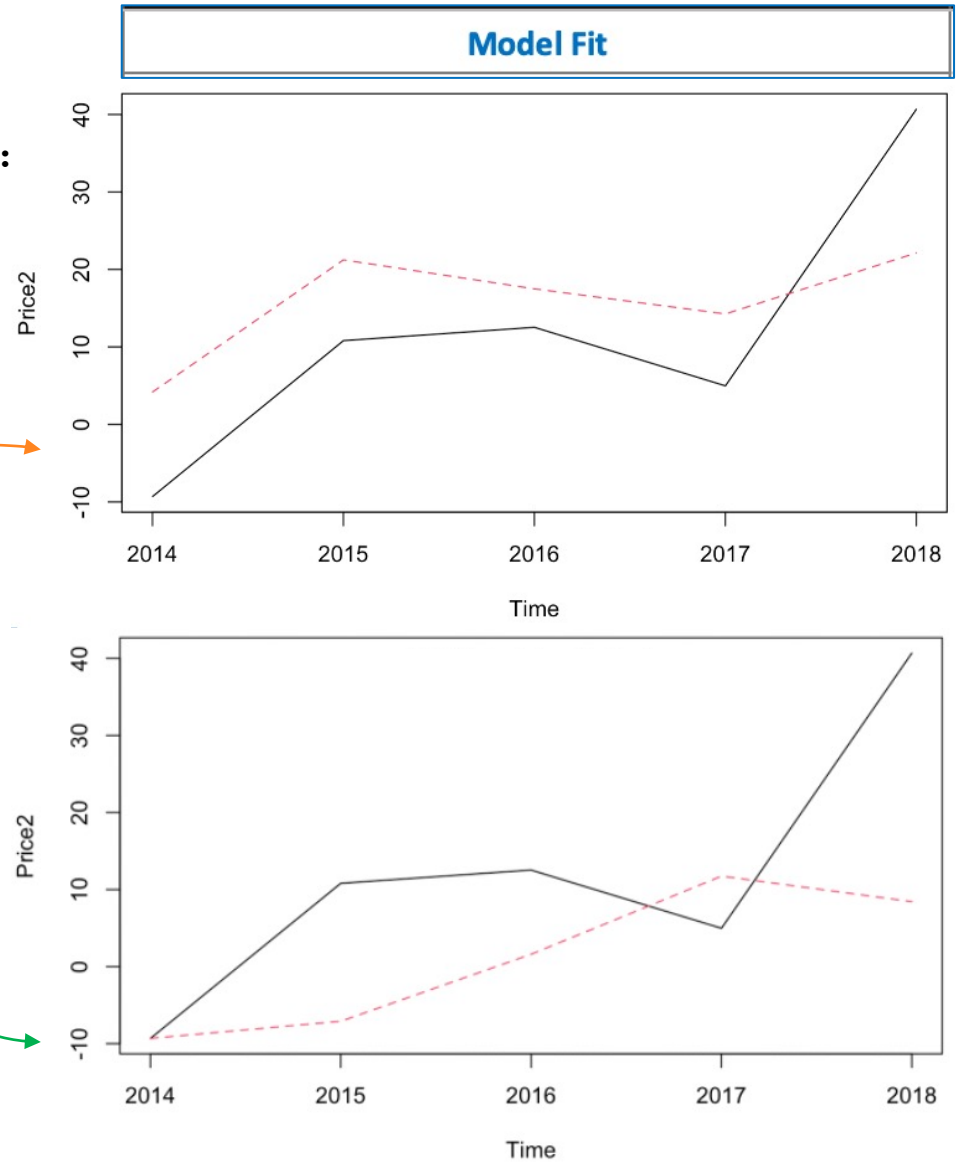


FINDINGS

Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC):

- Measures of model fit
- Lower the better

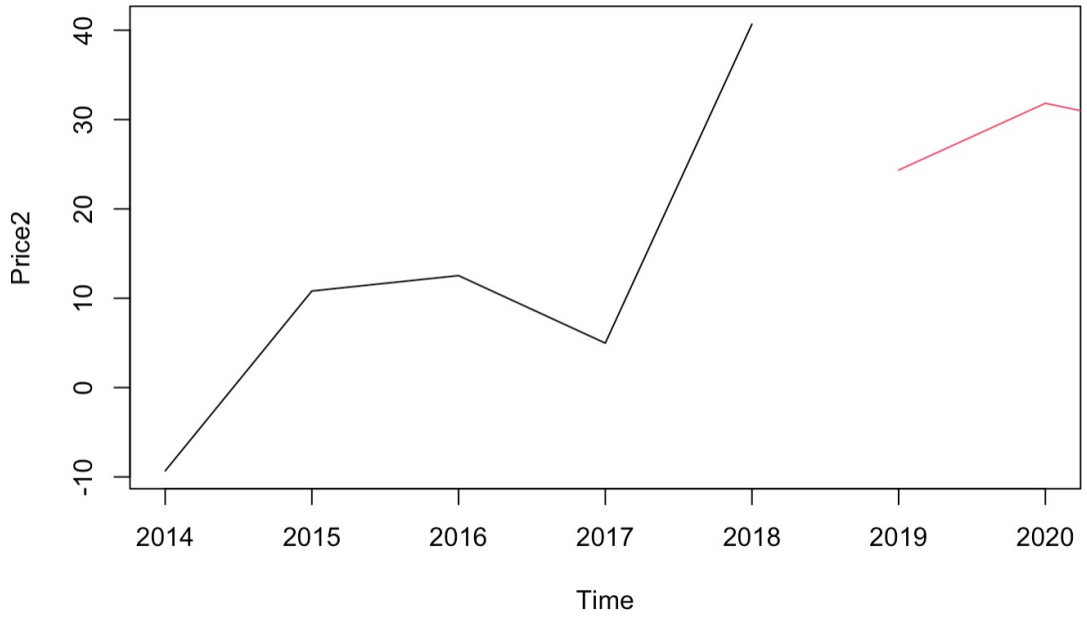
Stock	Model	AIC	BIC
PG	Autoregression (AR)	47.7	46.5
PG	Moving Average (MA)	47.2	46
PG	Autoregressive Moving Average (ARMA)	51.1	49.2
PG	Autoregressive Integrated Moving Average (ARIMA)	39.3	38.1



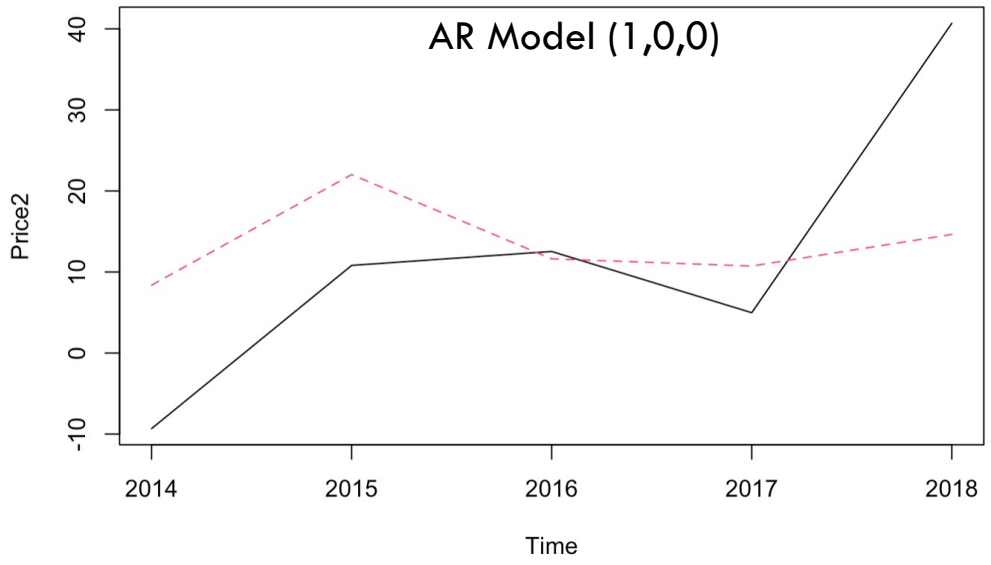
ARIMA > ARMA possibly due to the trend component in stationarity

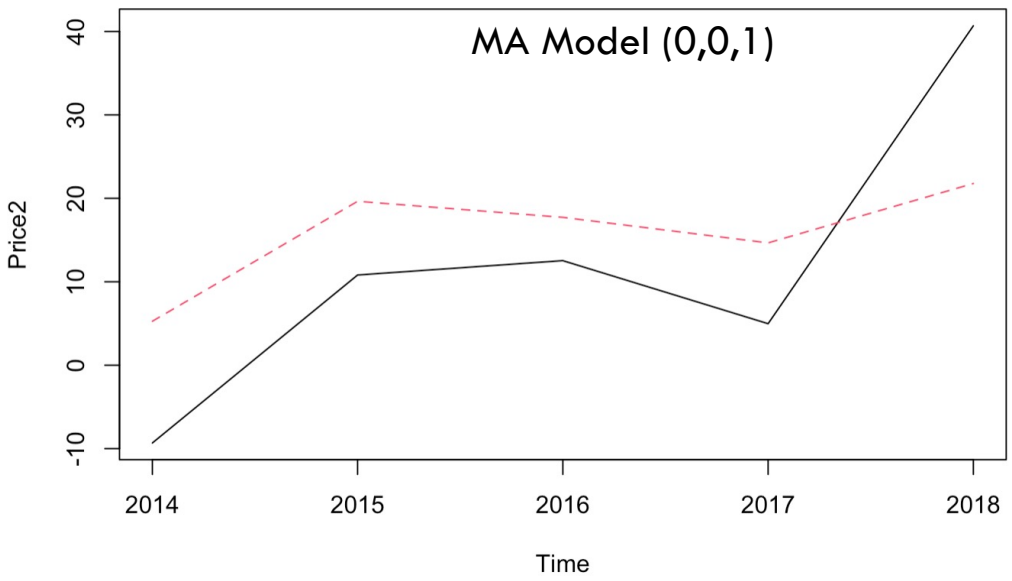
FORECASTING & INACCURACIES

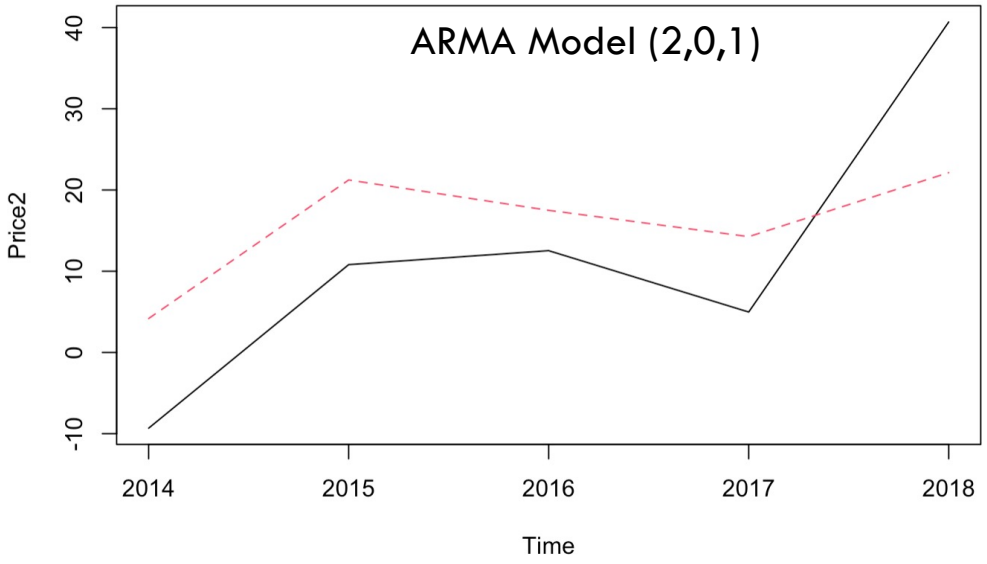
- Red – predicted forecast
- Gap in actual and forecasted data – yearly data
- Limited historical yearly data (5years). Average for high accuracy – 25 observations
- Frequency = 1. Monthly/Daily/Weekly data would have a higher frequency and better predictions
- Seasonal data – could use the SARIMA model (performs the same autoregression, moving average and differencing modeling at the seasonal level)
- Neural Network - could utilize multiple features instead of just the price variance over the years



APPENDIX







(p, d, q):

p → number of autoregressive terms
d → number of nonseasonal differences
q → number of moving-average terms

APPENDIX- ARIMA MODELS (Based on different p,d,q parameter values)