STA 9890 Final Project

"Bitcoin time-series analysis through other stocks"

Carlos Scuderi

May 19th, 2020

Baruch COLLEGE
The City University of New York

4.a) Nature of the Data

STA_9890_Final_Project_Scuderi

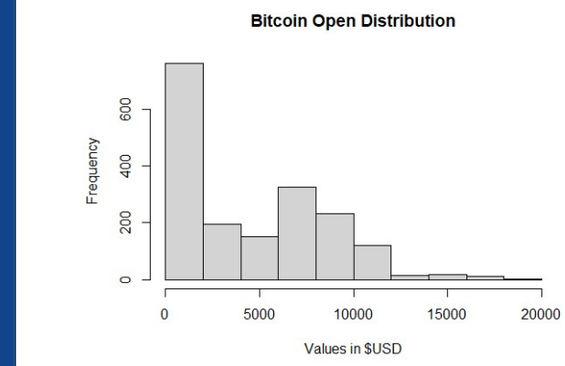
Project Description

The aim of this data science project is to analyze how different variables may impact the price of Bitcoin at market opening.

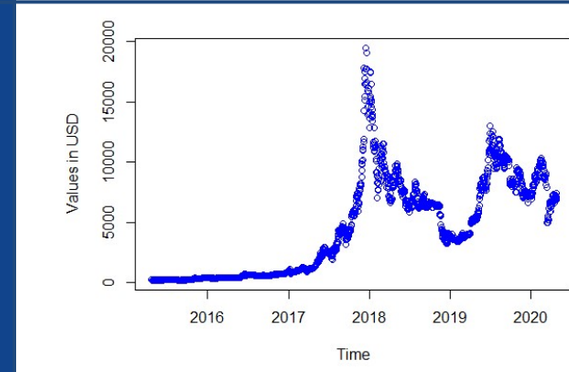
Variable	Role	Type	Description
Y_Open-Bitcoin	Response (Y)	FLOAT	Bitcoin price of the stock when the market opened on this day
P1_Open-Google	Predictor	FLOAT	7-day time series (t-1 to t-8) of Google Open stock price
P2_Volume-Google	Predictor	INTEGER	7-day time series (t-1 to t-8) of Google Volume (Number of shares traded)
P3_Open-IBM	Predictor	FLOAT	7-day time series (t-1 to t-8) of IBM Open stock price
P4_Volume-IBM	Predictor	INTEGER	7-day time series (t-1 to t-8) of IBM Volume
P5_Open-Shell Oil	Predictor	FLOAT	7-day time series (t-1 to t-8) of Shell Oil Open stock price
P6_Volume-Shell Oil	Predictor	INTEGER	7-day time series (t-1 to t-8) of Shell Oil Volume
P7_Open-General Motors	Predictor	FLOAT	7-day time series (t-1 to t-8) of GM Open stock price
P8_Volume-General Motors	Predictor	INTEGER	7-day time series of Shell Oil Volume (Number of shares traded))

n = 1820
p = 56

Histogram (y)

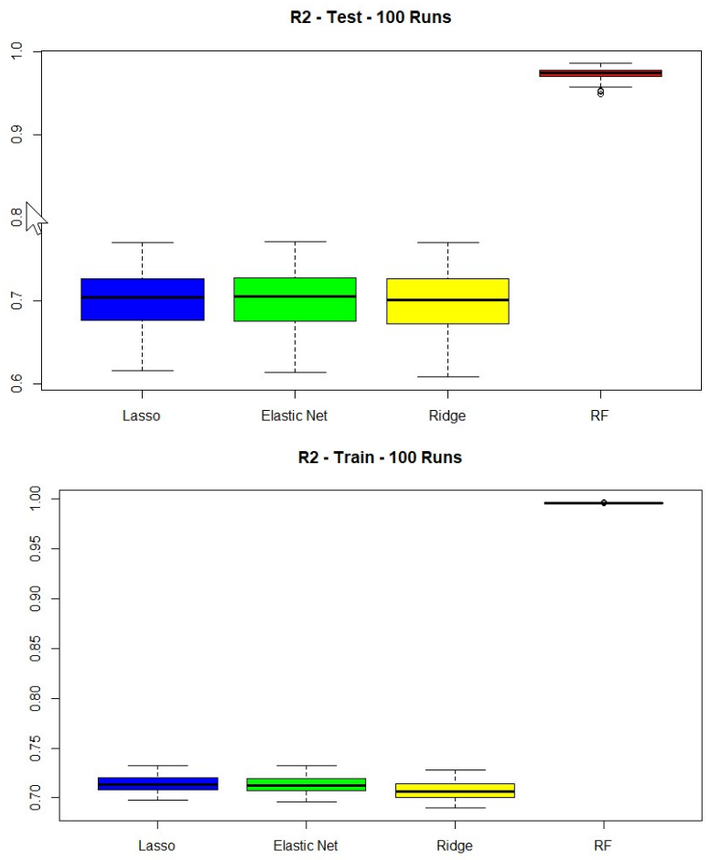


Time Series (y)

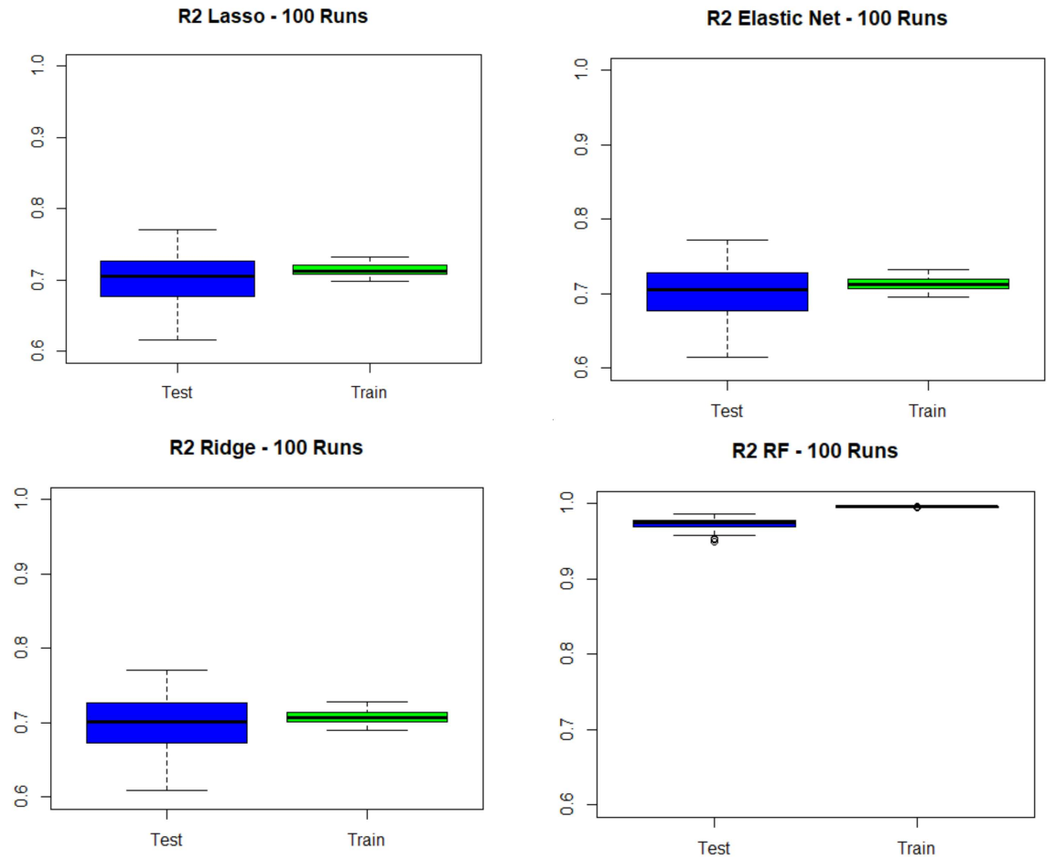


4.b) Side by Side Boxplot Analysis

R2 Boxplots Lasso, Elastic Net, Ridge, Random Forests Grouped by (Test and Train)



R2 Boxplots for Lasso, Elastic Net, Ridge, Random Forests Grouped by Model



Median R2 Test

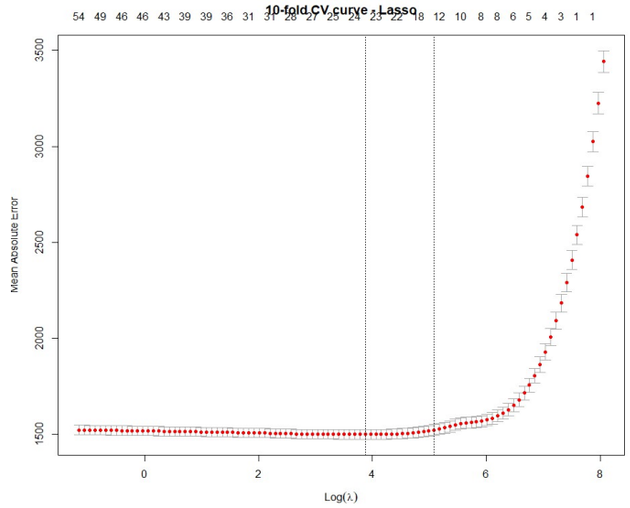
- Lasso: 0.705
- EN: 0.706
- Ridge: 0.705
- RF: 0.97

Median R2 Train

- Lasso: 0.713
- EN: 0.712
- Ridge: 0.706
- RF: 0.99

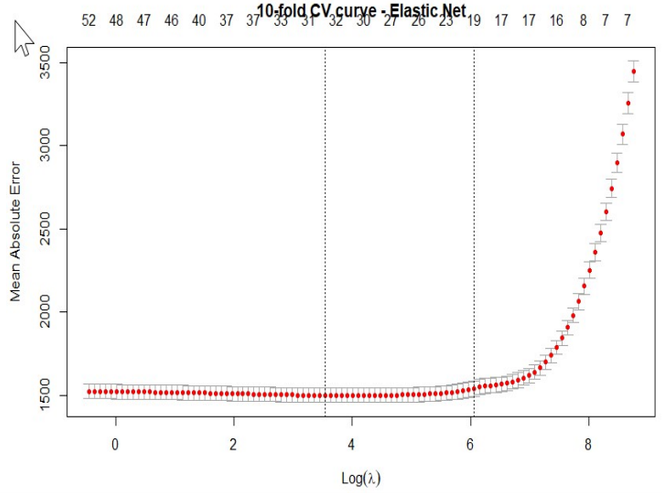
4.c) 10-Fold CV Curves: Lasso, Elastic Net, Ridge

10-Fold CV – Lasso Regression



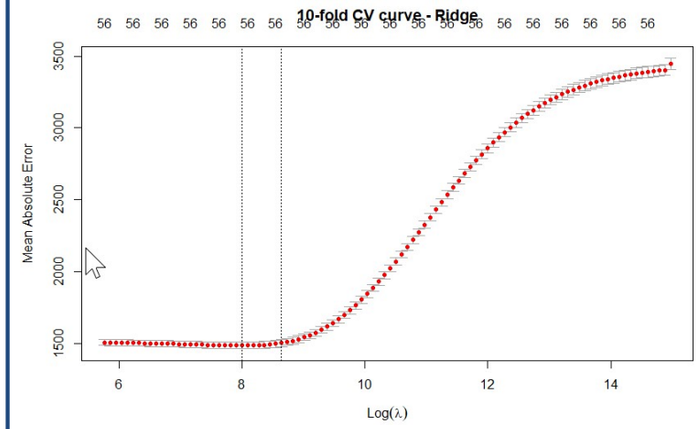
λmin: 48.25 λ1sd-rule: 161.72 Predictors: 24

10-Fold CV – Elastic Net Model



λmin: 34.68 λ1sd-rule: 427.57 Predictors: 32

10-Fold CV – Ridge Model

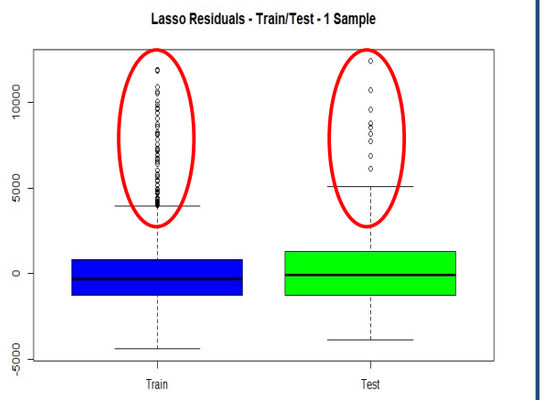


λmin: 2960.713 λ1sd-rule: 5678.38 Predictors: 56

Lasso / Elastic Net are performing variable selection taking 24/32 predictors. Still too many for model interpretation
Considering Lambda 1-standard rule could be a good option => Reduction in predictors without too much CV error increase
High Lambda values in Ridge could be an indication of the model trying to reduce many non-significant coefficients

4.d) Side-by-Side Boxplots for Residuals

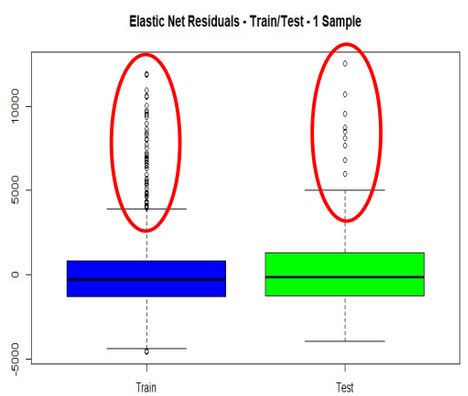
Lasso Regression Boxplots of Residuals



Max Res = 11910.33
Median = -315.19
Min = -4417.44

Max Res = 12414.72
Median = -107.40
Min Res = -3897.71

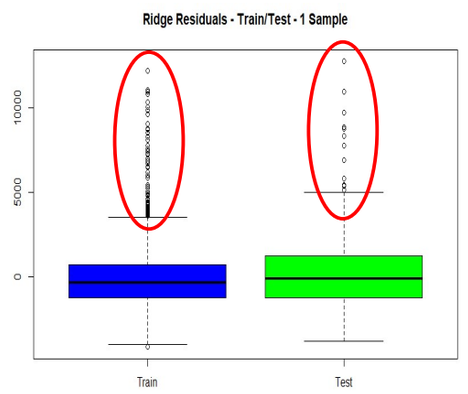
Elastic Net Model Boxplots of Residuals



Max Res = 11899.49
Median = -293.3
Min = -4577.73

Max Res = 12512.9
Median = -118.8
Min Res = -3951.25

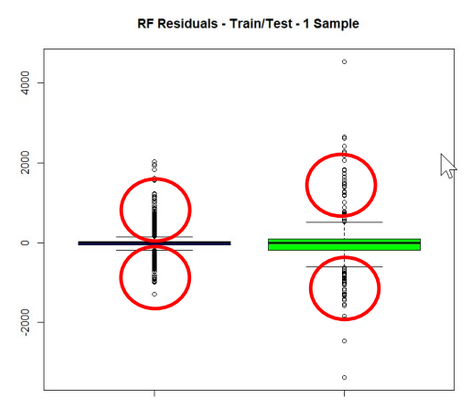
Ridge Model Boxplots of Residuals



Max Res = 12222.01
Median = -367.29
Min = -4191.24

Max Res = 12756.8
Median = -121.41
Min Res = -3859.33

Random Forests Model Boxplots of Residuals

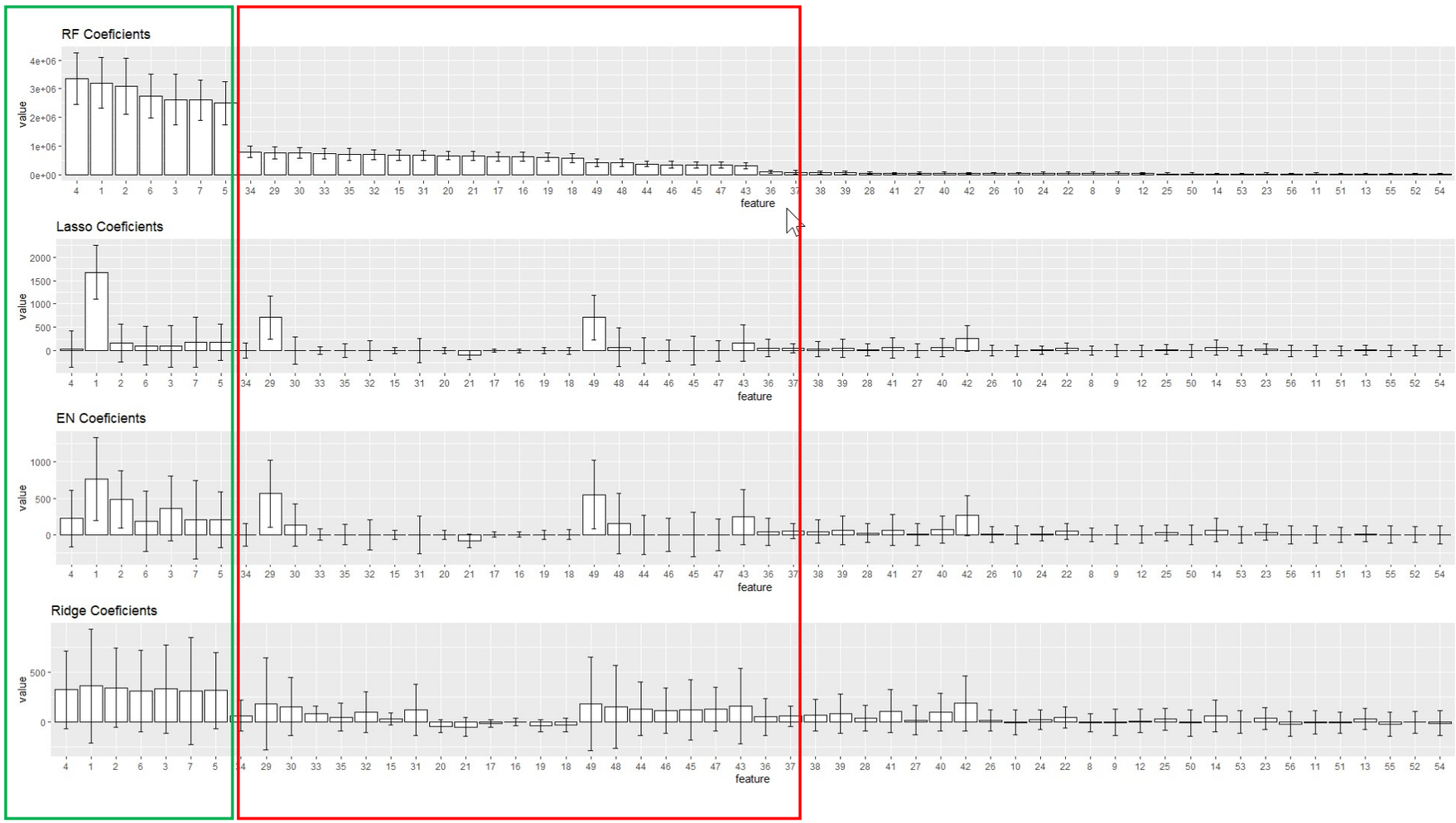


Max Res = 4534.32
Median = -8.16
Min = -3372.46

Max Res = 2026.417
Median = -3.64
Min Res = -1294.33

Linear models are failing to estimate Bitcoin Open due to variability (non-linear) in certain periods (little difference in residual performance)
Especially during peak times, regularization models consistently heavily undershoots (i.e. y.hat <<< y)
Random Forests show smaller residuals (=> higher R2), with outliers evenly distributed.

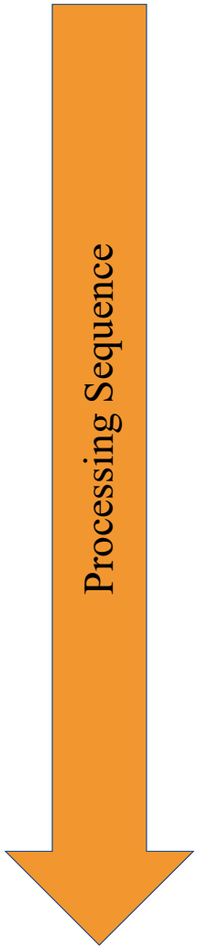
4.e) Estimated Coefficients / Importance



Summary

- **Green Square:** Models show a similar trend in identifying P1 – P7 as most relevant predictors (Google Open)
- **Red Square:** Many other predictors are not consistent across the models.
- Random Forests seem to do a better job at estimating given the less amount of coefficient variability (Box-plot error bars)
- Variability too much for Regularization. Models failing to find statistically relevant coefficients.
- Note: Adding Bitcoin (t-1..t-7) as regressor substantially improves R2 performance. Also if applying Log(y).

4.f) Performance and Time to Train



Description	Processing Time (Seconds)	Model Performance (Mean R2-Test)
Load, Standardize Data	0.062s	-
Running 100 Loops with Lasso, EN, Ridge, RF	821.97s (13.69min)	-
Boxplots for R2 Train / R2 Test	0.092s	-
10-Fold CV Curves for Lasso, Elastic, and Ridge	0.511s	-
Side-by-side boxplots of train/test residuals	8.04s	-
Bar-plots (100 Bootstrap resamples)	1074.38s (17.91min)	-
Time-to-Train: Random Forests	9.237s	0.973
Time-to-Train: Lasso	0.136s	0.7027
Time-to-Train: Elastic Net	0.141s	0.7025
Time-to-Train: Ridge	0.141s	0.698
Drawing Final Plots (e.g. Coefficients / Importance)	1.503s	
Total Processing Time	1916.22s (31.93min)	

Summary

- Random Forests show better performance but are computationally intensive (68x more resources than Lasso)
- Running loops to confirm relevance of R2 and other variables is good but also CPU intensive (43% of total running time)
- Bootstrap was also helpful to rule out not-significant coefficients but is also costly (56% of total running time)

Backup Slide - Reference Table of Predictors

Table of Predictors							
P1	Open-Google (Tau - 1)	P15	Open-IBM (Tau - 1)	P29	Open-Shell Oil (Tau - 1)	P43	Open-General Motors (Tau - 1)
P2	Open-Google(Tau - 2)	P16	Open-IBM(Tau - 2)	P30	Open-Shell Oil(Tau - 2)	P44	Open-General Motors(Tau - 2)
P3	Open-Google(Tau - 3)	P17	Open-IBM(Tau - 3)	P31	Open-Shell Oil(Tau - 3)	P45	Open-General Motors(Tau - 3)
P4	Open-Google(Tau - 4)	P18	Open-IBM(Tau - 4)	P32	Open-Shell Oil(Tau - 4)	P46	Open-General Motors(Tau - 4)
P5	Open-Google(Tau - 5)	P19	Open-IBM(Tau - 5)	P33	Open-Shell Oil(Tau - 5)	P47	Open-General Motors(Tau - 5)
P6	Open-Google(Tau - 6)	P20	Open-IBM(Tau - 6)	P34	Open-Shell Oil(Tau - 6)	P48	Open-General Motors(Tau - 6)
P7	Open-Google(Tau - 7)	P21	Open-IBM(Tau - 7)	P35	Open-Shell Oil(Tau - 7)	P49	Open-General Motors(Tau - 7)
P8	Volume-Google (Tau - 1)	P22	Volume-IBM (Tau - 1)	P36	Volume-Shell Oil (Tau - 1)	P50	Volume-General Motors (Tau - 1)
P9	Volume-Google(Tau - 2)	P23	Volume-IBM(Tau - 2)	P37	Volume-Shell Oil(Tau - 2)	P51	Volume-General Motors(Tau - 2)
P10	Volume-Google(Tau - 3)	P24	Volume-IBM(Tau - 3)	P38	Volume-Shell Oil(Tau - 3)	P52	Volume-General Motors(Tau - 3)
P11	Volume-Google(Tau - 4)	P25	Volume-IBM(Tau - 4)	P39	Volume-Shell Oil(Tau - 4)	P53	Volume-General Motors(Tau - 4)
P12	Volume-Google(Tau - 5)	P26	Volume-IBM(Tau - 5)	P40	Volume-Shell Oil(Tau - 5)	P54	Volume-General Motors(Tau - 5)
P13	Volume-Google(Tau - 6)	P27	Volume-IBM(Tau - 6)	P41	Volume-Shell Oil(Tau - 6)	P55	Volume-General Motors(Tau - 6)
P14	Volume-Google(Tau - 7)	P28	Volume-IBM(Tau - 7)	P42	Volume-Shell Oil(Tau - 7)	P56	Volume-General Motors(Tau - 7)

Thank you for your attention!

Baruch COLLEGE
The City University of New York