Forecasting Of Bitcoin Price

Javier Bernal
Manoj Nagabandi
Carmen Ortiz

OUTLINE

1. Introduction
2. Data Overview (Yahoo Dataset)
3. Pre-processing
4. Exploratory Data Analysis
5. Time Series Models
   a. BM, GBM, GGM
   b. Time Series Linear Model with Trend
   c. ARIMA models
6. Dataset (Indicators + Macroeconomic Variables)
7. Pre-processing and EDA
8. Different types of feature selection:
   a. Generalized Additive Regression
   b. Stepwise Regression
   c. Gradient Boosting Model
9. Forecasting Models
   a. Stepwise Regression
   b. Generalized Additive Regression
   c. Classification and Regression Trees
   d. Gradient Boosting Model
10. Comparison of Best Models
11. Conclusion

INTRODUCTION

Created in 2009, Bitcoin is the world's first decentralized digital currency that allows peer-to-peer transactions without the need for intermediaries like banks on a blockchain network.

There are a number of reasons why forecasting the price of BTC can be important:

1)Investment

2)Business

3)Innovation

DATASET

The variables considered are:

- Date
- Open
- High
- Close
- Volume

1. For Initial analysis BTCUSD dataset is taken from Yahoo Finance website.

2. Time period considered initially 01/12/2017 – 30/10/2022 (4 years 11 months)

3. Closing Price is considered as the price to be predicted while forecasting

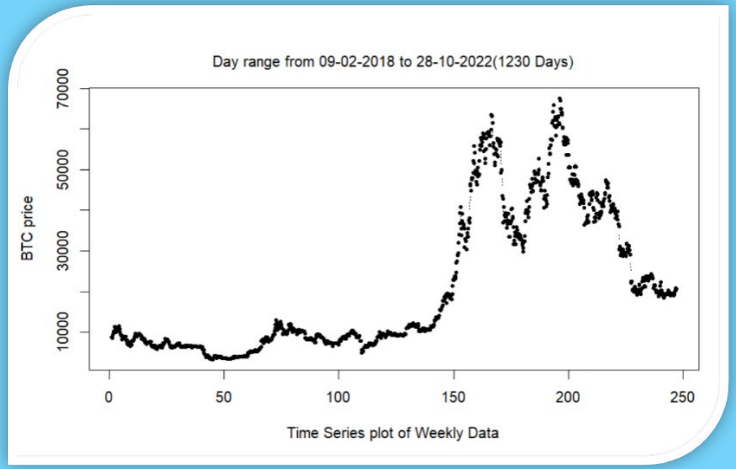
Pre-processing

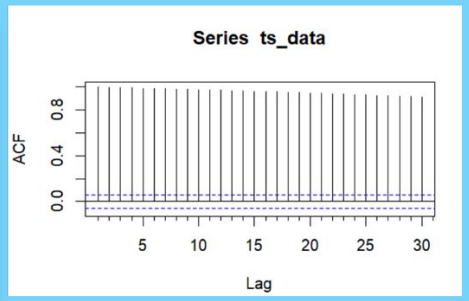
- Only 5 days Mon-Fri are considered as there is no much market movement during the weekends.

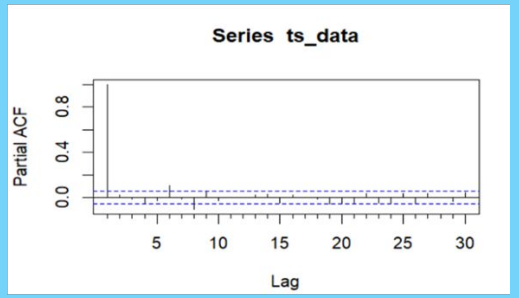
- We have 50 NA values in the considered range of 01/12/2017 - 30/10/2022 (4 years 11 months) so the rows are removed.

- The final day range is 07/02/2018 - 30/10/2022

EXPLORATORY DATA ANALYSIS





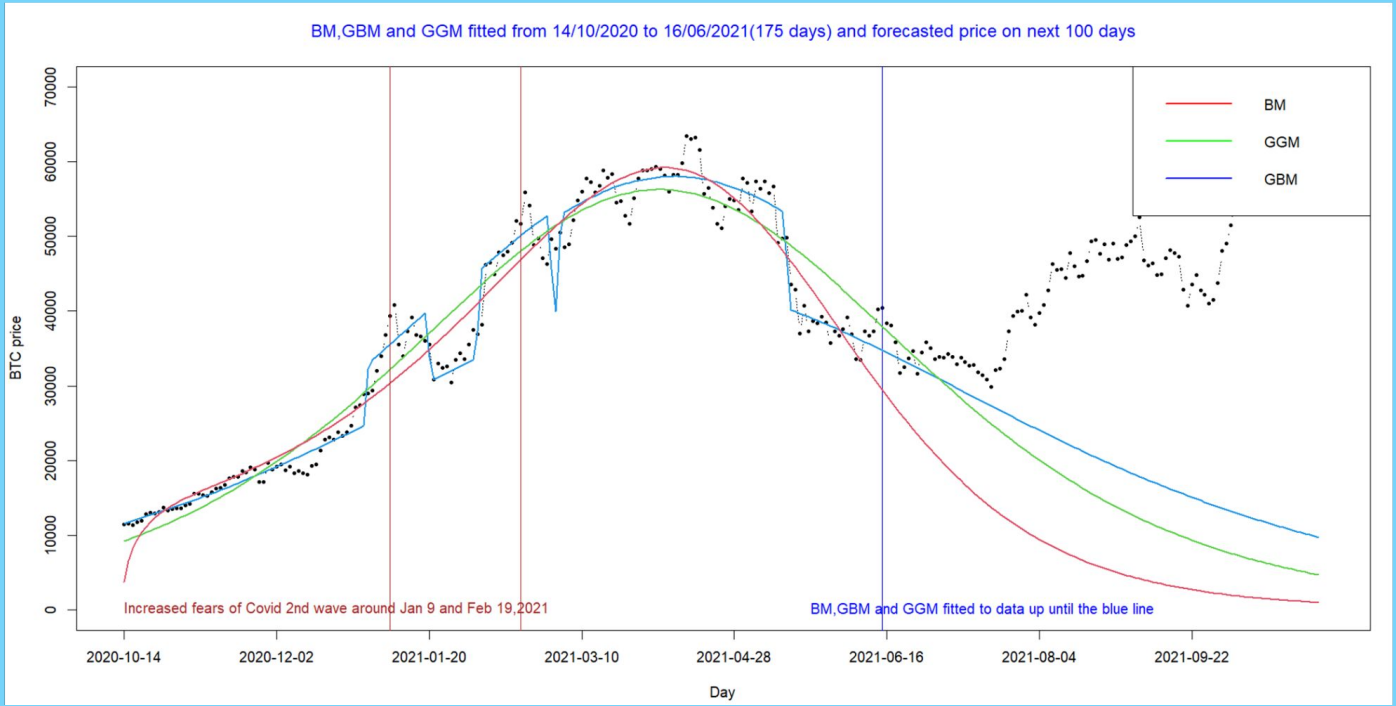



```
        Augmented Dickey-Fuller Test

data:  ts_data
Dickey-Fuller = -1.5127, Lag order = 10, p-value = 0.7846
alternative hypothesis: stationary
```

From ACF, PACF Correlograms and ADF test we can observe that

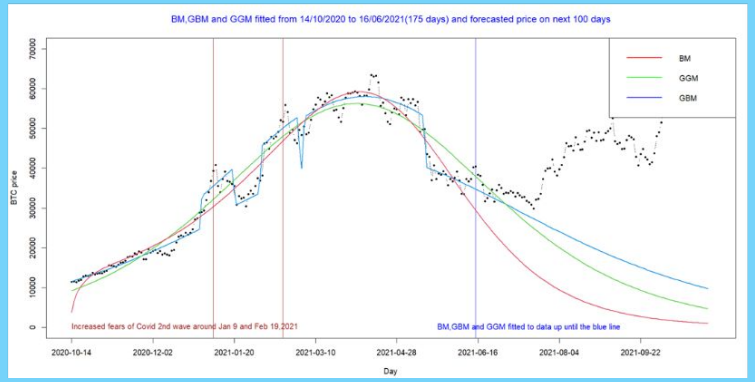
- the price of the next day of BTC is highly correlated on present day
- Every Monday BTC Price is positively correlated with previous Monday
- Every Wednesday BTC Price is negatively correlated with previous Monday
- there is no seasonality in our data.

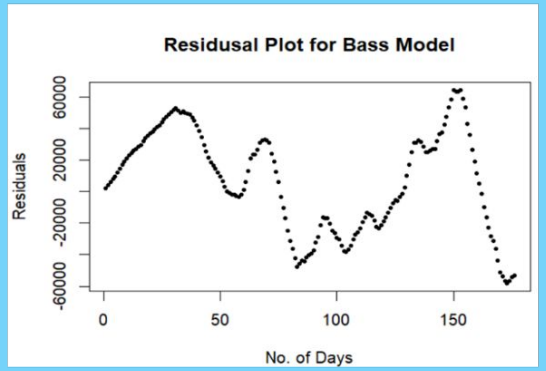
Non - Linear Models for EDA

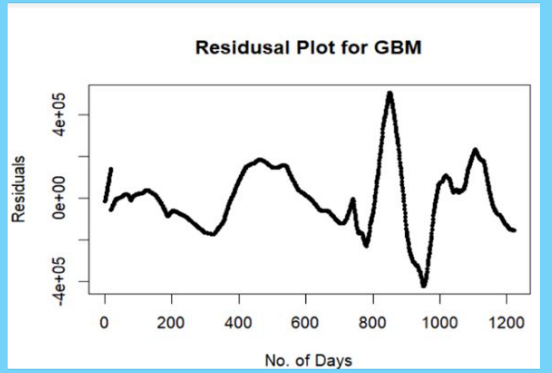


- Generalized Bass Model with 3 rectangular shocks fitted the data better when compared to GGM and BM.
- The presence of small peaks in the data on January 9th and February 19th, 2021 suggests that there were periods of increased concern about a potential second wave of COVID-19.

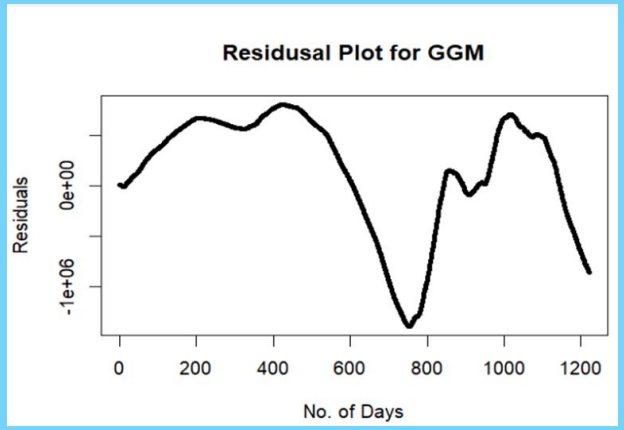
Non - Linear Models for EDA





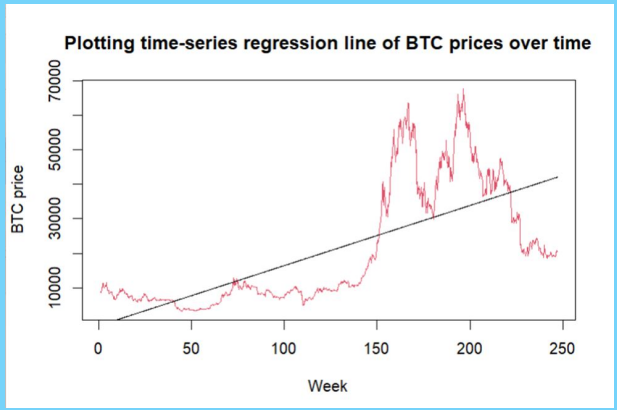


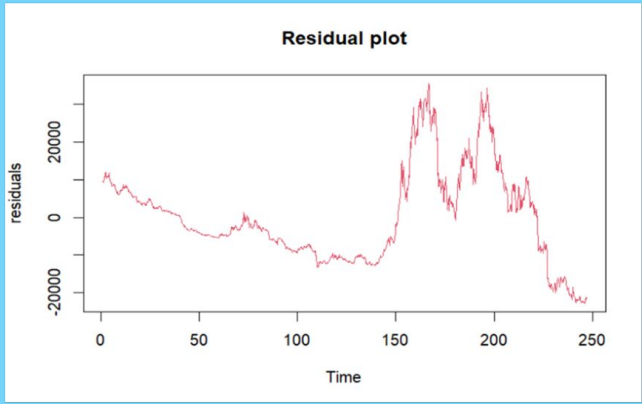
- We can see that GBM model best of 3 models when compared with residual plot could not forecast accurately as they consider market potential to be constant which implies that number of adopters will not change over time which not true in real world scenario.
- Due to this there are biased forecasts as market potential as time change.



TIME SERIES LINEAR MODEL

Fitting time series object with trend, we see that the global trend of Bitcoin prices is positive. However, this model is clearly underfitting, and we can see that there is a pattern in the residuals, which means that there is information that the model failed to capture.






```
Call:
tslm(formula = ts_data ~ trend)

Residuals:
   Min     1Q Median     3Q    Max
-22765  -8701  -2585   7039  35491

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -783.4845   699.9549  -1.119    0.263
trend         34.7776     0.9843  35.334   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12270 on 1229 degrees of freedom
Multiple R-squared:  0.5039,    Adjusted R-squared:  0.5035
F-statistic:  1248 on 1 and 1229 DF,  p-value: < 2.2e-16
```

Value after Durbin-Watson test = 0.009
-> Strong positive correlation
-> R^2 value is 0.5039 indicates that the model explains only about 50% of the variance in the response variable

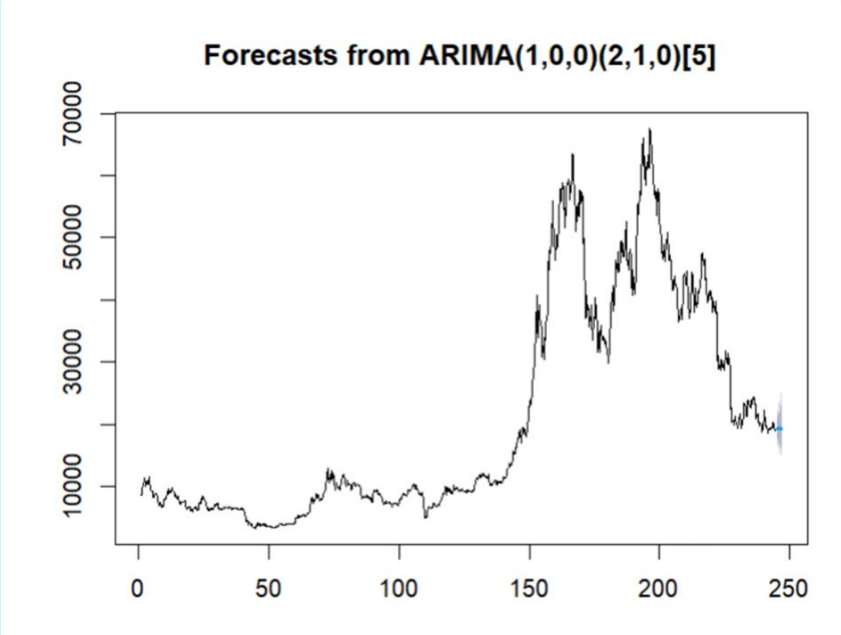
AIC: 26436.69

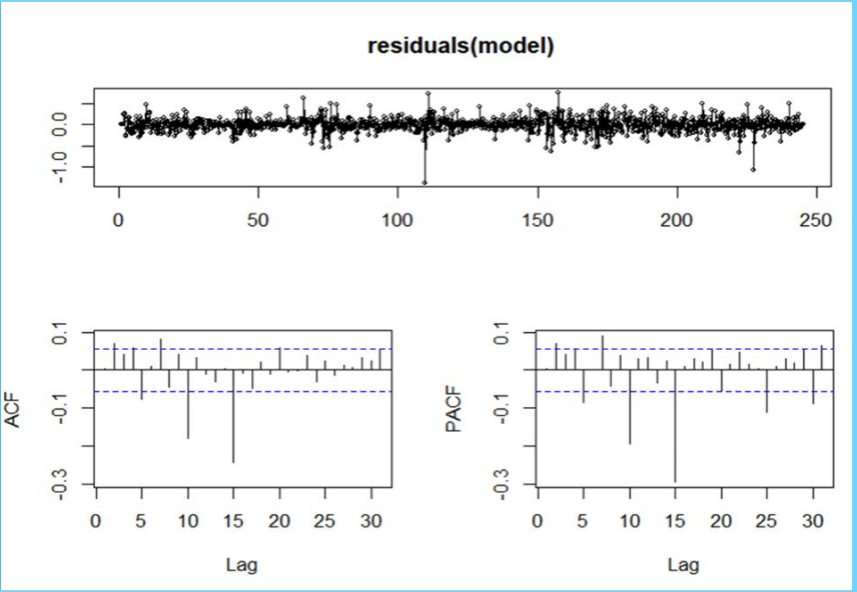
MAPE: 117.6073

RMSE: 23017.09

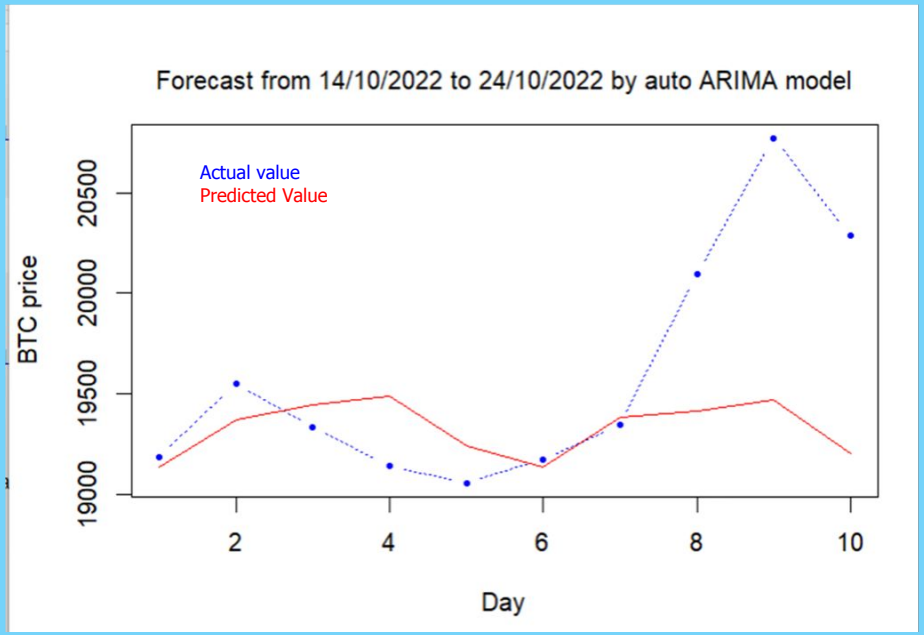
AUTO ARIMA

Auto ARIMA found that the best non seasonal parameters are (1,0,0) and seasonal parameters are (2,1,0).





AUTO ARIMA



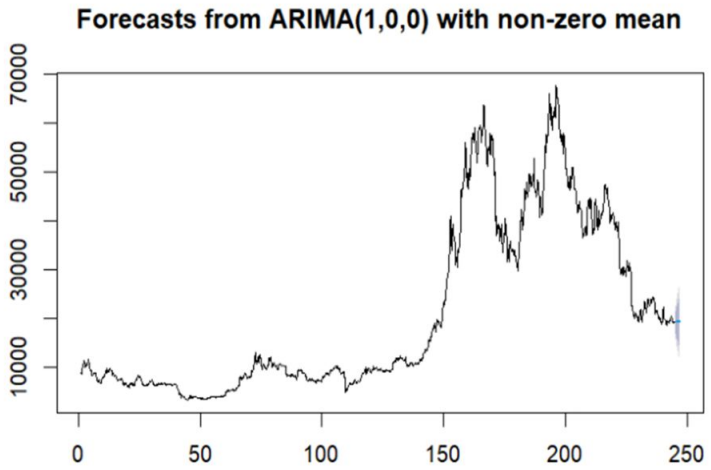

```
Series: ts_data_train
ARIMA(1,0,0)(2,1,0)[5]
Box Cox transformation: lambda= 0.1191897

Coefficients:
         ar1     sar1     sar2
      0.9210  -0.6948  -0.2956
s.e.  0.0115   0.0282   0.0280

sigma^2 = 0.02686:  log likelihood = 473.49
AIC=-938.98   AICc=-938.94   BIC=-918.56

Training set error measures:
                    ME     RMSE      MAE          MPE     MAPE      MASE          ACF1
Training set -4.294164 1381.86 766.9572 -0.08260945 3.563658 0.4796591 0.0003300313
```

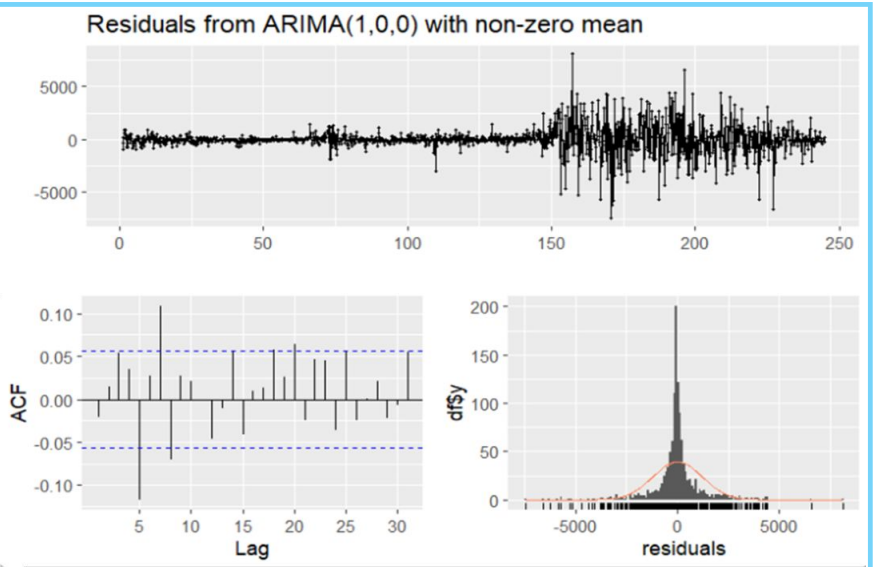
On Test Data:

AIC: -938.9767
MAPE: 2.000154
RMSE: 595.6623

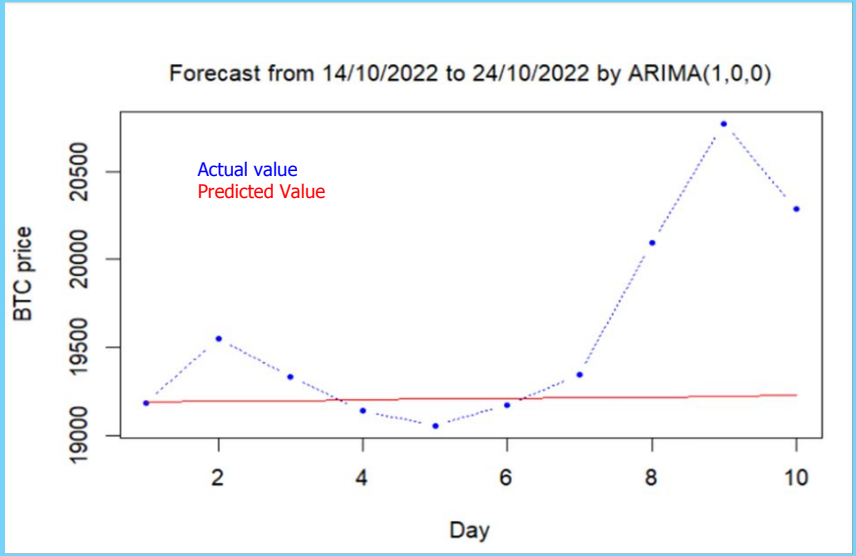
ARIMA(1,0,0)

Based on PACF parameter in EDA this ARIMA model non - seasonal parameters are considered.





ARIMA(1,0,0)




```
Call:
arima(x = ts_data_train, order = c(1, 0, 0))

Coefficients:
         ar1  intercept
      0.9971  20577.857
s.e.  0.0018   9639.878

sigma^2 estimated as 1458905:  log likelihood = -10400.05,  aic = 20806.09

Training set error measures:
                    ME     RMSE      MAE         MPE     MAPE     MASE         ACF1
Training set 8.023423 1207.851 667.1324 -0.3236026 3.075091 1.002656 -0.02125917
```

On Test Data:

AIC: 20806.09
MAPE: 2.160796
RMSE: 669.8516

SUMMARY

Models	RMSE	ADJR^2	MAPE	AIC
Bass Model	20711.11	0.9950559	105.8251	NA
GBM	1101.522	0.9996094	3.886277	NA
GGM	20711.57	0.9897177	105.8275	NA
TSLM with trend	23017.09	0.5175	117.6073	26436.69
ARIMA(1,0,0)	669.8516	NA	2.160796	20806.09
Auto.ARIMA	595.6623	NA	2.000154	-938.9767

DATASET-2

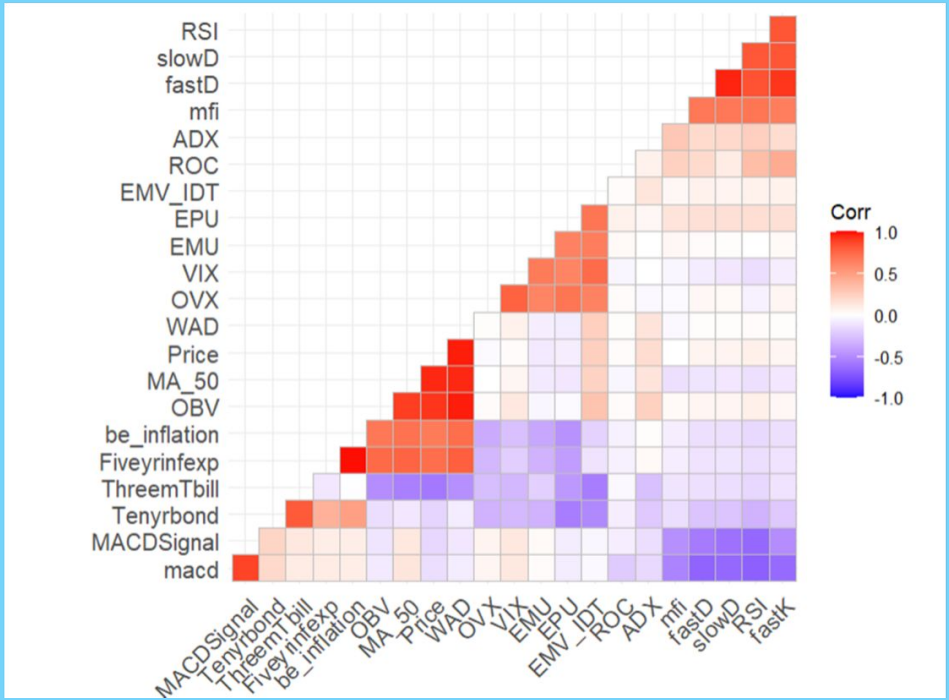
DATASET

A total of 23 features are considered for this Project and some of these features are considered from Yahoo Finance, St. Louis Federal Reserve, USA and Chicago Board Options Exchange, USA

No.	Features from Yahoo Finance	Features deduced from Yahoo Finance Dataset using TTR package	Macro Economic Features from S t. Louis Federal Reserve, USA
1	Date	Relative Strength Index Indicator (RSI)	Economic Policy Uncertainty(EPU)
2	Open	Stochastic Oscillator (stochOSC)	Economic Market Uncertainty (EMU)
3	High	Average Directional Index(ADX)	US Ten-Year Bond (Tenyrbond)
4	Low	Rate Of Change of Price (ROC)	Three-month T-bill (ThreemTbill)
5	Close	On-balance volume (OBV)	Cboe Volatility Index (VIX)
6	Volume	Money Flow Index (MFI)	Cboe Crude Oil ETF Volatility Index (OVX)
7		Williams Accumulation Distribution (WAD)	Equity Market Volatility Infectious Disease Tracker (EMV_IDT)
8		50 day Moving Average(MA_50)	Break-Even Inflation (be_inflation)
9		Moving Average Convergence/Divergence (MACD)	

EXPLORATORY DATA ANALYSIS

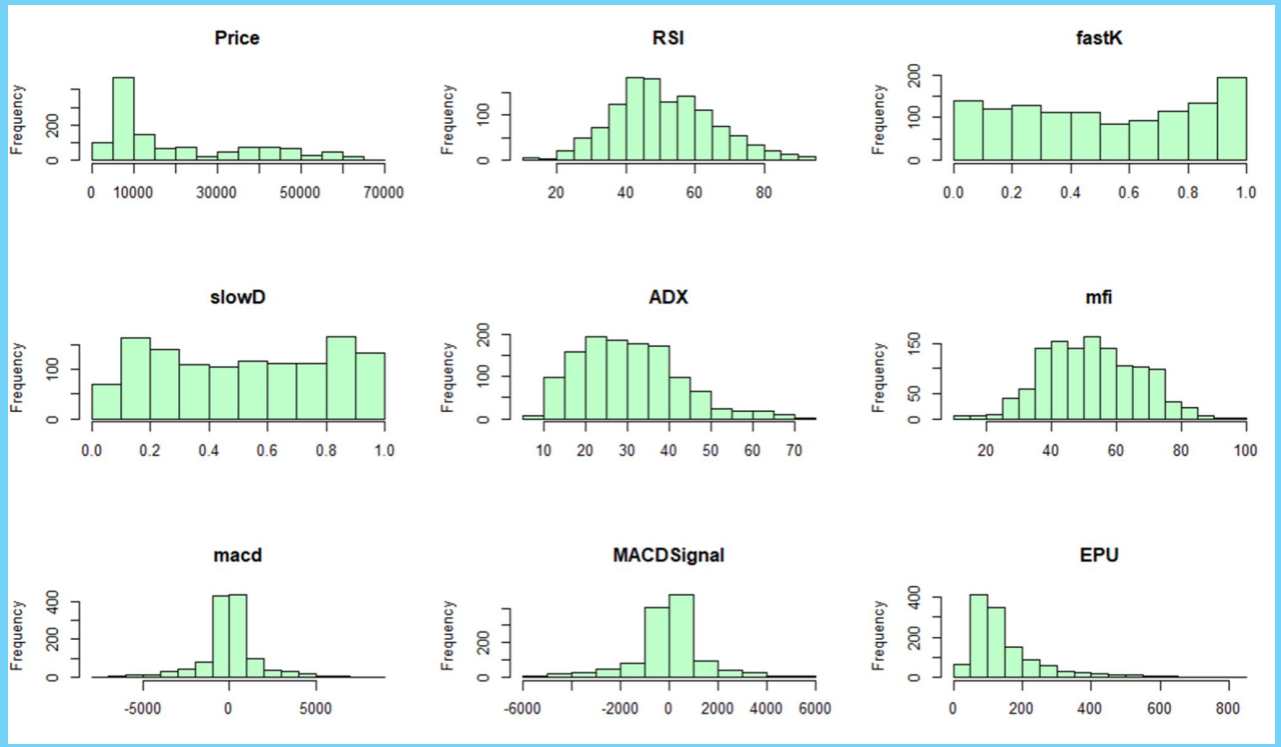
- Before starting feature selection we drop Open, High, Low, Volume columns since they have high correlation with closing price and are not known to us on the day of prediction.
- The correlation plotted below to remove features that are highly correlated to price and have high correlation among themselves.



- OBV, MA_50, WAD and ROC features have high correlation with price so they are removed.
- fastK, fastD and fastD, slowD are corelated among themselves so fastD feature is removed.

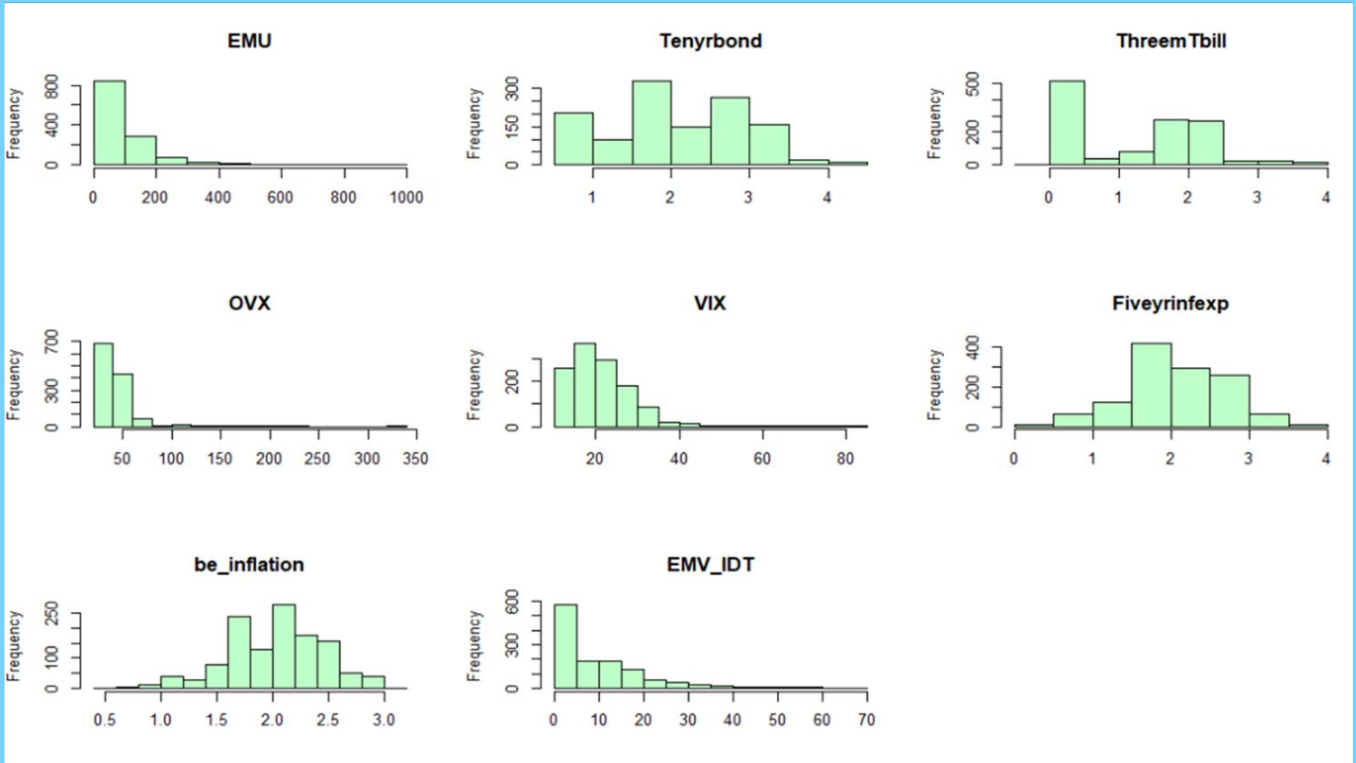
EXPLORATORY DATA ANALYSIS

- The distribution of the 17 remaining features excluding Date are:



EXPLORATORY DATA ANALYSIS

- The distribution of the 17 remaining features excluding Date are:



Pre-processing

- The Dataset is divided into 2 parts :
  - Train (70%)
  - Test (30%)

- When comparing relative importance between all features it is good to keep all the features on the same scale.

- So, min max scaler is fitted on training data and transformed on test data

Feature Selection

- The feature selection is one of the most important parts in the process of building the model.

- It is important as it helps in:
  - Reducing Dimensionality
  - Improving Model Performance
  - Enhancing Interpretability
  - Reducing Overfitting
  - Saving computational time

- Different types of feature selection techniques used are :
  - Generalized Additive Regression
  - Step-wise Regression
  - Gradient boosting Model

Feature Selection using Generalized Additive Regression

```
Family: gaussian
Link function: identity

Formula:
Price ~ s(date_col_train) + s(RSI) + s(fastK) + s(slowD) + s(ADX) +
    s(mfi) + s(macd) + s(MACDSignal) + s(EPU) + s(EMU) + s(Tenyrbond) +
    s(ThreemTbill) + s(OVX) + s(VIX) + s(Fiveyrinfexp) + s(be_inflation) +
    s(EMV_IDT)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.284371   0.001391   204.5   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                    edf Ref.df      F  p-value
s(date_col_train) 8.944  8.997 94.879  < 2e-16 ***
s(RSI)            6.946  8.047 12.522  < 2e-16 ***
s(fastK)          1.000  1.000  6.175 0.013167 *
s(slowD)          1.000  1.000 49.601  < 2e-16 ***
s(ADX)            3.205  4.080 10.979  < 2e-16 ***
s(mfi)            4.760  5.912  6.001 6.16e-06 ***
s(macd)           7.539  8.470  3.666 0.000204 ***
s(MACDSignal)     6.331  7.526 15.907  < 2e-16 ***
s(EPU)            2.908  3.730  2.332 0.052556 .
s(EMU)            2.326  2.984  0.820 0.451818
s(Tenyrbond)      8.549  8.935 53.510  < 2e-16 ***
s(ThreemTbill)    8.772  8.977 29.964  < 2e-16 ***
s(OVX)            2.000  2.536  1.569 0.148820
s(VIX)            7.603  8.506  5.624 4.97e-07 ***
s(Fiveyrinfexp)   8.966  8.997  6.347  < 2e-16 ***
s(be_inflation)   8.980  8.998  5.443 9.78e-07 ***
s(EMV_IDT)        7.023  8.072  2.237 0.022490 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.979   Deviance explained = 98.1%
```

After applying GAR the most significant variables are:

- date_col_train
- RSI
- slowD
- ADX
- mfi
- macd
- MACDSignal
- Tenyrbond
- ThreemTbill
- VIX
- Fiveyrinfexp
- be_inflation

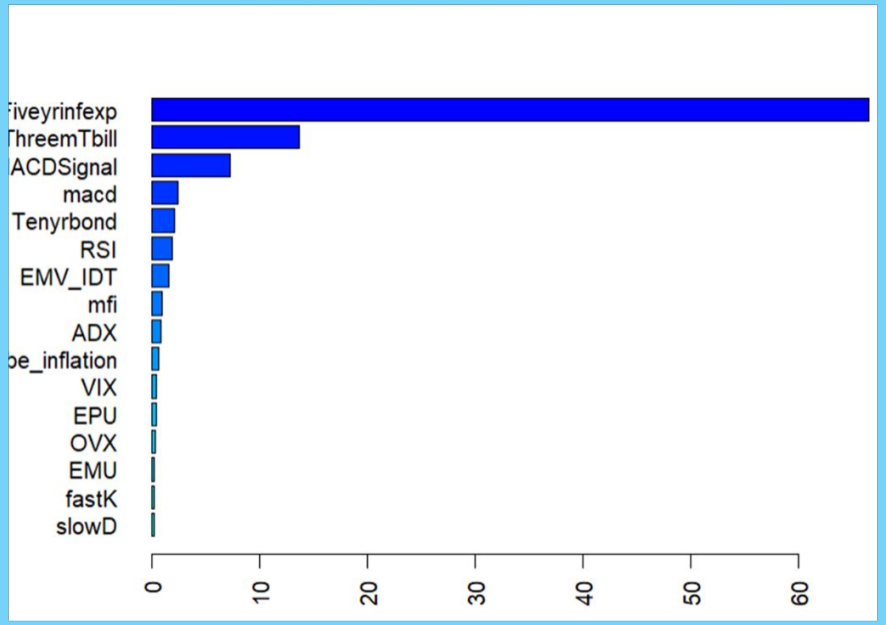
Feature Selection using Stepwise Linear Regression

```
Call:
lm(formula = Price ~ date_col_train + RSI + fastK + slowD + mfi +
    MACDSignal + EPU + EMU + Tenyrbond + ThreemTbill + OVX +
    VIX + be_inflation, data = df_train_scaled)

Residuals:
     Min       1Q   Median       3Q      Max
-0.21200 -0.07500 -0.00751  0.05625  0.47001

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -3.838e+00  2.431e-01 -15.788  < 2e-16 ***
date_col_train  2.374e-04  1.458e-05  16.282  < 2e-16 ***
RSI             1.124e-01  4.655e-02   2.415  0.01595 *
fastK           4.404e-02  2.380e-02   1.851  0.06455 .
slowD          -1.001e-01  2.481e-02  -4.034 5.97e-05 ***
mfi            -1.281e-01  3.002e-02  -4.267 2.21e-05 ***
MACDSignal     -5.183e-01  3.825e-02 -13.550  < 2e-16 ***
EPU            -3.490e-01  4.333e-02  -8.054 2.70e-15 ***
EMU             1.406e-01  4.732e-02   2.972  0.00305 **
Tenyrbond       4.049e-01  7.123e-02   5.685 1.80e-08 ***
ThreemTbill    -8.495e-01  5.209e-02 -16.307  < 2e-16 ***
OVX             3.202e-01  7.429e-02   4.310 1.82e-05 ***
VIX            -4.215e-01  5.886e-02  -7.161 1.74e-12 ***
be_inflation    3.625e-01  6.964e-02   5.206 2.42e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1034 on 848 degrees of freedom
Multiple R-squared:  0.8682,    Adjusted R-squared:  0.8662
F-statistic: 429.9 on 13 and 848 DF,  p-value: < 2.2e-16
```

After applying Stepwise Regression the most significant variables are:

- date_col_train
- slowD
- mfi
- MACDSignal
- EPU
- EMU
- Tenyrbond
- ThreemTbill
- OVX
- VIX
- be_inflation

Feature Selection using Gradient Boosting Models




```
> summary(boost.model, las=2, cBar=18)
                           var    rel.inf
Fiveyrinfexp     Fiveyrinfexp 66.5950213
ThreemTbill       ThreemTbill 13.7139535
MACDSignal         MACDSignal  7.2427827
macd                     macd  2.4461586
Tenyrbond           Tenyrbond  2.1604869
RSI                       RSI  1.8636159
EMV_IDT               EMV_IDT  1.5550848
mfi                       mfi  0.9946526
ADX                       ADX  0.8302276
be_inflation be_inflation      0.6969751
VIX                       VIX  0.4699025
EPU                       EPU  0.4166098
OVX                       OVX  0.2993686
EMU                       EMU  0.2528828
fastK                   fastK  0.2465567
slowD                   slowD  0.2157205
```

After getting feature importance from GBM we consider features with importance above 0.5% for building the model.

Model Building

Based on 3 feature selection techniques 4 models are built and are compared among each other with metrics like RMSE,MAPE,AIC.

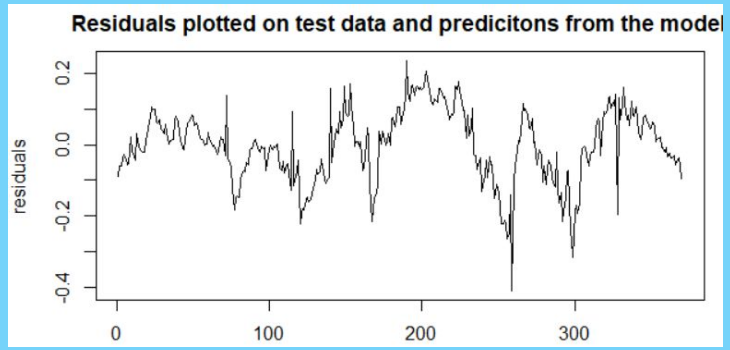
The 4 models are:

- Linear Model
- GAM
- CART
- Gradient Boosting

1. Model Building using features extracted from Generalized Additive Regression

1) Stepwise-Linear Model

```
Call:
lm(formula = Price ~ date_col_train + RSI + slowD + mfi + MACDSignal +
    Tenyrbond + ThreemTbill + VIX + Fiveyrinfexp, data = train_df1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.22993 -0.07282 -0.01064  0.06530  0.50252

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -3.651e+00  2.655e-01 -13.755  < 2e-16 ***
date_col_train  2.240e-04  1.574e-05  14.229  < 2e-16 ***
RSI             1.154e-01  4.281e-02   2.695  0.00718 **
slowD          -7.004e-02  2.241e-02  -3.125  0.00184 **
mfi            -1.338e-01  3.075e-02  -4.350 1.52e-05 ***
MACDSignal     -5.121e-01  3.883e-02 -13.187  < 2e-16 ***
Tenyrbond       4.231e-01  6.746e-02   6.271 5.70e-10 ***
ThreemTbill    -8.002e-01  5.133e-02 -15.590  < 2e-16 ***
VIX            -3.662e-01  4.858e-02  -7.537 1.23e-13 ***
Fiveyrinfexp    4.173e-01  6.675e-02   6.252 6.42e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1076 on 852 degrees of freedom
Multiple R-squared:  0.8564,    Adjusted R-squared:  0.8549
F-statistic: 564.6 on 9 and 852 DF,  p-value: < 2.2e-16
```

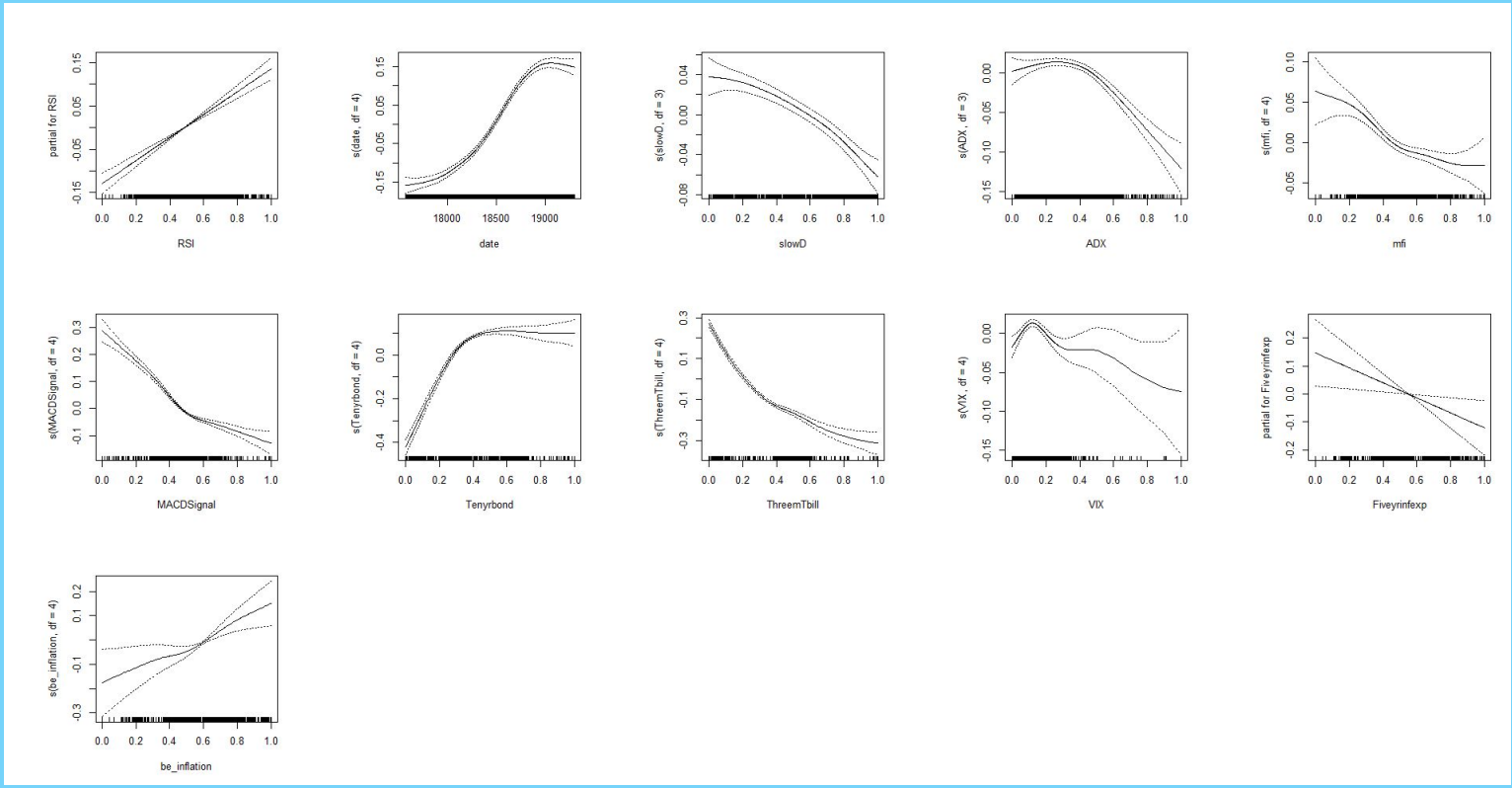


Results on test data:

- MAPE –> 115.0491
- RMSE –> 0.1005336
- AIC –> -1384.442
- Adj R^2 –> 0.8549

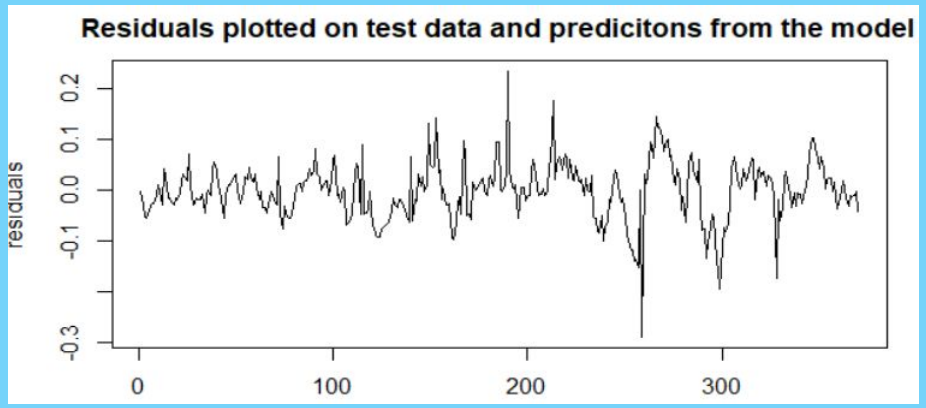
1. Model Building using features extracted from Generalized Additive Regression process

2) Stepwise Generalized Additive Model



1. Model Building using features extracted from Generalized Additive Regression process

2) Stepwise Generalized Additive Model



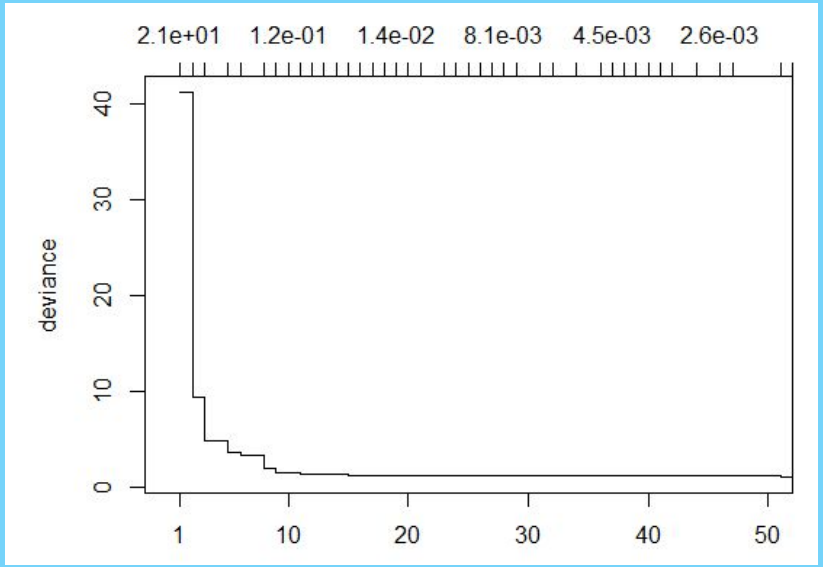
Results on test data:

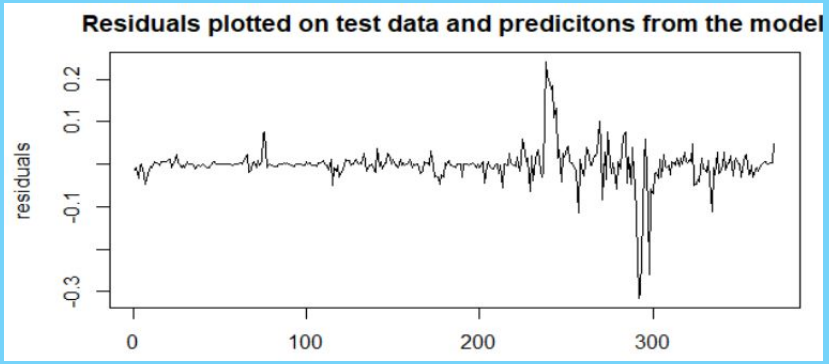
- MAPE –> 77.41214
- RMSE –> 0.055782

```
Call: gam(formula = train_df1$Price ~ RSI + s(date, df = 4) + s(slowD,
    df = 3) + s(ADX, df = 3) + s(mfi, df = 4) + s(MACDSignal,
    df = 4) + s(Tenyrbond, df = 4) + s(ThreemTbill, df = 4) +
    s(VIX, df = 4) + Fiveyrinfexp + s(be_inflation, df = 4),
    data = train_df1, trace = FALSE)
Deviance Residuals:
      Min        1Q    Median        3Q       Max
-0.185756 -0.033863 -0.002405  0.025085  0.412478

(Dispersion Parameter for gaussian family taken to be 0.0036)

    Null Deviance: 68.7558 on 861 degrees of freedom
Residual Deviance: 3.0064 on 825.0001 degrees of freedom
AIC: -2355.379

Number of Local Scoring Iterations: NA

Anova for Parametric Effects
                          Df  Sum Sq Mean Sq   F value    Pr(>F)
RSI                        1  0.2099  0.2099   57.5960 8.699e-14 ***
s(date, df = 4)            1 24.5343 24.5343 6732.5019 < 2.2e-16 ***
s(slowD, df = 3)           1  0.0578  0.0578   15.8579 7.430e-05 ***
s(ADX, df = 3)             1  0.0365  0.0365   10.0213  0.001604 **
s(mfi, df = 4)             1  0.0196  0.0196    5.3811  0.020599 *
s(MACDSignal, df = 4)      1  1.9894  1.9894  545.9198 < 2.2e-16 ***
s(Tenyrbond, df = 4)       1  0.0052  0.0052    1.4355  0.231209
s(ThreemTbill, df = 4)     1  9.3047  9.3047 2553.3320 < 2.2e-16 ***
s(VIX, df = 4)             1  0.1119  0.1119   30.7119 4.028e-08 ***
Fiveyrinfexp               1  0.0079  0.0079    2.1779  0.140390
s(be_inflation, df = 4)    1  0.0372  0.0372   10.1968  0.001460 **
Residuals                825  3.0064  0.0036
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Anova for Nonparametric Effects
                          Npar Df Npar F     Pr(F)
(Intercept)
RSI
s(date, df = 4)                3  41.78 < 2.2e-16 ***
s(slowD, df = 3)               2   5.44  0.004488 **
s(ADX, df = 3)                 2  23.48 1.206e-10 ***
s(mfi, df = 4)                 3   4.98  0.001983 **
s(MACDSignal, df = 4)          3  26.10 4.441e-16 ***
s(Tenyrbond, df = 4)           3 353.24 < 2.2e-16 ***
s(ThreemTbill, df = 4)         3 115.74 < 2.2e-16 ***
s(VIX, df = 4)                 3  14.57 2.960e-09 ***
Fiveyrinfexp
s(be_inflation, df = 4)        3   8.45 1.551e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1. Model Building using features extracted from Generalized Additive Regression process

3) CART (Classification and Regression Trees)





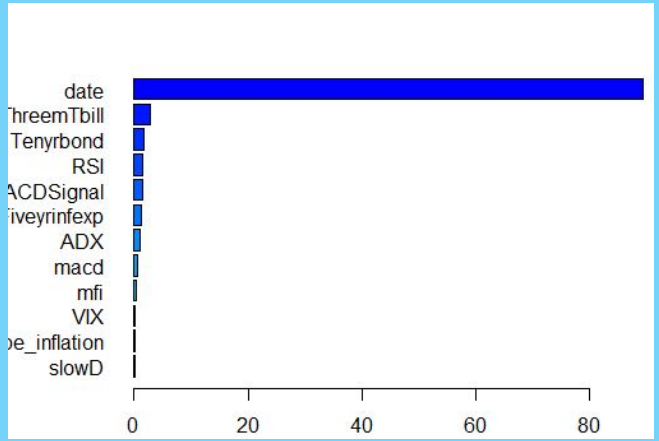
Results on test data:

- MAPE –> 36.68
- RMSE –> 0.0552

Model Building using features extracted from Generalized Additive Regression process

4) Gradient Boosting Model

```
> summary(boost.model, las=1, cBar=20)
                               var     rel.inf
date                          date 89.34372298
ThreemTbill        ThreemTbill  2.72583657
Tenyrbond            Tenyrbond  1.58374397
RSI                          RSI  1.51110523
MACDSignal          MACDSignal  1.45073632
Fiveyrinfexp      Fiveyrinfexp  1.13125089
ADX                          ADX  1.03588192
macd                        macd  0.50666699
mfi                          mfi  0.31064601
VIX                          VIX  0.19208014
be_inflation    be_inflation  0.14687829
slowD                      slowD  0.06145071
```



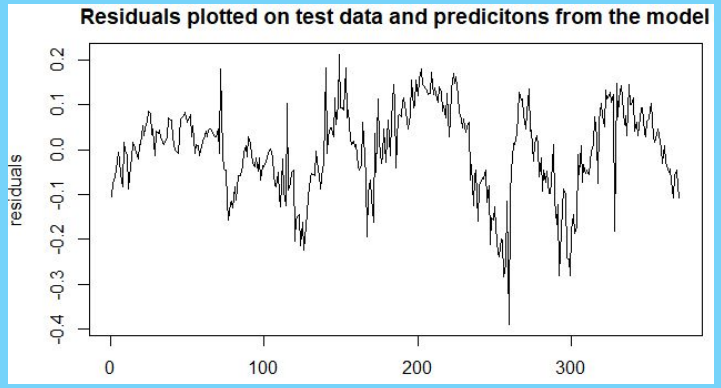
Results on test data with 5000 trees, depth = 4 and shrinkage=0.01

- MAPE min. –> 8.581
- RMSE min. –> 0.016

2. Model Building using features extracted from Stepwise Linear Regression

1) Stepwise-Linear Model

```
Call:
lm(formula = Price ~ date + slowD + mfi + MACDSignal + EPU +
    EMU + Tenyrbond + ThreemTbill + OVX + VIX + be_inflation,
    data = train_df2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.22310 -0.07402 -0.00907  0.05672  0.48621

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.842e+00  2.437e-01 -15.763  < 2e-16 ***
date         2.395e-04  1.468e-05  16.317  < 2e-16 ***
slowD       -3.853e-02  1.915e-02  -2.012  0.04451 *
mfi         -8.127e-02  2.787e-02  -2.916  0.00363 **
MACDSignal  -5.413e-01  3.615e-02 -14.973  < 2e-16 ***
EPU         -3.355e-01  4.336e-02  -7.736 2.90e-14 ***
EMU          1.384e-01  4.770e-02   2.901  0.00381 **
Tenyrbond    3.543e-01  6.945e-02   5.102 4.16e-07 ***
ThreemTbill -8.148e-01  5.106e-02 -15.956  < 2e-16 ***
OVX          3.173e-01  7.436e-02   4.267 2.20e-05 ***
VIX         -4.434e-01  5.911e-02  -7.502 1.59e-13 ***
be_inflation 3.797e-01  6.955e-02   5.459 6.29e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1043 on 850 degrees of freedom
Multiple R-squared:  0.8656,    Adjusted R-squared:  0.8639
F-statistic: 497.8 on 11 and 850 DF,  p-value: < 2.2e-16
```

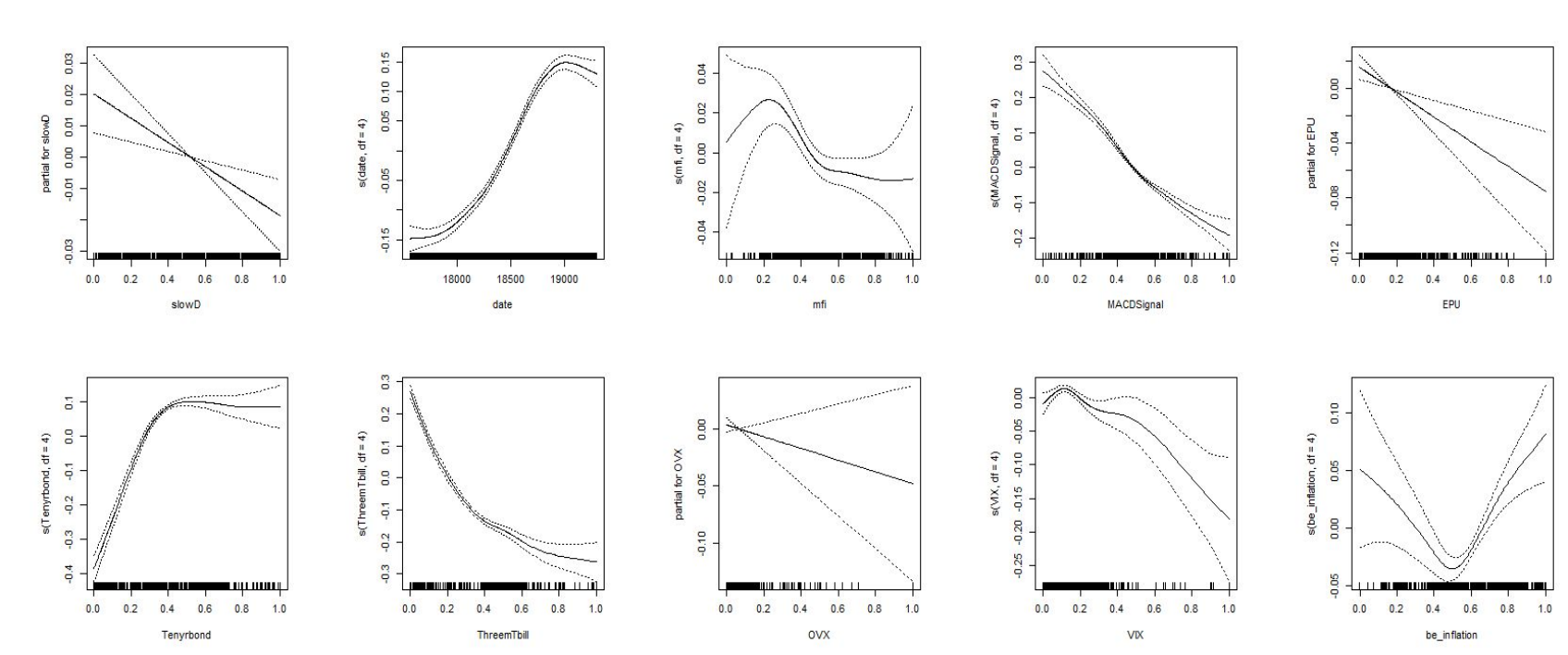


Results on test data:

- MAPE –> 109.3437
- RMSE –> 0.09725803
- AIC –> -1437.706
- Adj R^2 –> 0.8639

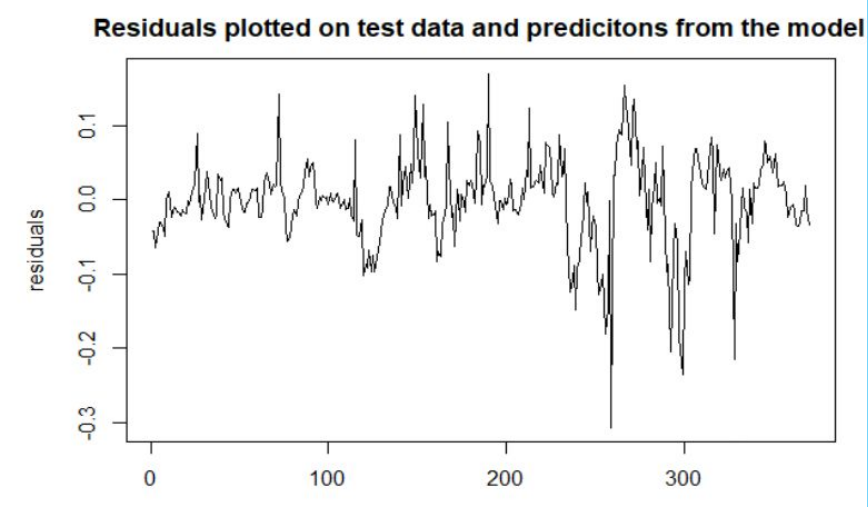
2. Model Building using features extracted from Stepwise Linear Regression

2) Stepwise Generalized Additive Model



2. Model Building using features extracted from Step-wise Linear Regression

2) Stepwise Generalized Additive Model



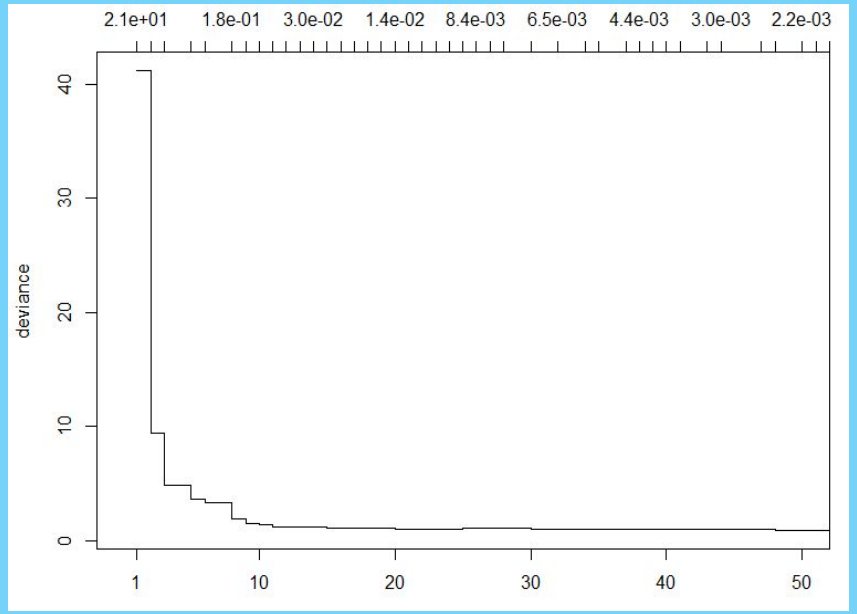
Results on test data:

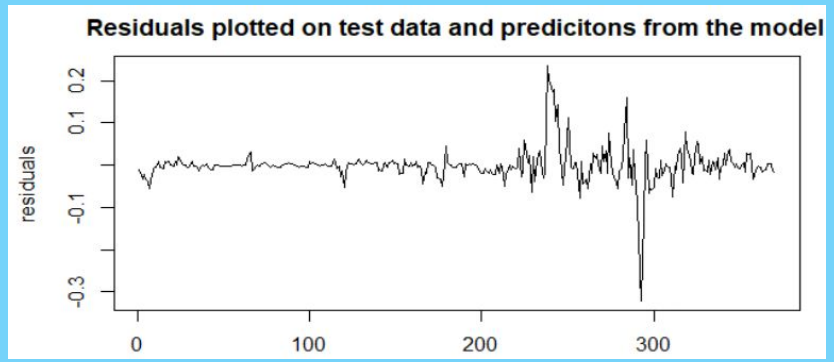
- MAPE –> 56.9786
- RMSE –> 0.058855

```
Call: gam(formula = train_df2$Price ~ slowD + s(date, df = 4) + s(mfi,
    df = 4) + s(MACDSignal, df = 4) + EPU + s(Tenyrbond, df = 4) +
    s(ThreemTbill, df = 4) + OVX + s(VIX, df = 4) + s(be_inflation,
    df = 4), data = train_df2, trace = FALSE)
Deviance Residuals:
      Min        1Q    Median        3Q       Max
-0.158745 -0.034944 -0.001262  0.026383  0.473853

(Dispersion Parameter for gaussian family taken to be 0.0042)

    Null Deviance: 68.7558 on 861 degrees of freedom
Residual Deviance: 3.5003 on 830.0001 degrees of freedom
AIC: -2234.279

Number of Local Scoring Iterations: NA

Anova for Parametric Effects
                      Df  Sum Sq Mean Sq   F value    Pr(>F)
slowD                  1  0.0475  0.0475   11.2685 0.0008242 ***
s(date, df = 4)        1 23.1289 23.1289 5484.4144 < 2.2e-16 ***
s(mfi, df = 4)         1  0.0926  0.0926   21.9494 3.270e-06 ***
s(MACDSignal, df = 4)  1  2.3573  2.3573  558.9684 < 2.2e-16 ***
EPU                    1  0.7812  0.7812  185.2347 < 2.2e-16 ***
s(Tenyrbond, df = 4)   1  0.1337  0.1337   31.7138 2.445e-08 ***
s(ThreemTbill, df = 4) 1  9.3306  9.3306 2212.5103 < 2.2e-16 ***
OVX                    1  0.1210  0.1210   28.6927 1.099e-07 ***
s(VIX, df = 4)         1  0.1021  0.1021   24.2209 1.036e-06 ***
s(be_inflation, df = 4) 1  0.0253  0.0253    6.0107 0.0144246 *
Residuals            830  3.5003  0.0042
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Anova for Nonparametric Effects
                      Npar Df  Npar F     Pr(F)
(Intercept)
slowD
s(date, df = 4)             3  38.438 < 2.2e-16 ***
s(mfi, df = 4)              3   5.004  0.001916 **
s(MACDSignal, df = 4)       3  10.704 6.596e-07 ***
EPU
s(Tenyrbond, df = 4)        3 315.237 < 2.2e-16 ***
s(ThreemTbill, df = 4)      3 148.999 < 2.2e-16 ***
OVX
s(VIX, df = 4)              3  11.633 1.793e-07 ***
s(be_inflation, df = 4)     3  20.386 9.419e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2. Model Building using features extracted from Stepwise Linear Regression

3) CART (Classification and Regression Trees)





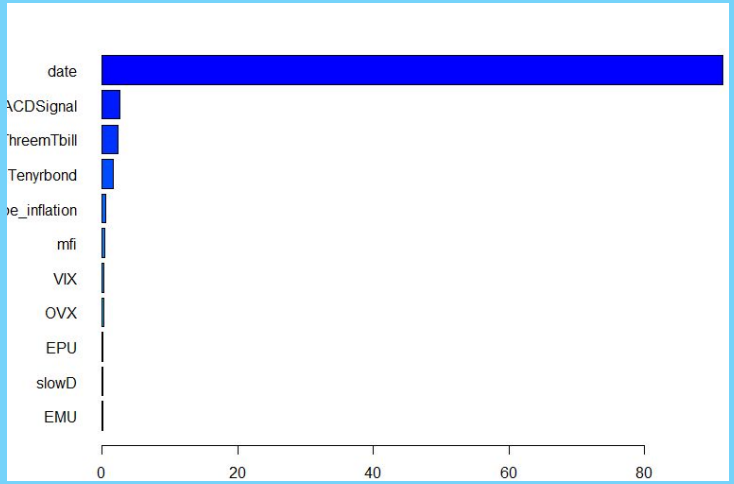
Results on test data:

- MAPE –> 13.59625
- RMSE –> 0.0419

2. Model Building using features extracted from Generalized Additive Regression process

4) Gradient Boosting Model

```
> summary(boost.model, las=1, cBar=20)
                             var     rel.inf
date                        date 91.48776442
MACDSignal           MACDSignal  2.65915420
ThreemTbill         ThreemTbill  2.39001818
Tenyrbond             Tenyrbond  1.70914519
be_inflation       be_inflation  0.56601427
mfi                         mfi  0.35020330
VIX                         VIX  0.31966422
OVX                         OVX  0.22282062
EPU                         EPU  0.10069874
slowD                     slowD  0.09762656
EMU                         EMU  0.09689028
```



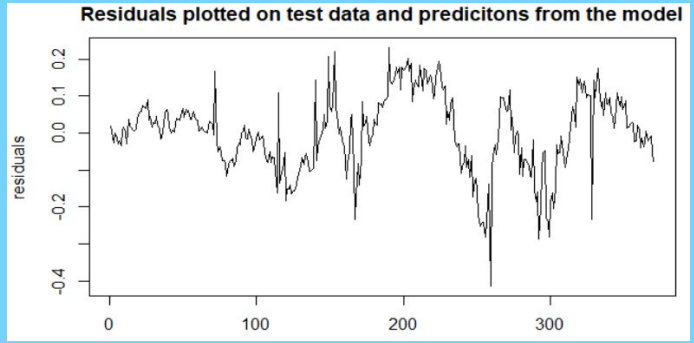
Results on test data with 5000 trees, depth = 4 and shrinkage=0.01

- MAPE min. –> 10.41613
- RMSE min. –> 0.0203

3. Model Building using features extracted from Gradient Boosting Models

1) Stepwise-Linear Model

```
Call:
lm(formula = Price ~ date + RSI + mfi + MACDSignal + Tenyrbond +
    ThreemTbill + be_inflation + EMV_IDT, data = train_df3)

Residuals:
     Min       1Q   Median       3Q      Max
-0.25254 -0.06334 -0.00991  0.06468  0.52734

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.190e+00  2.354e-01 -13.549  < 2e-16 ***
date         1.919e-04  1.386e-05  13.845  < 2e-16 ***
RSI          9.032e-02  3.735e-02   2.418 0.015796 *
mfi         -1.887e-01  3.007e-02  -6.274 5.59e-10 ***
MACDSignal  -4.886e-01  3.908e-02 -12.501  < 2e-16 ***
Tenyrbond    2.317e-01  6.882e-02   3.367 0.000795 ***
ThreemTbill -7.070e-01  5.268e-02 -13.422  < 2e-16 ***
be_inflation 6.208e-01  6.163e-02  10.073  < 2e-16 ***
EMV_IDT     -1.564e-01  3.754e-02  -4.165 3.42e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.11 on 853 degrees of freedom
Multiple R-squared:   0.85,     Adjusted R-squared:  0.8486
F-statistic: 604.2 on 8 and 853 DF,  p-value: < 2.2e-16
```

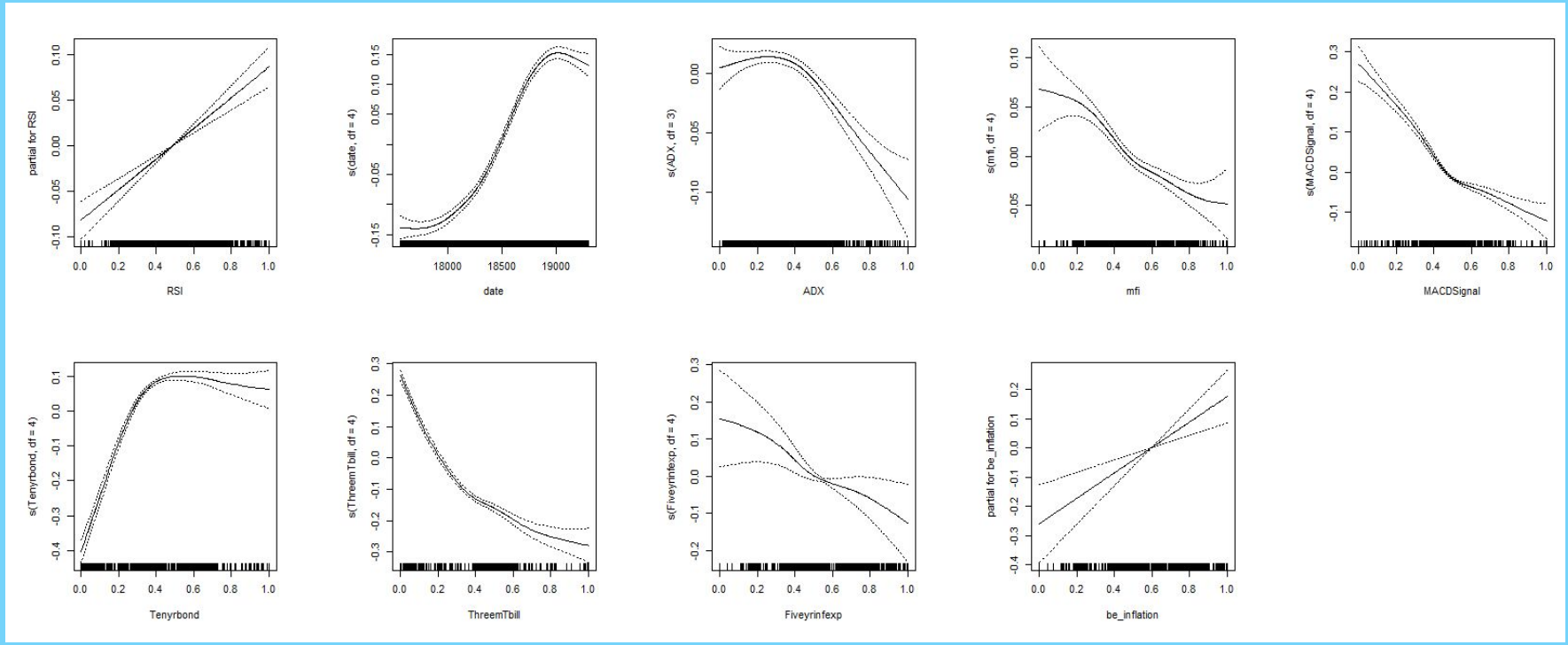


Results on test data:

- MAPE –> 108.4665
- RMSE –> 0.1023907
- AIC –> -1348.3
- Adj R^2 –> 0.8486

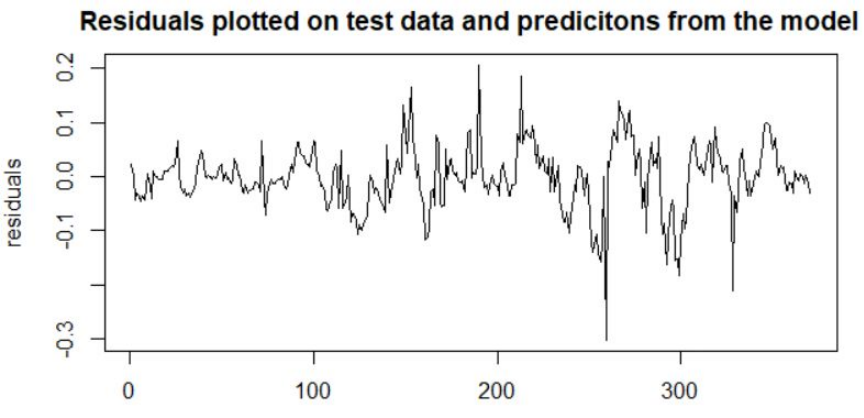
3. Model Building using features extracted from Gradient Boosting Models

2) Stepwise Generalized Additive Model



3. Model Building using features extracted from Gradient Boosting Models

2) Stepwise Generalized Additive Model



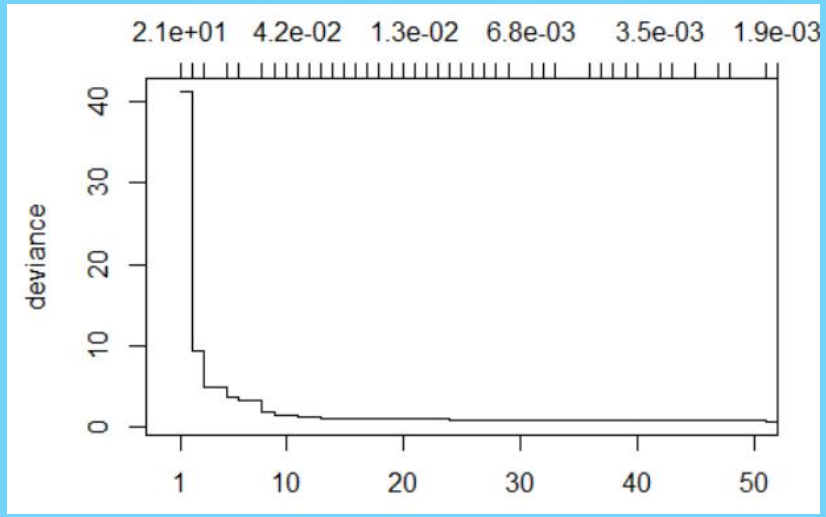
Results on test data:

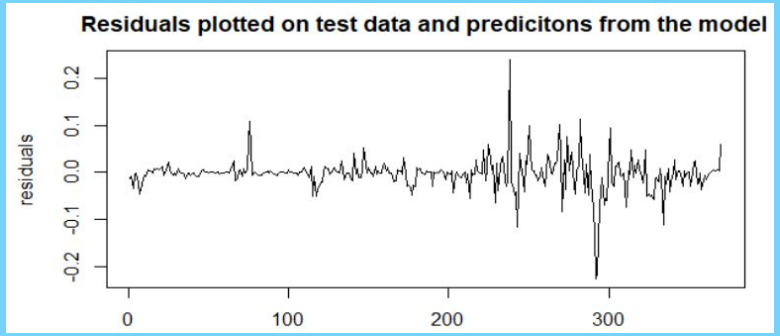
- MAPE –> 69.37234
- RMSE –> 0.058453

```
Call: gam(formula = train_df3$Price ~ RSI + s(date, df = 4) + s(ADX,
    df = 3) + s(mfi, df = 4) + s(MACDSignal, df = 4) + s(Tenyrbond,
    df = 4) + s(ThreemTbill, df = 4) + s(Fiveyrinfexp, df = 4) +
    be_inflation, data = train_df3, trace = FALSE)
Deviance Residuals:
      Min        1Q    Median        3Q       Max
-0.172433 -0.036861 -0.002651  0.027874  0.442551

(Dispersion Parameter for gaussian family taken to be 0.004)

    Null Deviance: 68.7558 on 861 degrees of freedom
Residual Deviance: 3.2941 on 831.9999 degrees of freedom
AIC: -2290.6

Number of Local Scoring Iterations: NA

Anova for Parametric Effects
                          Df  Sum Sq Mean Sq   F value    Pr(>F)
RSI                        1  0.1945  0.1945   49.1272 4.960e-12 ***
s(date, df = 4)            1 25.1988 25.1988 6364.4685 < 2.2e-16 ***
s(ADX, df = 3)             1  0.0414  0.0414   10.4560  0.001271 **
s(mfi, df = 4)             1  0.0063  0.0063    1.5973  0.206634
s(MACDSignal, df = 4)      1  1.7744  1.7744  448.1540 < 2.2e-16 ***
s(Tenyrbond, df = 4)       1  0.0024  0.0024    0.6049  0.436943
s(ThreemTbill, df = 4)     1  9.5007  9.5007 2399.5836 < 2.2e-16 ***
s(Fiveyrinfexp, df = 4)    1  0.1014  0.1014   25.5993 5.173e-07 ***
be_inflation               1  0.0599  0.0599   15.1188  0.000109 ***
Residuals                832  3.2941  0.0040
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Anova for Nonparametric Effects
                          Npar Df Npar F     Pr(F)
(Intercept)
RSI
s(date, df = 4)                3  48.06 < 2.2e-16 ***
s(ADX, df = 3)                 2  17.48 3.666e-08 ***
s(mfi, df = 4)                 3   4.18   0.00598 **
s(MACDSignal, df = 4)          3  23.92 7.438e-15 ***
s(Tenyrbond, df = 4)           3 380.79 < 2.2e-16 ***
s(ThreemTbill, df = 4)         3 119.36 < 2.2e-16 ***
s(Fiveyrinfexp, df = 4)        3   9.83 2.230e-06 ***
be_inflation
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

3. Model Building using features extracted from Gradient Boosting Models

3) CART (Classification and Regression Trees)





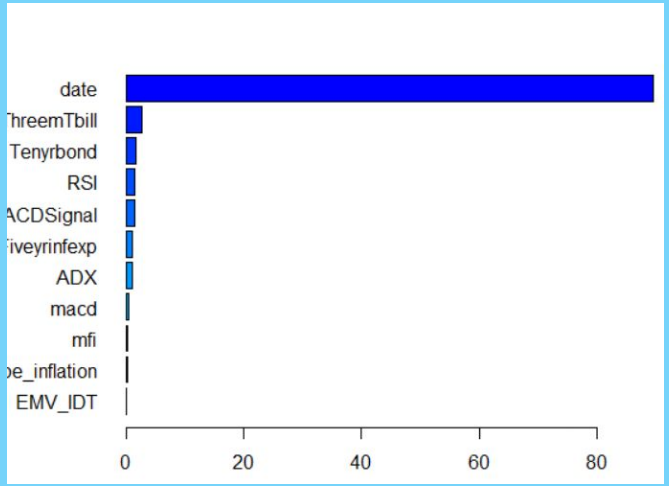
Results on test data:

- MAPE –> 17.352
- RMSE –> 0.034

3. Model Building using features extracted from Gradient Boosting Models

4) Gradient Boosting Model

```
> summary(boost.model, las=1, cBar=20)
                               var     rel.inf
date                          date 89.52983287
ThreemTbill    ThreemTbill 2.69160335
Tenyrbond        Tenyrbond 1.58488335
RSI                          RSI 1.51623977
MACDSignal      MACDSignal 1.46723575
Fiveyrinfexp Fiveyrinfexp 1.11586554
ADX                          ADX 1.02847148
macd                        macd 0.50455509
mfi                          mfi 0.31805362
be_inflation be_inflation 0.14741773
EMV_IDT                  EMV_IDT 0.09584144
```



Results on test data with 5000 trees, depth = 4 and shrinkage=0.01

- MAPE min. –> 8.028302
- RMSE min. –> 0.01716198

Final Comparison

Feature Selection Technique	Model	MAPE	RMSE	AIC	Adj. R^2
Generalized Additive Regression	Stepwise-Linear Model	115.0491	0.1005336	-1384.442	0.8549
	Step-wise Generalized Additive Model	77.41214	0.055782	NA	NA
	Classification and Regression Trees	36.68	0.0552	NA	NA
	Gradient Boosting Model	8.581	0.016	NA	NA
Step-wise Linear Regression	Stepwise-Linear Model	109.3437	0.09725803	-1437.706	0.8639
	Step-wise Generalized Additive Model	56.9786	0.058855	NA	NA
	Classification and Regression Trees	13.59625	0.0419	NA	NA
	Gradient Boosting Model	10.41613	0.0203	NA	NA
Gradient Boosting Models	Stepwise-Linear Model	108.4665	0.1023907	-1348.3	0.8446
	Step-wise Generalized Additive Model	69.37234	0.058453	NA	NA
	Classification and Regression Trees	17.352	0.034	NA	NA
	Gradient Boosting Model	8.028302	0.01716198	NA	NA

CONCLUSION

- A comprehensive analysis was conducted to forecast the outcome variable using various advanced modeling techniques such as BM, GBM, GGM, Time Series Linear Model with trend and ARIMA.
- The models were evaluated and interpreted using diagnostic techniques such as residual plots, Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) correlograms.
- Different types of Feature Selection methods were also applied, and it was found that GBM feature selection performed the best among all other methods.
- As a result, the GBM model with feature selection was determined to be the most suitable model for forecasting the outcome variable with the lowest Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE).

Thank you!

Any questions?

GLOSSARY

- **Relative Strength Indicator (RSI):** is a momentum oscillator that measures the speed and change of price movements.
- **Stochastic Oscillator (stochOSC):** it measures the momentum (rate of acceleration) in price movement.
- **Average Directional Index (ADX):** is a technical analysis indicator used by some traders to determine the strength of a trend.
- **Rate Of Change of Price (ROC):** is an unbounded momentum indicator used in technical analysis set against a zero-level midpoint.
- **On-Balance Volume (OBV):** is a technical analysis indicator intended to relate price and volume in the stock market.
- **Money Flow Index (MFI):** is a momentum indicator that measures the flow of money into and out of a security over a specified period of time.

GLOSSARY

- **Williams Accumulation Distribution (WAD):** is an indicator used in technical analysis to gauge bullish and bearish price pressure by comparing positive (accumulation) and negative (distribution) price movements.
- **Moving Average Convergence/Divergence:** is a trend-following momentum indicator that shows the relationship between two exponential moving averages (EMAs) of a security's price.
- **Economic Policy Uncertainty (EPU):** is an index developed to measure economic policy uncertainty and is calculated by calculating the relative frequency of each country's newspaper articles, including terms' economy', 'policy', and 'uncertainty'.
- **Economic Market Uncertainty (EMU):** is a measure of the level of uncertainty in financial markets.

GLOSSARY

- **Treasury Bill (T-Bill):** is a short-term debt security issued by the U.S. Department of the Treasury, backed by the full faith and credit of the U.S. government.
- **CBOE Volatility Index (VIX):** is a measure of the expected volatility of the U.S. stock market over the next 30 days
- **CBOE Crude Oil ETF Volatility Index (OVX):** is a measure of the expected volatility of crude oil prices over the next 30 days.
- **Equity Market Volatility Infectious Disease Tracker (EMV_IDT):** is included to account for the impact of the COVID-19 pandemic.
- **Break-Even Inflation (be_inflation):** is a measure of the expected rate of inflation over the life of a bond or other fixed-income security.

REFERENCES

- **Forecasting Bitcoin price direction with random forests: How important are interest rates, inflation, and market volatility? Syed Abul Basher a , Perry Sadorsky**