Capstone HarvardX - Project: Algorithms Comparison for Financial Shares Forecasting

Papacosmas Nicholas

May 2019

Contents

01.Abstract

The financial markets contain a plethora of statistical patterns. The behavior of those patterns is similar with the behavior of the natural phenomena patterns. That means that both are affected by unknown and unstable variables. Which leads to high unpredictability and volatility. That makes almost impossible to forecast future behavior.

As Burton Malkiel, who argues in his 1973 book, "A Random Walk Down Wall Street,"

Nevertheles, one forecasting methodology is: To use the past performance of markets as a predictor for the future. That can be achieved by observing the changes of small seasonal intervals, when the time series is stationary.

02.Introduction

The purpose of this project is to analyze 3 different algorithms for the financial forecasting of Daimler share. Disclaimer I am working on Advanced Analytics of Daimler AG. This analysis is for educational purposes and not for financial advising.

03.Methodology

We will use Daimler historical share market datasets (from 2010). For forecasting future values. On those nature of forecasting we assume that some patterns of our sets have carriage on future short linear interims. The same approach is applied on the weather forecasting.

We will apply mathematical technical indicators in our datasets on the below domains:
-**Support & resistance**
-**Trend**
-**Momentum**
-**Volume**
-**Volatility**

Some of them are the:
-**Moving average convergence/divergence**
-**Relative strength index**
-**Stochastic oscillator**
-**Ease of movement**
-**Larry Williams oscilator. Etc.**

The 3 algorithms that will we compare to predict the Daimler financial share behavior are:

-(LASSO) Least Absolute Shrinkage and Selection Operator. This method is based on a linear regression model is proposed as a novel method to predict financial market behavior

-Deep Learning (Long Short Term Memory Neural Network of linear
stack densely connected layers.

-And eXtreme Gradient Boosting

Apreciation to the colleagues from:
Cornell University (arXiv:1512.04916v3) [q-fin.CP] (Ruoxua, 2016)

Cornell University Social and Information Networks (cs.SI); Computational Finance (q-fin.CP) (Jichang Zhao,2019)

04.Disclaimer

This article is intended for academic and educational purposes and is not an investment recommendation. The information that we provide or should not be a substitute for advice from an investment professional. The models discussed in this paper do not reflect the investment performance. A decision to invest in any product or strategy should not be based on the information or conclusions contained herein. This is neither an offer to sell / buy nor a solicitation for an offer to buy interests in securities.

05.Data Observation

We will use as data sets the Daimler AG Symbol (DDAIF)

Display of the 6 first entries of our dataset

```
          DDAIF.Open               DDAIF.High               DDAIF.Low
               FALSE                    FALSE                   FALSE
         DDAIF.Close             DDAIF.Volume          DDAIF.Adjusted
               FALSE                    FALSE                   FALSE
       Avg_volume_10            Avg_volume_20       Volume_perc_avg_60
                TRUE                     TRUE                    TRUE
               Range      perc_change_closing        change_from_yest
               FALSE                     TRUE                    TRUE
       moving_avg_10            moving_avg_20            moving_avg_60
                TRUE                     TRUE                    TRUE
  perc_moving_avg_10       perc_moving_avg_20       perc_moving_avg_60
                TRUE                     TRUE                    TRUE
         cash_tradet      avg_cash_trated_10      avg_cash_trated_20
               FALSE                     TRUE                    TRUE
  avg_cash_trated_60 Avg_Dollar_volume_pct_10 Avg_Dollar_volume_pct_20
                TRUE                     TRUE                    TRUE
Avg_Dollar_volume_pct_60                 nightgap          night_gap_perc
                TRUE                     TRUE                    TRUE
  perc_range_previous          perc_range_atpr     perc_range_williams
               FALSE                    FALSE                   FALSE
 one_month_range_perc                    EMA10                   EMA20
```

```
                 TRUE                     TRUE                     TRUE
                EMA60                    WMA10                   EVWMA10
                 TRUE                     TRUE                     TRUE
              ZLEMA10                   VWAP10                     HMA10
                 TRUE                     TRUE                     TRUE
               ALMA10
                 TRUE

           DDAIF.Open DDAIF.High DDAIF.Low DDAIF.Close DDAIF.Volume
2010-05-03      50.90      51.62     50.68       51.26       753100
2010-05-04      48.92      49.31     48.30       48.86      1917000
2010-05-05      47.27      48.27     46.97       47.33      1866200
2010-05-06      47.13      47.82     42.22       46.13      2363700
2010-05-07      46.22      46.82     44.61       45.55      1697900
2010-05-10      48.44      48.99     48.00       48.39      1220800
           DDAIF.Adjusted Avg_volume_10 Avg_volume_20 Volume_perc_avg_60
2010-05-03       37.62413            NA            NA                 NA
2010-05-04       35.86256            NA            NA                 NA
2010-05-05       34.73956            NA            NA                 NA
2010-05-06       33.85878            NA            NA                 NA
2010-05-07       33.43306            NA            NA                 NA
2010-05-10       35.51759            NA            NA                 NA
              Range perc_change_closing change_from_yest moving_avg_10
2010-05-03 0.939999                  NA               NA            NA
2010-05-04 1.010002           -4.682008        -2.399997            NA
2010-05-05 1.299999           -3.131394        -1.529999            NA
2010-05-06 5.599999           -2.535392        -1.200001            NA
2010-05-07 2.209999           -1.257321        -0.580002            NA
2010-05-10 0.990002            6.234907         2.840000            NA
           moving_avg_20 moving_avg_60 perc_moving_avg_10
2010-05-03            NA            NA                 NA
2010-05-04            NA            NA                 NA
2010-05-05            NA            NA                 NA
2010-05-06            NA            NA                 NA
2010-05-07            NA            NA                 NA
2010-05-10            NA            NA                 NA
           perc_moving_avg_20 perc_moving_avg_60 cash_tradet
2010-05-03                 NA                 NA    38603904
2010-05-04                 NA                 NA    93664622
2010-05-05                 NA                 NA    88327250
2010-05-06                 NA                 NA   109037483
2010-05-07                 NA                 NA    77339343
2010-05-10                 NA                 NA    59074511
           avg_cash_trated_10 avg_cash_trated_20 avg_cash_trated_60
2010-05-03                 NA                 NA                 NA
```

```
2010-05-04                 NA                 NA                 NA
2010-05-05                 NA                 NA                 NA
2010-05-06                 NA                 NA                 NA
2010-05-07                 NA                 NA                 NA
2010-05-10                 NA                 NA                 NA
           Avg_Dollar_volume_pct_10 Avg_Dollar_volume_pct_20
2010-05-03                       NA                       NA
2010-05-04                       NA                       NA
2010-05-05                       NA                       NA
2010-05-06                       NA                       NA
2010-05-07                       NA                       NA
2010-05-10                       NA                       NA
           Avg_Dollar_volume_pct_60  nightgap night_gap_perc
2010-05-03                       NA        NA             NA
2010-05-04                       NA -2.340000     -4.5649631
2010-05-05                       NA -1.590001     -3.2541976
2010-05-06                       NA -0.200001     -0.4225671
2010-05-07                       NA  0.090000      0.1951008
2010-05-10                       NA  2.890000      6.3446763
           perc_range_previous perc_range_atpr perc_range_williams
2010-05-03           38.297488        1.833787            38.29802
2010-05-04            5.940285        2.067135            44.55437
2010-05-05            4.615542        2.746670            72.30759
2010-05-06           17.857146       12.139603            30.17856
2010-05-07           30.316846        4.851809            57.46613
2010-05-10            5.050495        2.045881            60.60624
           one_month_range_perc EMA10 EMA20 EMA60 WMA10 EVWMA10 ZLEMA10
2010-05-03                   NA    NA    NA    NA    NA      NA      NA
2010-05-04                   NA    NA    NA    NA    NA      NA      NA
2010-05-05                   NA    NA    NA    NA    NA      NA      NA
2010-05-06                   NA    NA    NA    NA    NA      NA      NA
2010-05-07                   NA    NA    NA    NA    NA      NA      NA
2010-05-10                   NA    NA    NA    NA    NA      NA      NA
           VWAP10 HMA10 ALMA10
2010-05-03     NA    NA     NA
2010-05-04     NA    NA     NA
2010-05-05     NA    NA     NA
2010-05-06     NA    NA     NA
2010-05-07     NA    NA     NA
2010-05-10     NA    NA     NA

An 'xts' object on 2010-05-03/2019-04-30 containing:
  Data: num [1:2264, 1:40] 50.9 48.9 47.3 47.1 46.2 ...
 - attr(*, "dimnames")=List of 2
  ..$ : NULL
```

```
  ..$ : chr [1:40] "DDAIF.Open" "DDAIF.High" "DDAIF.Low" "DDAIF.Close" ...
  Indexed by objects of class: [Date] TZ: UTC
  xts Attributes:
List of 2
 $ src    : chr "yahoo"
 $ updated: POSIXct[1:1], format: "2019-05-15 18:38:05"
```

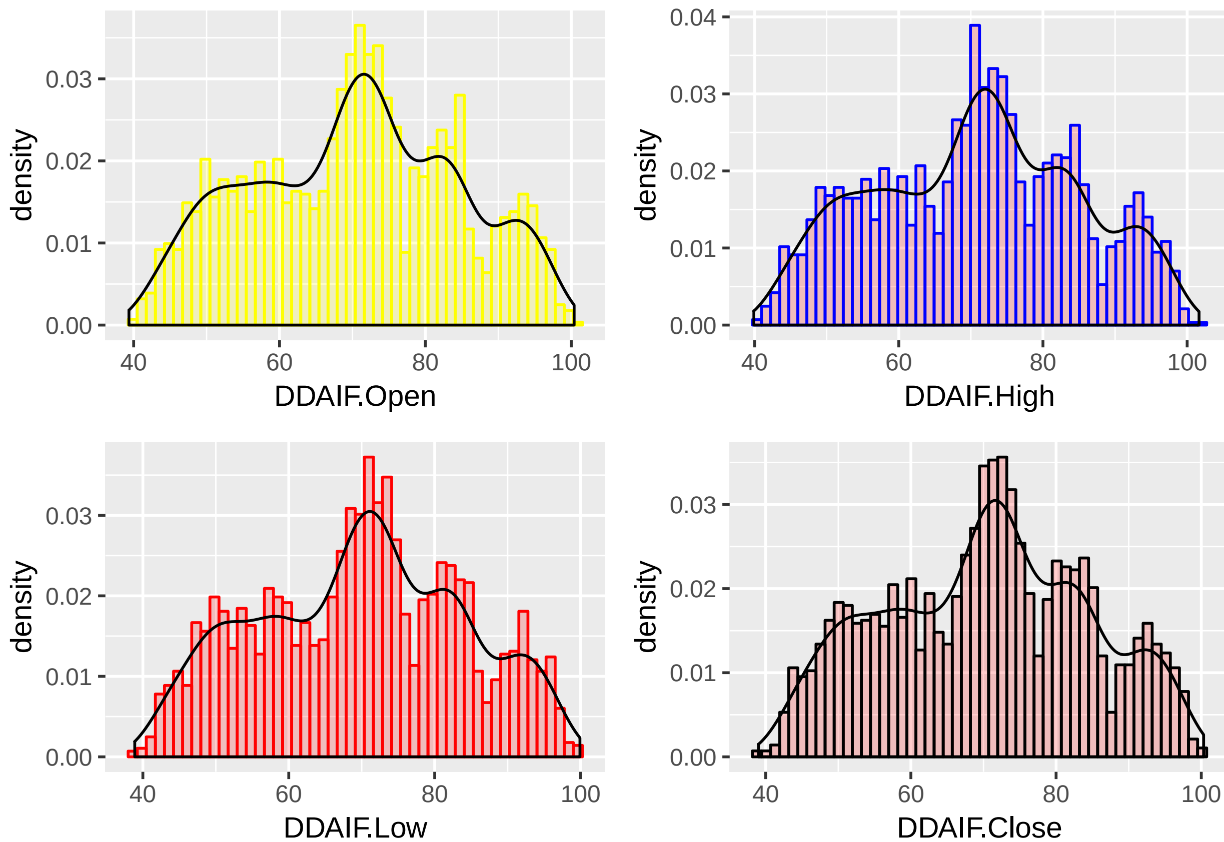
Chart series technical analysis graph

Bollinger Bands, and Moving Average Convergence Divergence

Graph with opening Prices since 2010 pro semester

Graph with the Adjusted Closing Prices

Plot of an An open-high-low-close chart



06.Keras with LSTM

We will apply deep learning networks of linear stack densely connected layers

```
keras_model <- keras_model_sequential()

keras_model %>%
  #We ddd a densely-connected NN layer to an output
  #ReLU (Rectified Linear Unit) Activation Function
  layer_dense(units = 60, activation = 'relu', input_shape = ker) %>%
  layer_dropout(rate = 0.2) %>% #We apply dropout  to prevent overfitting
  layer_dense(units = 50, activation = 'relu') %>%
  layer_dropout(rate = 0.2) %>%
  layer_dense(units = 1, activation = 'linear')
```

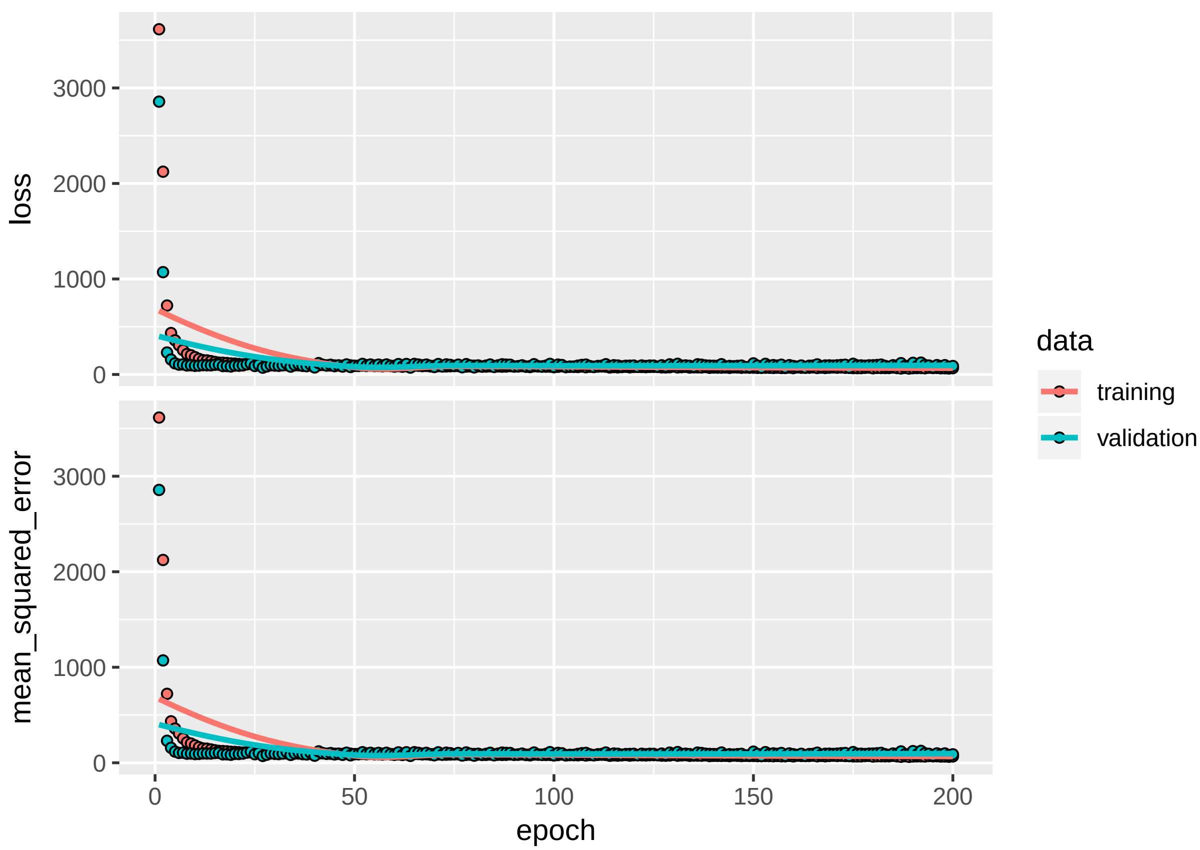
```r
keras_model %>% compile(
  optimizer = 'rmsprop',
  loss = 'mse',
  metrics = 'mse')
```

We train the NN model

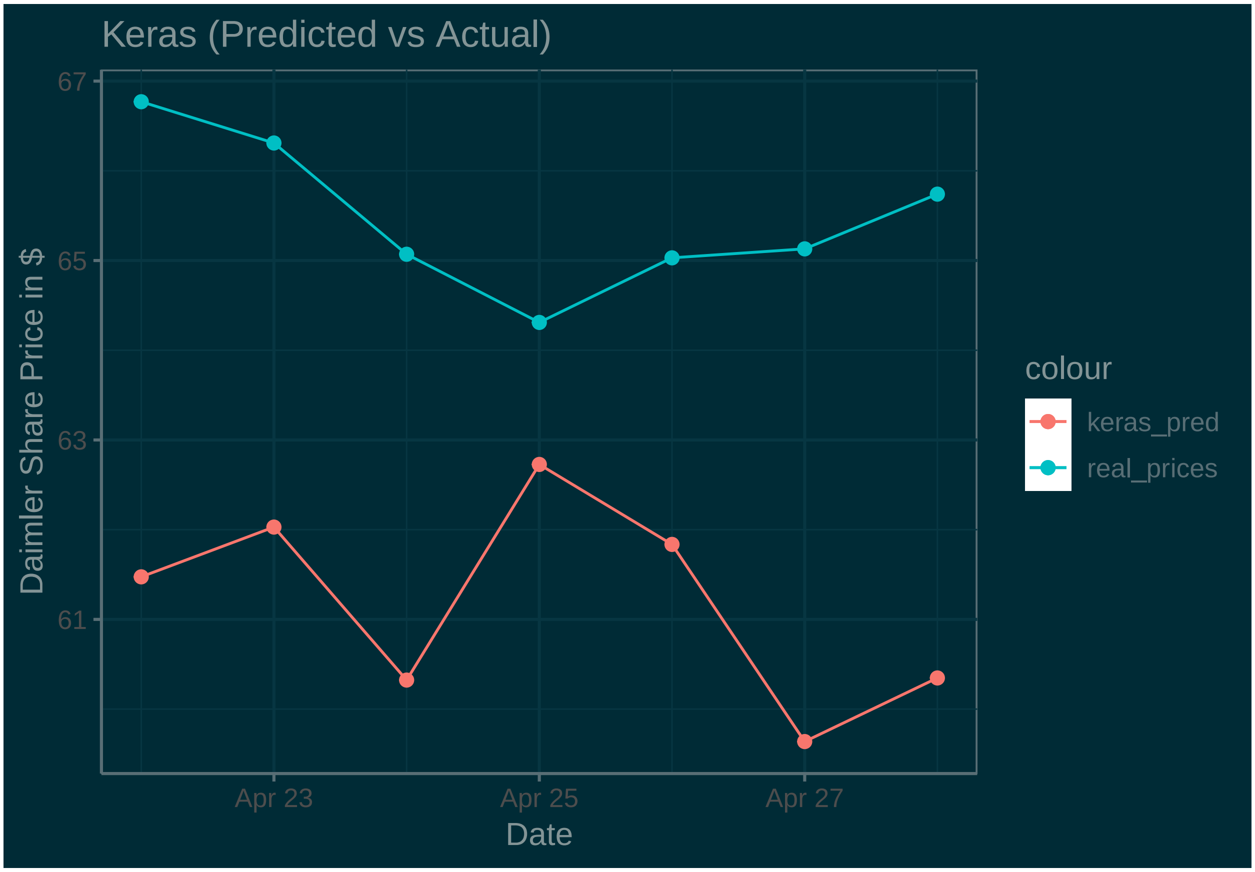
```r
keras_history <- keras_model %>% fit(X_train, y_train, epochs=200,
                                    batch_size=28, validation_split = 0.1,
                                    callbacks = callback_tensorboard("logs/run_a"))
```

```r
keras_pred <- keras_model %>% predict(X_test, batch_size = 28)
```

Plot of Keras Model History



Plot of Keras Model Predictions vs Actuals



KERAS PRED	REAL PRICES
61.47425	66.77
62.02896	66.31
60.32361	65.07
62.72711	64.31
61.83586	65.03
59.63830	65.13
60.34718	65.74

07.Lasso regression model

With caret package we will apply cross validation in order to find the optimal hyperparameters

```
require(caret)
```

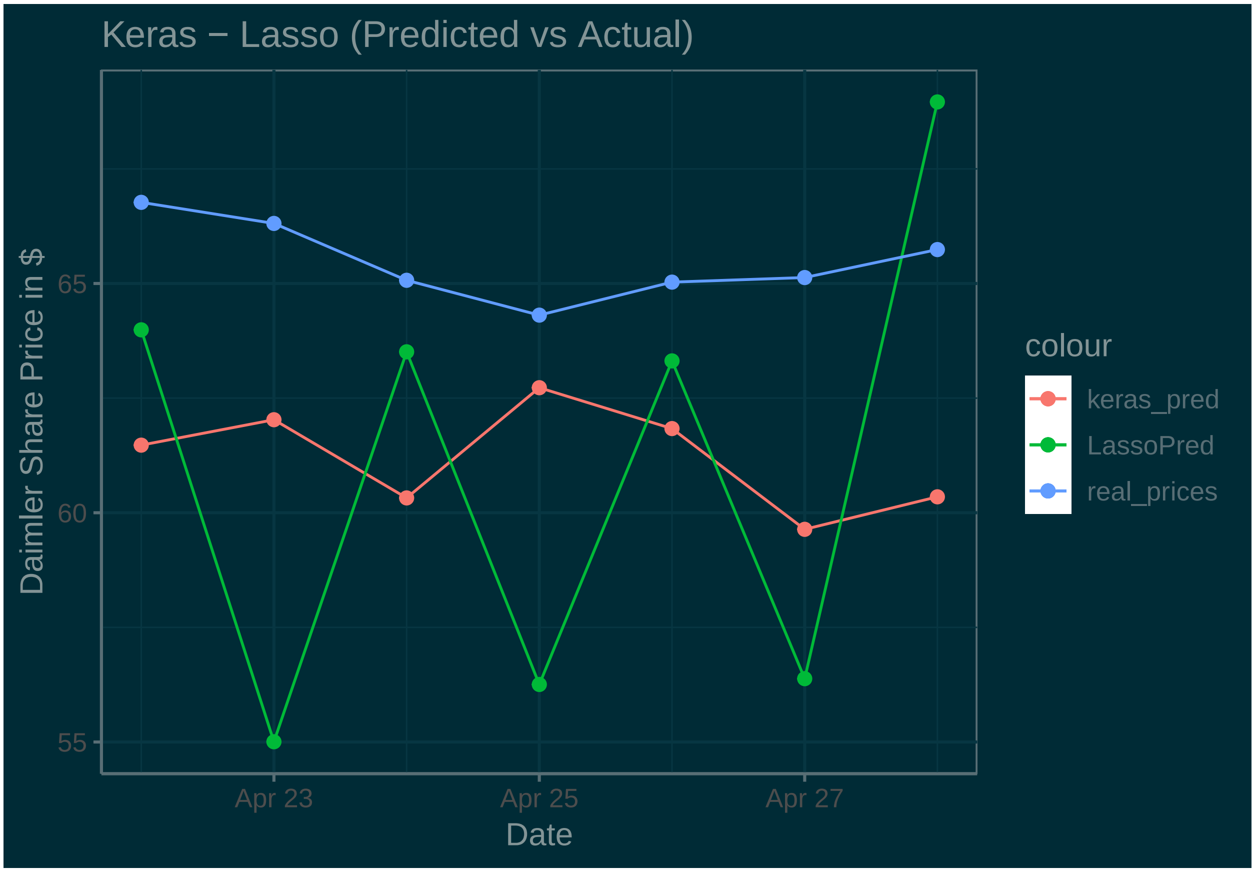
```r
train$DDAIF.Adjusted <- as.numeric(train$DDAIF.Adjusted)
set.seed(123)#algorithm for reproducability
trainControl <-trainControl(method="cv", number=5)
lassoGrid <- expand.grid(alpha = 1, lambda = seq(0.001,0.1,by = 0.0005))

lassomod <- train(DDAIF.Adjusted ~., data = na.omit(train), method='glmnet', trControl=
                  tuneGrid=lassoGrid)
#we display the optimal alpha and lambda penalties
lassomod$bestTune

#We display the root mean squared error
min(lassomod$results$RMSE)

#From caret package we use () varImp. Is a generic method for calculating
#variable importance for objects produced by train and method specific methods
lasso_VarImp <- varImp(lassomod,scale=F)
lasso_Importance <- lasso_VarImp$importance
vars_Selected <- length(which(lasso_Importance$Overall!=0))
vars_NotSelected <- length(which(lasso_Importance$Overall==0))
```

Display of the Laso Regression vars penalty

```
The Lasso regression used 11 variables, and did not used 5 variables.
```

```r
#Run the prediction for the next 7 days
LassoPred <- predict(lassomod, X_test)
```

Plot of Keras and Lasso regression:Predicted vs Actual Prices



	KERAS PRED	LASSO PRED	REAL PRICES
2019-04-22	61.47425	**63.98916**	66.77
2019-04-23	62.02896	**55.00517**	66.31
2019-04-24	60.32361	**63.50995**	65.07
2019-04-25	62.72711	**56.25471**	64.31
2019-04-26	61.83586	**63.31208**	65.03
2019-04-29	59.63830	**56.38149**	65.13
2019-04-30	60.34718	**68.96010**	65.74

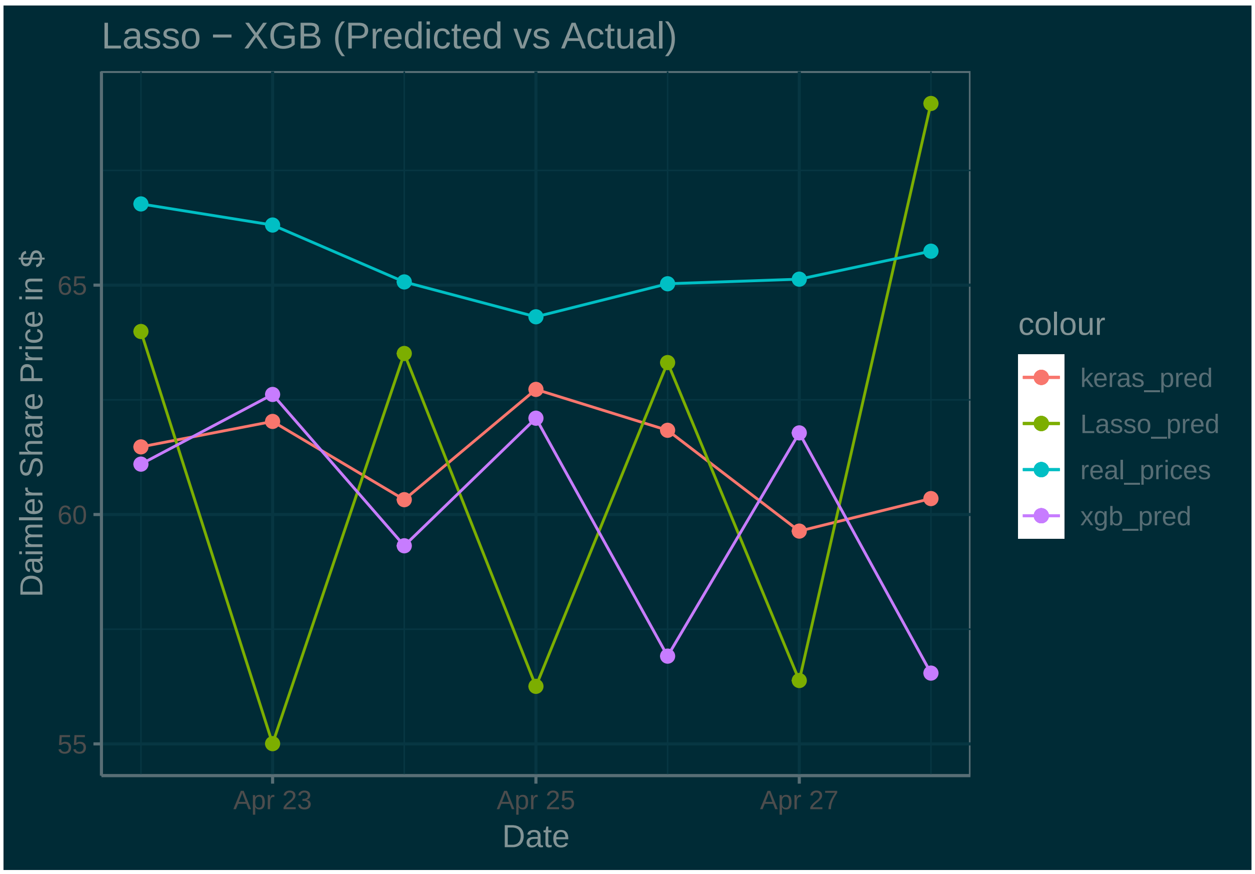
08.XGBoost model

We define the parameters that caret will use in the finding of hyperparameters

```
library(xgboost)
xgb_grid = expand.grid(
  nrounds = 1000,
```

```r
  eta = c(0.1, 0.05, 0.01),
  max_depth = c(2, 3, 4, 5, 6),
  gamma = 0,
  colsample_bytree=1,
  min_child_weight=c(1, 2, 3, 4 ,5),
  subsample=1)

# With the 5 fold cross validation, caret package can find the optimal hyperparameters
# for our model (takes a lot of time...)
xgb_hyperparam <- train(DDAIF.Adjusted~., data = na.omit(train), method='xgbTree', trCon
```

09.Results of all models



	KERAS PRED	LASSO PRED	XGB PRED	REAL PRICES
2019-04-22	61.47425	63.98916	**61.09808**	66.77
2019-04-23	62.02896	55.00517	**62.61695**	66.31
2019-04-24	60.32361	63.50995	**59.31833**	65.07
2019-04-25	62.72711	56.25471	**62.09966**	64.31
2019-04-26	61.83586	63.31208	**56.91272**	65.03
2019-04-29	59.63830	56.38149	**61.77945**	65.13
2019-04-30	60.34718	68.96010	**56.54294**	65.74

10.Conclusion

Usually the share price daily fluctuation is between 1 - 2 percent in ordinary time periods. Unfortunately the above models presented high daily fluctuation. Regardless from our application of the mathematical technical indicators. Into our datasets before the training of the models.

Unfortunately currently our models are not adequate to forecast time series of markets successfully.

11.Proposal

On another paper, i have also created some models to analyze the correlation of social media sentiment and Daimler share price. There my models forecasting was significantly more successful. I would propose to create a model that would analyze and combine the results of:

-Social media sentiment
-Economic News
-Market Time Series Analysis

Thank you for reading my analysis
KR
Niko

REFERENCES

[1] Ruoxuan Xiong, Eric P. Nichols, Yuan Shen. Deep Learning Stock Volatility with Google Domestic Trends. Cornell University (arXiv:1512.04916v3) [q-fin.CP] (2016)

[2] Junran Wu, Ke Xu, Jichang Zhao.Online reviews can predict long-term returns of individual stocks . Cornell University (2019)