ds-4100-project-modeling.Rmd

Veronica Shei, Edward Wang, Ethan Tang

3/23/2019

```
# initialize empty vector for holding all observations
master_dataset = c()
# database contains 619622 observations
# iterate over all the stocks and pull in the features and responses
for (ticker in dbListTables(db)) {
  curr_data = dbReadTable(db,ticker)
  master_dataset = rbind(master_dataset, curr_data[,8:length(curr_data)])
}
# out put the number of observations
num_og_obs = dim(master_dataset)[1]
print(num_og_obs)
```

```
## [1] 619622
```

```
# current number of observations
dim(master_dataset)[1]
```

```
## [1] 619622
```

```
# Remove all rows with an NA
# This can result due to division by zero or transformations resulting from indicator calculations,
master_dataset = na.omit(master_dataset)
# current number of observations
dim(master_dataset)[1]
```

```
## [1] 501293
```

```
# Remove all volume change deltas equal to zero as this is faulty data
master_dataset = master_dataset[master_dataset$PERCENT_VOL != 0,]
# current number of observations
dim(master_dataset)[1]
```

```
## [1] 500754
```

```
# Only keep perent volume under 1000% to remove outliers
master_dataset = master_dataset[master_dataset$PERCENT_VOL < 1000,]
dim(master_dataset)[1]
```

```
## [1] 500638
```

```
# Normalisation function for SMA, EWMA, Percent Price, Percent Vol
# Takes the takes value subtracts minimum, and divides by the range
normalize <- function(ticker_df, type) {
  if(type == 'SMA') {
    sma <- ticker_df$SMA
    return ((sma - min(sma)) / (max(sma) - min(sma)))
  } else if (type == 'EWMA') {
      ewma <- ticker_df$EWMA
      return ((ewma - min(ewma)) / (max(ewma) - min(ewma)))
  } else if (type == 'PERCENT_PRICE') {
      price <- ticker_df$PERCENT_PRICE
```

```
    return ((price - min(price)) / (max(price) - min(price)))
  } else if (type == 'PERCENT_VOL') {
    vol <- ticker_df$PERCENT_VOL
    return ((vol - min(vol)) / (max(vol) - min(vol)))
  }
}
```

```
# Normalize the percent price, percent vol, sma, and ewma
master_dataset = cbind(master_dataset,
                       NORM_PERC_PRICE = normalize(master_dataset, "PERCENT_PRICE"),
                       NORM_PERC_VOL = normalize(master_dataset, "PERCENT_VOL"),
                       NORM_SMA = normalize(master_dataset, "SMA"),
                       NORM_EWMA = normalize(master_dataset, "EWMA")
                       )
```

```
# current number of observation
dim(master_dataset)[1]
```

```
## [1] 500638
```

```
# removes the 0 padding used in calculating the initial sma values
master_dataset = master_dataset[master_dataset$SMA != 0,]
# current number of observation
dim(master_dataset)[1]
```

```
## [1] 496700
```

```
# removes the 0 padding used in calculating the initial momentum values
master_dataset = master_dataset[master_dataset$MOM != 0,]
# current number of observation
dim(master_dataset)[1]
```

```
## [1] 488459
```

```
# removes the 0 padding used in calculating the initial rsi values
master_dataset = master_dataset[master_dataset$RSI != 0,]
# current number of observation
dim(master_dataset)[1]
```

```
## [1] 488458
```

```
# removes the 0 padding used in calculating the initial rsi values
master_dataset = master_dataset[master_dataset$RSI != 1,]
# current number of observation
dim(master_dataset)[1]
```

```
## [1] 488447
```

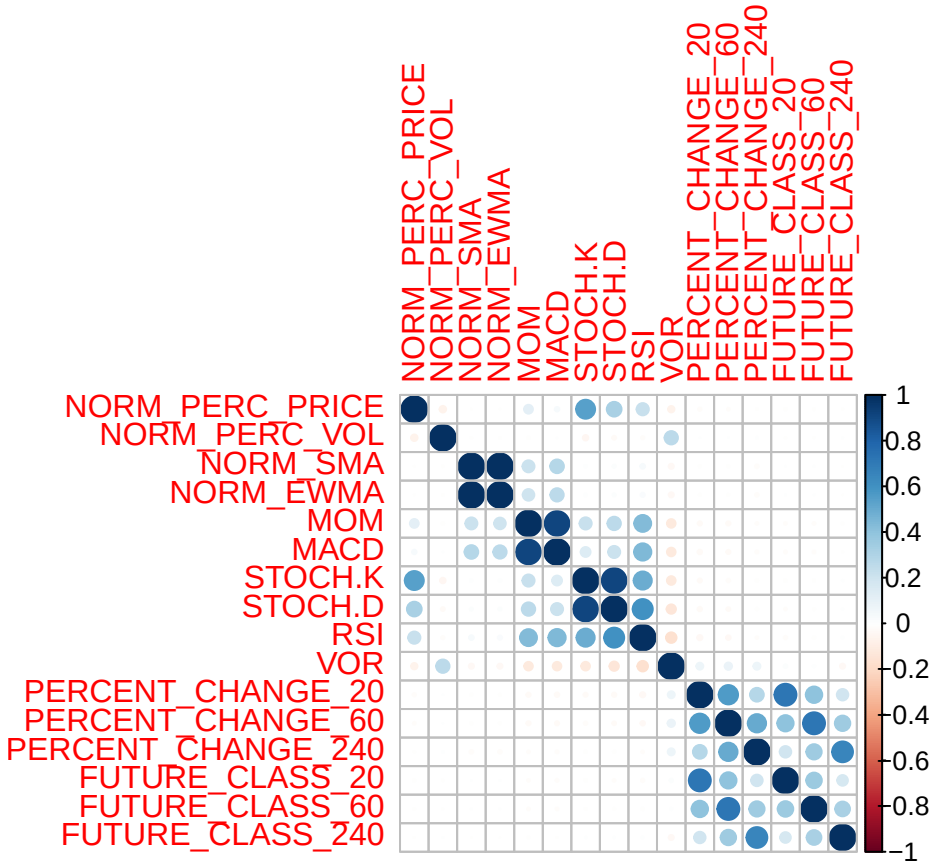
```
# determining a training size based on 70% of the dataset
training_perc = .70
training_size = round(dim(master_dataset)[[1]] * training_perc,0)

# creating a random index with the determined training size for selecting the training/testing set
training_idx = sample(nrow(master_dataset),size=training_size,replace=FALSE)
train_df = master_dataset[training_idx,]
test_df = master_dataset[-training_idx,]
```

```
# Selecting only relevant features and response variables
master_dataset <- master_dataset %>%
```

```
  select(NORM_PERC_PRICE, NORM_PERC_VOL, NORM_SMA, NORM_EWMA, MOM, MACD, STOCH.K, STOCH.D, RSI, VOR,
         PERCENT_CHANGE_20, PERCENT_CHANGE_60, PERCENT_CHANGE_240,
         FUTURE_CLASS_20, FUTURE_CLASS_60, FUTURE_CLASS_240)
```

```
# Looking for correlations
correlations <-  cor(master_dataset)
corrplot(correlations,"circle")
```




```
# Linear Model Short
pred_lm_short <- step(lm(PERCENT_CHANGE_20 ~ NORM_PERC_PRICE + NORM_PERC_VOL + NORM_SMA + NORM_EWMA + M
```

```
## Start:  AIC=1328278
## PERCENT_CHANGE_20 ~ NORM_PERC_PRICE + NORM_PERC_VOL + NORM_SMA +
##     NORM_EWMA + MOM + MACD + STOCH.K + STOCH.D + RSI + VOR
##
##                   Df Sum of Sq      RSS     AIC
## - STOCH.D          1        10 16636151 1328276
## - STOCH.K          1        24 16636165 1328276
## <none>                         16636141 1328278
## - NORM_PERC_PRICE  1       401 16636542 1328284
## - NORM_EWMA        1       418 16636559 1328284
## - NORM_SMA         1       419 16636560 1328284
## - MOM              1       829 16636970 1328293
## - MACD             1      1886 16638026 1328314
## - RSI              1      2854 16638995 1328334
## - NORM_PERC_VOL    1      5878 16642019 1328396
## - VOR              1     78788 16714929 1329891
```

```
##
## Step:  AIC=1328276
## PERCENT_CHANGE_20 ~ NORM_PERC_PRICE + NORM_PERC_VOL + NORM_SMA +
##     NORM_EWMA + MOM + MACD + STOCH.K + RSI + VOR
##
##                   Df Sum of Sq      RSS     AIC
## - STOCH.K          1        23 16636174 1328274
## <none>                         16636151 1328276
## + STOCH.D          1        10 16636141 1328278
## - NORM_EWMA        1       416 16636567 1328282
## - NORM_SMA         1       416 16636568 1328282
## - NORM_PERC_PRICE  1       674 16636825 1328288
## - MOM              1       827 16636979 1328291
## - MACD             1      1905 16638057 1328313
## - RSI              1      3445 16639596 1328345
## - NORM_PERC_VOL    1      5869 16642020 1328394
## - VOR              1     78780 16714932 1329889
##
## Step:  AIC=1328274
## PERCENT_CHANGE_20 ~ NORM_PERC_PRICE + NORM_PERC_VOL + NORM_SMA +
##     NORM_EWMA + MOM + MACD + RSI + VOR
##
##                   Df Sum of Sq      RSS     AIC
## <none>                         16636174 1328274
## + STOCH.K          1        23 16636151 1328276
## + STOCH.D          1         9 16636165 1328276
## - NORM_EWMA        1       522 16636696 1328283
## - NORM_SMA         1       523 16636697 1328283
## - MOM              1       809 16636983 1328289
## - NORM_PERC_PRICE  1      1039 16637213 1328294
## - MACD             1      1975 16638149 1328313
## - RSI              1      4691 16640865 1328369
## - NORM_PERC_VOL    1      5863 16642037 1328393
## - VOR              1     78797 16714971 1329888
```

```
# Model removed no variables
# Model dominated by SMA, EWMA, MOM, RSI
print(summary(pred_lm_short))
```

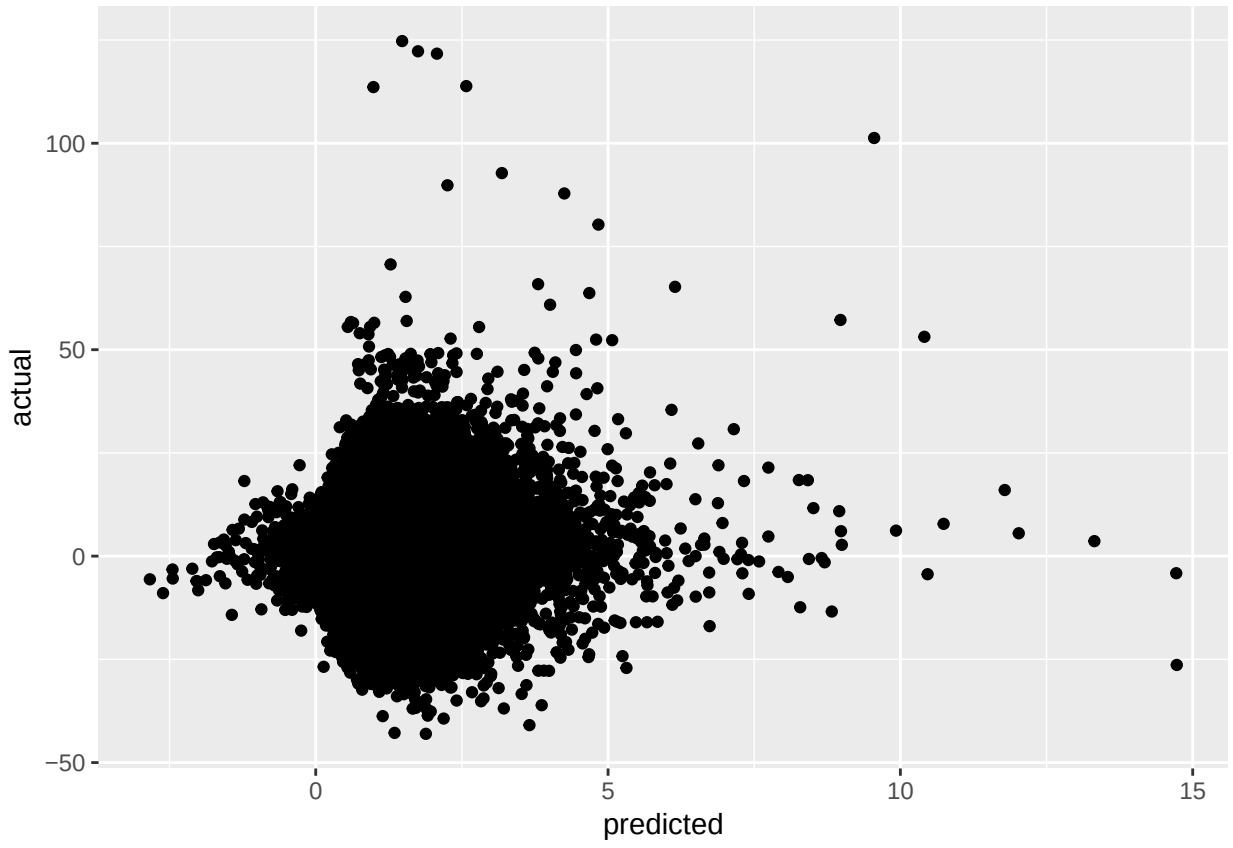
```
##
## Call:
## lm(formula = PERCENT_CHANGE_20 ~ NORM_PERC_PRICE + NORM_PERC_VOL +
##     NORM_SMA + NORM_EWMA + MOM + MACD + RSI + VOR, data = train_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.561  -3.857   0.062   3.862 128.361
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.190e+00  2.732e-01   8.017 1.09e-15 ***
## NORM_PERC_PRICE -3.149e+00  6.815e-01  -4.621 3.82e-06 ***
## NORM_PERC_VOL   -2.825e+00  2.574e-01 -10.977  < 2e-16 ***
## NORM_SMA         1.157e+02  3.531e+01   3.278  0.00105 **
## NORM_EWMA       -1.130e+02  3.449e+01  -3.276  0.00105 **
```

```
## MOM              1.276e-02  3.129e-03   4.077 4.56e-05 ***
## MACD            -1.280e-01  2.009e-02  -6.371 1.88e-10 ***
## RSI             -8.343e-01  8.497e-02  -9.819  < 2e-16 ***
## VOR              3.891e+01  9.669e-01  40.242  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.975 on 341904 degrees of freedom
## Multiple R-squared:  0.006485,   Adjusted R-squared:  0.006461
## F-statistic: 278.9 on 8 and 341904 DF,  p-value: < 2.2e-16
```

```
# testing model against test dataset
# rounding applies a threshold of 50% probability for buy
predicted <- predict(pred_lm_short, test_df, type="response")
# place predictions and actual class into a dataframe
results <- data.frame(predicted, actual=test_df$PERCENT_CHANGE_20)
head(results)
```

```
##    predicted    actual
## 28 0.7919676  2.434246
## 31 0.6848999  1.306823
## 33 0.7192153  2.126029
## 36 0.7057885 -1.192874
## 39 0.5130089 -3.800932
## 41 0.6418055 -2.560806
```

```
# plotting the predicted against the actual
ggplot(results) + geom_point(aes(x=predicted, y=actual))
```



```r
# Linear Model Medium
pred_lm_med <- step(lm(PERCENT_CHANGE_60 ~ NORM_PERC_PRICE + NORM_PERC_VOL + NORM_SMA + NORM_EWMA + MOM
```

```
## Start:  AIC=1697189
## PERCENT_CHANGE_60 ~ NORM_PERC_PRICE + NORM_PERC_VOL + NORM_SMA +
##     NORM_EWMA + MOM + MACD + STOCH.K + STOCH.D + RSI + VOR
##
##                   Df Sum of Sq      RSS     AIC
## - STOCH.D          1         2 48937344 1697187
## - STOCH.K          1        22 48937363 1697187
## - NORM_PERC_PRICE  1       139 48937480 1697188
## - MACD             1       257 48937599 1697189
## <none>                         48937342 1697189
## - NORM_EWMA        1       297 48937638 1697189
## - NORM_SMA         1       299 48937640 1697189
## - MOM              1      1224 48938565 1697196
## - RSI              1      2133 48939475 1697202
## - NORM_PERC_VOL    1     24228 48961570 1697356
## - VOR              1    314593 49251935 1699378
##
## Step:  AIC=1697187
## PERCENT_CHANGE_60 ~ NORM_PERC_PRICE + NORM_PERC_VOL + NORM_SMA +
##     NORM_EWMA + MOM + MACD + STOCH.K + RSI + VOR
##
##                   Df Sum of Sq      RSS     AIC
## - STOCH.K          1        71 48937415 1697186
## - NORM_PERC_PRICE  1       221 48937565 1697187
## - MACD             1       276 48937620 1697187
## <none>                         48937344 1697187
## - NORM_EWMA        1       338 48937681 1697188
## - NORM_SMA         1       340 48937683 1697188
## + STOCH.D          1         2 48937342 1697189
## - MOM              1      1225 48938568 1697194
## - RSI              1      2655 48939998 1697204
## - NORM_PERC_VOL    1     24236 48961580 1697354
## - VOR              1    314661 49252004 1699377
##
## Step:  AIC=1697186
## PERCENT_CHANGE_60 ~ NORM_PERC_PRICE + NORM_PERC_VOL + NORM_SMA +
##     NORM_EWMA + MOM + MACD + RSI + VOR
##
##                   Df Sum of Sq      RSS     AIC
## - MACD             1       249 48937664 1697185
## - NORM_EWMA        1       276 48937691 1697186
## - NORM_SMA         1       278 48937693 1697186
## <none>                         48937415 1697186
## - NORM_PERC_PRICE  1       467 48937882 1697187
## + STOCH.K          1        71 48937344 1697187
## + STOCH.D          1        51 48937363 1697187
## - MOM              1      1292 48938706 1697193
## - RSI              1      3883 48941298 1697211
## - NORM_PERC_VOL    1     24216 48961631 1697353
## - VOR              1    314720 49252134 1699375
##
```

```
## Step:  AIC=1697185
## PERCENT_CHANGE_60 ~ NORM_PERC_PRICE + NORM_PERC_VOL + NORM_SMA +
##     NORM_EWMA + MOM + RSI + VOR
##
##                    Df Sum of Sq      RSS     AIC
## - NORM_EWMA         1        59 48937723 1697184
## - NORM_SMA          1        60 48937724 1697184
## <none>                          48937664 1697185
## + MACD              1       249 48937415 1697186
## + STOCH.K           1        44 48937620 1697187
## - NORM_PERC_PRICE   1       553 48938217 1697187
## + STOCH.D           1        18 48937646 1697187
## - MOM               1      1271 48938935 1697192
## - RSI               1      3697 48941361 1697209
## - NORM_PERC_VOL     1     24135 48961799 1697352
## - VOR               1    314568 49252232 1699374
##
## Step:  AIC=1697184
## PERCENT_CHANGE_60 ~ NORM_PERC_PRICE + NORM_PERC_VOL + NORM_SMA +
##     MOM + RSI + VOR
##
##                    Df Sum of Sq      RSS     AIC
## - NORM_SMA          1        65 48937788 1697182
## <none>                          48937723 1697184
## - NORM_PERC_PRICE   1       501 48938224 1697185
## + NORM_EWMA         1        59 48937664 1697185
## + MACD              1        32 48937691 1697186
## + STOCH.K           1        15 48937708 1697186
## + STOCH.D           1         1 48937722 1697186
## - RSI               1      3950 48941673 1697209
## - MOM               1      4203 48941926 1697211
## - NORM_PERC_VOL     1     24130 48961853 1697350
## - VOR               1    314662 49252385 1699373
##
## Step:  AIC=1697182
## PERCENT_CHANGE_60 ~ NORM_PERC_PRICE + NORM_PERC_VOL + MOM + RSI +
##     VOR
##
##                    Df Sum of Sq      RSS     AIC
## <none>                          48937788 1697182
## - NORM_PERC_PRICE   1       496 48938285 1697184
## + NORM_SMA          1        65 48937723 1697184
## + NORM_EWMA         1        64 48937724 1697184
## + STOCH.K           1        15 48937773 1697184
## + MACD              1        14 48937774 1697184
## + STOCH.D           1         1 48937787 1697184
## - RSI               1      3901 48941689 1697208
## - MOM               1      4633 48942421 1697213
## - NORM_PERC_VOL     1     24138 48961927 1697349
## - VOR               1    315086 49252874 1699375
```

```r
# Model removed SMA, EWMA, MACD, and Percent Price
# Model dominated by volatility ratio
print(summary(pred_lm_med))
```

```
##
## Call:
## lm(formula = PERCENT_CHANGE_60 ~ NORM_PERC_PRICE + NORM_PERC_VOL +
##     MOM + RSI + VOR, data = train_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -94.498  -7.011   0.168   6.820 250.779
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.522626   0.451813   7.797 6.38e-15 ***
## NORM_PERC_PRICE -2.111714   1.133944  -1.862   0.0626 .
## NORM_PERC_VOL   -5.732116   0.441396 -12.986  < 2e-16 ***
## MOM             -0.013117   0.002306  -5.689 1.28e-08 ***
## RSI             -0.743396   0.142397  -5.221 1.78e-07 ***
## VOR             77.652736   1.655046  46.919  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.96 on 341907 degrees of freedom
## Multiple R-squared:  0.007439,   Adjusted R-squared:  0.007425
## F-statistic: 512.5 on 5 and 341907 DF,  p-value: < 2.2e-16
```

```
# Linear Model Medium removed vor
pred_lm_med <- step(lm(PERCENT_CHANGE_60 ~ NORM_PERC_PRICE + NORM_PERC_VOL + NORM_SMA + NORM_EWMA + MOM
```

```
## Start:  AIC=1699378
## PERCENT_CHANGE_60 ~ NORM_PERC_PRICE + NORM_PERC_VOL + NORM_SMA +
##     NORM_EWMA + MOM + MACD + STOCH.K + STOCH.D + RSI
##
##                   Df Sum of Sq      RSS     AIC
## - STOCH.K          1       8.8 49251944 1699376
## - NORM_PERC_VOL    1      11.6 49251946 1699376
## - MACD             1      45.1 49251980 1699376
## - STOCH.D          1      69.3 49252004 1699377
## - NORM_SMA         1      73.5 49252008 1699377
## - NORM_EWMA        1      76.0 49252011 1699377
## <none>                         49251935 1699378
## - NORM_PERC_PRICE  1     354.4 49252289 1699379
## - MOM              1    1326.6 49253261 1699385
## - RSI              1    9286.8 49261222 1699441
##
## Step:  AIC=1699376
## PERCENT_CHANGE_60 ~ NORM_PERC_PRICE + NORM_PERC_VOL + NORM_SMA +
##     NORM_EWMA + MOM + MACD + STOCH.D + RSI
##
##                   Df Sum of Sq      RSS     AIC
## - NORM_PERC_VOL    1      12.2 49251956 1699374
## - MACD             1      51.7 49251995 1699375
## - NORM_SMA         1      82.9 49252027 1699375
## - NORM_EWMA        1      85.6 49252029 1699375
## - STOCH.D          1     190.9 49252134 1699375
## <none>                         49251944 1699376
## - NORM_PERC_PRICE  1     504.9 49252448 1699378
```

```
## + STOCH.K          1       8.8 49251935 1699378
## - MOM              1    1319.3 49253263 1699383
## - RSI              1   10144.1 49262088 1699445
##
## Step:  AIC=1699374
## PERCENT_CHANGE_60 ~ NORM_PERC_PRICE + NORM_SMA + NORM_EWMA +
##     MOM + MACD + STOCH.D + RSI
##
##                   Df Sum of Sq      RSS     AIC
## - MACD             1      52.0 49252008 1699373
## - NORM_SMA         1      82.8 49252039 1699373
## - NORM_EWMA        1      85.5 49252041 1699373
## - STOCH.D          1     191.0 49252147 1699374
## <none>                         49251956 1699374
## - NORM_PERC_PRICE  1     497.3 49252453 1699376
## + NORM_PERC_VOL    1      12.2 49251944 1699376
## + STOCH.K          1       9.3 49251946 1699376
## - MOM              1    1317.3 49253273 1699381
## - RSI              1   10143.3 49262099 1699443
##
## Step:  AIC=1699373
## PERCENT_CHANGE_60 ~ NORM_PERC_PRICE + NORM_SMA + NORM_EWMA +
##     MOM + STOCH.D + RSI
##
##                   Df Sum of Sq      RSS     AIC
## - NORM_SMA         1      31.9 49252040 1699371
## - NORM_EWMA        1      34.3 49252042 1699371
## - STOCH.D          1     236.5 49252244 1699372
## <none>                         49252008 1699373
## - NORM_PERC_PRICE  1     461.7 49252469 1699374
## + MACD             1      52.0 49251956 1699374
## + STOCH.K          1      16.2 49251992 1699374
## + NORM_PERC_VOL    1      12.5 49251995 1699375
## - MOM              1    3347.9 49255356 1699394
## - RSI              1   10144.5 49262152 1699441
##
## Step:  AIC=1699371
## PERCENT_CHANGE_60 ~ NORM_PERC_PRICE + NORM_EWMA + MOM + STOCH.D +
##     RSI
##
##                   Df Sum of Sq      RSS     AIC
## <none>                         49252040 1699371
## - STOCH.D          1     352.4 49252392 1699371
## - NORM_EWMA        1     499.7 49252539 1699372
## - NORM_PERC_PRICE  1     503.8 49252543 1699372
## + NORM_SMA         1      31.9 49252008 1699373
## + STOCH.K          1      19.4 49252020 1699373
## + NORM_PERC_VOL    1      12.1 49252027 1699373
## + MACD             1       1.1 49252039 1699373
## - MOM              1    7629.1 49259669 1699422
## - RSI              1   11005.0 49263045 1699445
```

```
# Model removed Stochastics and percent volume
# Model dominated by SMA, EWMA, but R^2 drops significantly
```

```r
print(summary(pred_lm_med))
```

```
##
## Call:
## lm(formula = PERCENT_CHANGE_60 ~ NORM_PERC_PRICE + NORM_EWMA +
##     MOM + STOCH.D + RSI, data = train_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -73.123  -7.078   0.046   6.738 252.064
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.043407   0.458840  10.992  < 2e-16 ***
## NORM_PERC_PRICE -2.189742   1.170925  -1.870   0.0615 .
## NORM_EWMA       -0.774360   0.415775  -1.862   0.0625 .
## MOM             -0.017165   0.002359  -7.277 3.41e-13 ***
## STOCH.D         -0.169976   0.108667  -1.564   0.1178
## RSI             -1.471123   0.168311  -8.740  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12 on 341907 degrees of freedom
## Multiple R-squared:  0.001066,   Adjusted R-squared:  0.001051
## F-statistic: 72.96 on 5 and 341907 DF,  p-value: < 2.2e-16
```

```r
# Linear Model
pred_lm_long <- step(lm(PERCENT_CHANGE_240 ~ NORM_PERC_PRICE + NORM_PERC_VOL + NORM_SMA + NORM_EWMA + M
```

```
## Start:  AIC=2277462
## PERCENT_CHANGE_240 ~ NORM_PERC_PRICE + NORM_PERC_VOL + NORM_SMA +
##     NORM_EWMA + MOM + MACD + STOCH.K + STOCH.D + RSI + VOR
##
##                   Df Sum of Sq       RSS     AIC
## - MOM              1        75 267114250 2277460
## - NORM_PERC_PRICE  1        85 267114260 2277460
## - MACD             1       161 267114336 2277460
## - NORM_SMA         1       382 267114557 2277460
## - STOCH.K          1       451 267114626 2277460
## - NORM_EWMA        1       452 267114627 2277460
## - RSI              1       531 267114706 2277460
## - STOCH.D          1       629 267114804 2277461
## <none>                         267114175 2277462
## - NORM_PERC_VOL    1     70310 267184485 2277550
## - VOR              1   1429433 268543608 2279285
##
## Step:  AIC=2277460
## PERCENT_CHANGE_240 ~ NORM_PERC_PRICE + NORM_PERC_VOL + NORM_SMA +
##     NORM_EWMA + MACD + STOCH.K + STOCH.D + RSI + VOR
##
##                   Df Sum of Sq       RSS     AIC
## - NORM_PERC_PRICE  1        70 267114320 2277458
## - NORM_SMA         1       424 267114675 2277458
## - STOCH.K          1       470 267114720 2277459
```

```
## - NORM_EWMA         1       499 267114749 2277459
## - RSI               1       543 267114794 2277459
## - STOCH.D           1       634 267114885 2277459
## - MACD              1       686 267114937 2277459
## <none>                          267114250 2277460
## + MOM               1        75 267114175 2277462
## - NORM_PERC_VOL     1     70285 267184535 2277548
## - VOR               1   1429496 268543747 2279283
##
## Step:  AIC=2277458
## PERCENT_CHANGE_240 ~ NORM_PERC_VOL + NORM_SMA + NORM_EWMA + MACD +
##     STOCH.K + STOCH.D + RSI + VOR
##
##                   Df Sum of Sq       RSS     AIC
## - NORM_SMA         1       394 267114714 2277456
## - NORM_EWMA        1       466 267114786 2277457
## - STOCH.K          1       470 267114790 2277457
## - STOCH.D          1       607 267114927 2277457
## - MACD             1       655 267114976 2277457
## - RSI              1       665 267114985 2277457
## <none>                         267114320 2277458
## + NORM_PERC_PRICE  1        70 267114250 2277460
## + MOM              1        60 267114260 2277460
## - NORM_PERC_VOL    1     70427 267184747 2277546
## - VOR              1   1429477 268543797 2279281
##
## Step:  AIC=2277456
## PERCENT_CHANGE_240 ~ NORM_PERC_VOL + NORM_EWMA + MACD + STOCH.K +
##     STOCH.D + RSI + VOR
##
##                   Df Sum of Sq       RSS     AIC
## - MACD             1       320 267115035 2277455
## - STOCH.K          1       392 267115106 2277455
## - STOCH.D          1       396 267115110 2277455
## - RSI              1      1019 267115733 2277456
## <none>                         267114714 2277456
## + NORM_SMA         1       394 267114320 2277458
## + MOM              1       101 267114613 2277458
## + NORM_PERC_PRICE  1        40 267114675 2277458
## - NORM_EWMA        1     69571 267184285 2277544
## - NORM_PERC_VOL    1     70473 267185188 2277545
## - VOR              1   1435101 268549815 2279287
##
## Step:  AIC=2277455
## PERCENT_CHANGE_240 ~ NORM_PERC_VOL + NORM_EWMA + STOCH.K + STOCH.D +
##     RSI + VOR
##
##                   Df Sum of Sq       RSS     AIC
## - STOCH.K          1       342 267115377 2277453
## - STOCH.D          1       373 267115408 2277453
## - RSI              1       731 267115766 2277454
## <none>                         267115035 2277455
## + MOM              1       420 267114614 2277456
## + MACD             1       320 267114714 2277456
```

```
## + NORM_SMA          1        59 267114976 2277457
## + NORM_PERC_PRICE   1        45 267114990 2277457
## - NORM_PERC_VOL     1     70627 267185662 2277543
## - NORM_EWMA         1     77949 267192984 2277553
## - VOR               1   1439024 268554058 2279290
##
## Step:  AIC=2277453
## PERCENT_CHANGE_240 ~ NORM_PERC_VOL + NORM_EWMA + STOCH.D + RSI +
##     VOR
##
##                    Df Sum of Sq       RSS     AIC
## - STOCH.D           1        39 267115416 2277451
## - RSI               1       960 267116338 2277453
## <none>                          267115377 2277453
## + MOM               1       419 267114958 2277455
## + STOCH.K           1       342 267115035 2277455
## + MACD              1       271 267115106 2277455
## + NORM_PERC_PRICE   1        54 267115323 2277455
## + NORM_SMA          1        52 267115325 2277455
## - NORM_PERC_VOL     1     70287 267185664 2277541
## - NORM_EWMA         1     77934 267193312 2277551
## - VOR               1   1438683 268554060 2279288
##
## Step:  AIC=2277451
## PERCENT_CHANGE_240 ~ NORM_PERC_VOL + NORM_EWMA + RSI + VOR
##
##                    Df Sum of Sq       RSS     AIC
## <none>                          267115416 2277451
## - RSI               1      1871 267117287 2277452
## + MOM               1       416 267114999 2277453
## + MACD              1       285 267115130 2277453
## + NORM_SMA          1        73 267115342 2277453
## + STOCH.D           1        39 267115377 2277453
## + NORM_PERC_PRICE   1        32 267115384 2277453
## + STOCH.K           1         8 267115408 2277453
## - NORM_PERC_VOL     1     70314 267185730 2277539
## - NORM_EWMA         1     77968 267193384 2277549
## - VOR               1   1439808 268555223 2279287
```

```
# Model removed SMA, MACD, MOM, Stochastics
# Model dominated by volatility ratio
print(summary(pred_lm_long))
```

```
##
## Call:
## lm(formula = PERCENT_CHANGE_240 ~ NORM_PERC_VOL + NORM_EWMA +
##     RSI + VOR, data = train_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -177.96  -16.30   -1.43   13.44  562.23
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    10.1079     0.2083  48.523   <2e-16 ***
```

```
## NORM_PERC_VOL  -9.7652     1.0293  -9.487   <2e-16 ***
## NORM_EWMA      -9.4808     0.9490  -9.990   <2e-16 ***
## RSI             0.4575     0.2956   1.548    0.122
## VOR           165.9066     3.8646  42.930   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.95 on 341908 degrees of freedom
## Multiple R-squared:  0.005861,   Adjusted R-squared:  0.005849
## F-statistic: 503.9 on 4 and 341908 DF,  p-value: < 2.2e-16
```

```
# Logistic Model short
pred_glm_short <- step(glm(FUTURE_CLASS_20 ~ NORM_PERC_PRICE + NORM_PERC_VOL + NORM_SMA + NORM_EWMA + M
```

```
## Start:  AIC=466457.8
## FUTURE_CLASS_20 ~ NORM_PERC_PRICE + NORM_PERC_VOL + NORM_SMA +
##     NORM_EWMA + MOM + MACD + STOCH.K + STOCH.D + RSI + VOR
##
##                   Df Deviance    AIC
## - STOCH.D          1   466436 466456
## - STOCH.K          1   466437 466457
## <none>                 466436 466458
## - NORM_EWMA        1   466448 466468
## - NORM_SMA         1   466448 466468
## - NORM_PERC_PRICE  1   466450 466470
## - RSI              1   466451 466471
## - NORM_PERC_VOL    1   466457 466477
## - MOM              1   466463 466483
## - MACD             1   466479 466499
## - VOR              1   466542 466562
##
## Step:  AIC=466456
## FUTURE_CLASS_20 ~ NORM_PERC_PRICE + NORM_PERC_VOL + NORM_SMA +
##     NORM_EWMA + MOM + MACD + STOCH.K + RSI + VOR
##
##                   Df Deviance    AIC
## <none>                 466436 466456
## - STOCH.K          1   466439 466457
## + STOCH.D          1   466436 466458
## - NORM_EWMA        1   466450 466468
## - NORM_SMA         1   466450 466468
## - NORM_PERC_PRICE  1   466454 466472
## - NORM_PERC_VOL    1   466457 466475
## - RSI              1   466458 466476
## - MOM              1   466463 466481
## - MACD             1   466482 466500
## - VOR              1   466542 466560
```

```
# Model removed stochastics
# Model dominated by momentum
print(summary(pred_glm_short))
```

```
##
## Call:
## glm(formula = FUTURE_CLASS_20 ~ NORM_PERC_PRICE + NORM_PERC_VOL +
##     NORM_SMA + NORM_EWMA + MOM + MACD + STOCH.K + RSI + VOR,
```

```
##     family = binomial, data = train_df)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.516  -1.297   1.034   1.061   1.279
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      6.695e-01  8.868e-02   7.550 4.34e-14 ***
## NORM_PERC_PRICE -9.616e-01  2.270e-01  -4.237 2.27e-05 ***
## NORM_PERC_VOL   -3.435e-01  7.462e-02  -4.603 4.16e-06 ***
## NORM_SMA         4.100e+01  1.084e+01   3.783 0.000155 ***
## NORM_EWMA       -3.983e+01  1.059e+01  -3.763 0.000168 ***
## MOM              4.817e-03  9.231e-04   5.219 1.80e-07 ***
## MACD            -4.004e-02  5.951e-03  -6.728 1.72e-11 ***
## STOCH.K          2.678e-02  1.636e-02   1.636 0.101743
## RSI             -1.299e-01  2.775e-02  -4.681 2.86e-06 ***
## VOR              2.930e+00  2.866e-01  10.222  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 466698  on 341912  degrees of freedom
## Residual deviance: 466436  on 341903  degrees of freedom
## AIC: 466456
##
## Number of Fisher Scoring iterations: 4
```

```
# testing model against test dataset
# rounding applies a threshold of 50% probability for buy
predicted <- round(predict(pred_glm_short, test_df, type="response"),0)
# place predictions and actual class into a dataframe
results <- data.frame(predicted, actual=test_df$FUTURE_CLASS_20)
head(results)
```

```
##    predicted actual
## 28         1      1
## 31         1      1
## 33         1      1
## 36         1      0
## 39         1      0
## 41         1      0
```

```
# Confusion Matrix with stats
confusionMatrix(factor(predicted), factor(test_df$FUTURE_CLASS_20))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0    29    29
##          1 62665 83811
##
##                Accuracy : 0.5722
##                  95% CI : (0.5696, 0.5747)
```

```
##     No Information Rate : 0.5722
##     P-Value [Acc > NIR] : 0.5011
##
##                   Kappa : 1e-04
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.0004626
##             Specificity : 0.9996541
##          Pos Pred Value : 0.5000000
##          Neg Pred Value : 0.5721825
##              Prevalence : 0.4278461
##          Detection Rate : 0.0001979
##    Detection Prevalence : 0.0003958
##       Balanced Accuracy : 0.5000583
##
##        'Positive' Class : 0
##
```

```
# Model almost always predicts true
# Accuracy of .5707
```

```
# Logistic model medium
pred_glm_med <- step(glm(FUTURE_CLASS_60 ~ NORM_PERC_PRICE + NORM_PERC_VOL + NORM_SMA + NORM_EWMA + MOM
```

```
## Start:  AIC=452436.5
## FUTURE_CLASS_60 ~ NORM_PERC_PRICE + NORM_PERC_VOL + NORM_SMA +
##     NORM_EWMA + MOM + MACD + STOCH.K + STOCH.D + RSI + VOR
##
##                   Df Deviance    AIC
## - NORM_EWMA        1   452415 452435
## - NORM_SMA         1   452415 452435
## - RSI              1   452415 452435
## - MACD             1   452415 452435
## - MOM              1   452416 452436
## - STOCH.K          1   452417 452437
## <none>                 452415 452437
## - STOCH.D          1   452417 452437
## - NORM_PERC_VOL    1   452419 452439
## - NORM_PERC_PRICE  1   452420 452440
## - VOR              1   452433 452453
##
## Step:  AIC=452434.6
## FUTURE_CLASS_60 ~ NORM_PERC_PRICE + NORM_PERC_VOL + NORM_SMA +
##     MOM + MACD + STOCH.K + STOCH.D + RSI + VOR
##
##                   Df Deviance    AIC
## - RSI              1   452415 452433
## - MACD             1   452415 452433
## - MOM              1   452417 452435
## <none>                 452415 452435
## - STOCH.K          1   452417 452435
## - STOCH.D          1   452417 452435
## - NORM_SMA         1   452418 452436
## + NORM_EWMA        1   452415 452437
```

```
## - NORM_PERC_VOL    1   452420 452438
## - NORM_PERC_PRICE  1   452420 452438
## - VOR              1   452433 452451
##
## Step:  AIC=452433
## FUTURE_CLASS_60 ~ NORM_PERC_PRICE + NORM_PERC_VOL + NORM_SMA +
##     MOM + MACD + STOCH.K + STOCH.D + VOR
##
##                   Df Deviance    AIC
## - MACD             1   452416 452432
## - STOCH.K          1   452417 452433
## <none>                 452415 452433
## - MOM              1   452417 452433
## - STOCH.D          1   452417 452433
## - NORM_SMA         1   452418 452434
## + RSI              1   452415 452435
## + NORM_EWMA        1   452415 452435
## - NORM_PERC_VOL    1   452420 452436
## - NORM_PERC_PRICE  1   452420 452436
## - VOR              1   452433 452449
##
## Step:  AIC=452431.6
## FUTURE_CLASS_60 ~ NORM_PERC_PRICE + NORM_PERC_VOL + NORM_SMA +
##     MOM + STOCH.K + STOCH.D + VOR
##
##                   Df Deviance    AIC
## <none>                 452416 452432
## - STOCH.K          1   452418 452432
## - NORM_SMA         1   452418 452432
## - STOCH.D          1   452418 452432
## + MACD             1   452415 452433
## + RSI              1   452415 452433
## + NORM_EWMA        1   452415 452433
## - NORM_PERC_PRICE  1   452420 452434
## - NORM_PERC_VOL    1   452421 452435
## - VOR              1   452434 452448
## - MOM              1   452442 452456
```

```r
# Model removed MOM, RSI, EWMA, MACD
# Model is not heavily dominated
print(summary(pred_glm_med))
```

```
##
## Call:
## glm(formula = FUTURE_CLASS_60 ~ NORM_PERC_PRICE + NORM_PERC_VOL +
##     NORM_SMA + MOM + STOCH.K + STOCH.D + VOR, family = binomial,
##     data = train_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9369  -1.3952   0.9642   0.9730   1.1950
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      0.7385416  0.1066411   6.925 4.34e-12 ***
```

```
## NORM_PERC_PRICE -0.5797029  0.2675420  -2.167   0.0303 *
## NORM_PERC_VOL   -0.1695735  0.0761448  -2.227   0.0259 *
## NORM_SMA         0.1152364  0.0737778   1.562   0.1183
## MOM             -0.0019950  0.0003876  -5.146 2.65e-07 ***
## STOCH.K          0.0560799  0.0392935   1.427   0.1535
## STOCH.D         -0.0681715  0.0435016  -1.567   0.1171
## VOR              1.2214444  0.2883398   4.236 2.27e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 452485  on 341912  degrees of freedom
## Residual deviance: 452416  on 341905  degrees of freedom
## AIC: 452432
##
## Number of Fisher Scoring iterations: 4
```

```
# testing model against test dataset
# rounding applies a threshold of 50% probability for buy
predicted <- round(predict(pred_glm_med, test_df, type="response"),0)
# place predictions and actual class into a dataframe
results <- data.frame(predicted, actual=test_df$FUTURE_CLASS_60)
head(results)
```

```
##    predicted actual
## 28         1      1
## 31         1      1
## 33         1      1
## 36         1      0
## 39         1      0
## 41         1      0
```

```
# Confusion Matrix with stats
confusionMatrix(factor(predicted), factor(test_df$FUTURE_CLASS_60))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0     3     2
##          1 55187 91342
##
##                Accuracy : 0.6234
##                  95% CI : (0.6209, 0.6259)
##     No Information Rate : 0.6234
##     P-Value [Acc > NIR] : 0.499
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 5.436e-05
##             Specificity : 1.000e+00
##          Pos Pred Value : 6.000e-01
##          Neg Pred Value : 6.234e-01
```

```
##              Prevalence : 3.766e-01
##          Detection Rate : 2.047e-05
##    Detection Prevalence : 3.412e-05
##       Balanced Accuracy : 5.000e-01
##
##        'Positive' Class : 0
##
```

```
# Model almost always predicts true
# Accuracy of .6244
```

```
# logistic model long
pred_glm_long <- step(glm(FUTURE_CLASS_240 ~ NORM_PERC_PRICE + NORM_PERC_VOL + NORM_SMA + NORM_EWMA + M
```

```
## Start:  AIC=423918.8
## FUTURE_CLASS_240 ~ NORM_PERC_PRICE + NORM_PERC_VOL + NORM_SMA +
##     NORM_EWMA + MOM + MACD + STOCH.K + STOCH.D + RSI + VOR
##
##                   Df Deviance    AIC
## - NORM_SMA         1   423897 423917
## - NORM_EWMA        1   423897 423917
## - MACD             1   423898 423918
## <none>                 423897 423919
## - STOCH.K          1   423900 423920
## - STOCH.D          1   423900 423920
## - NORM_PERC_PRICE  1   423900 423920
## - RSI              1   423903 423923
## - MOM              1   423909 423929
## - NORM_PERC_VOL    1   423943 423963
## - VOR              1   424408 424428
##
## Step:  AIC=423916.8
## FUTURE_CLASS_240 ~ NORM_PERC_PRICE + NORM_PERC_VOL + NORM_EWMA +
##     MOM + MACD + STOCH.K + STOCH.D + RSI + VOR
##
##                   Df Deviance    AIC
## - NORM_EWMA        1   423897 423915
## <none>                 423897 423917
## - MACD             1   423899 423917
## - STOCH.K          1   423900 423918
## - NORM_PERC_PRICE  1   423901 423919
## + NORM_SMA         1   423897 423919
## - STOCH.D          1   423901 423919
## - RSI              1   423904 423922
## - MOM              1   423910 423928
## - NORM_PERC_VOL    1   423943 423961
## - VOR              1   424408 424426
##
## Step:  AIC=423915
## FUTURE_CLASS_240 ~ NORM_PERC_PRICE + NORM_PERC_VOL + MOM + MACD +
##     STOCH.K + STOCH.D + RSI + VOR
##
##                   Df Deviance    AIC
## <none>                 423897 423915
## - MACD             1   423899 423915
```

```
## - STOCH.K          1   423900 423916
## - NORM_PERC_PRICE  1   423901 423917
## + NORM_EWMA        1   423897 423917
## + NORM_SMA         1   423897 423917
## - STOCH.D          1   423901 423917
## - RSI              1   423904 423920
## - MOM              1   423910 423926
## - NORM_PERC_VOL    1   423943 423959
## - VOR              1   424408 424424
```

```
# Model removed SMA, EWMA, MACD
# Model dominated by volatility ratio
print(summary(pred_glm_long))
```

```
##
## Call:
## glm(formula = FUTURE_CLASS_240 ~ NORM_PERC_PRICE + NORM_PERC_VOL +
##     MOM + MACD + STOCH.K + STOCH.D + RSI + VOR, family = binomial,
##     data = train_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7271  -1.5033   0.8467   0.8650   1.9479
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.0571675  0.1096436   9.642  < 2e-16 ***
## NORM_PERC_PRICE -0.5437163  0.2796228  -1.944 0.051840 .
## NORM_PERC_VOL    0.5418652  0.0803286   6.746 1.52e-11 ***
## MOM             -0.0033964  0.0009588  -3.542 0.000397 ***
## MACD             0.0064351  0.0043128   1.492 0.135679
## STOCH.K          0.0762945  0.0422542   1.806 0.070979 .
## STOCH.D         -0.1000811  0.0502114  -1.993 0.046240 *
## RSI              0.0858657  0.0321261   2.673 0.007523 **
## VOR             -6.6586534  0.2939248 -22.654  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 424448  on 341912  degrees of freedom
## Residual deviance: 423897  on 341904  degrees of freedom
## AIC: 423915
##
## Number of Fisher Scoring iterations: 4
```

```
# testing model against test dataset
# rounding applies a threshold of 50% probability for buy
predicted <- round(predict(pred_glm_long, test_df, type="response"),0)
# place predictions and actual class into a dataframe
results <- data.frame(predicted, actual=test_df$FUTURE_CLASS_240)
head(results)
```

```
##    predicted actual
## 28         1      1
## 31         1      1
```

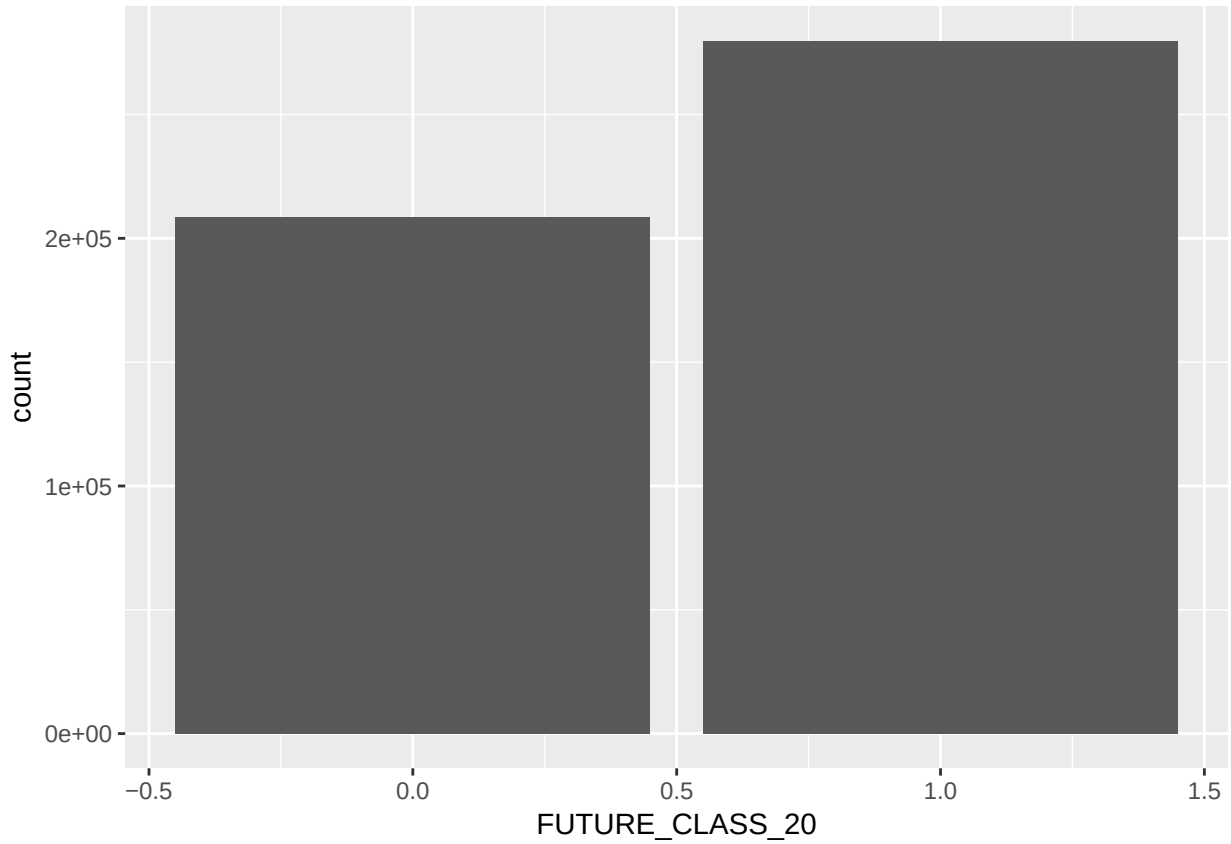
```
## 33         1      1
## 36         1      1
## 39         1      1
## 41         1      1
```

```
# Confusion Matrix with stats
confusionMatrix(factor(predicted), factor(test_df$FUTURE_CLASS_240))
```

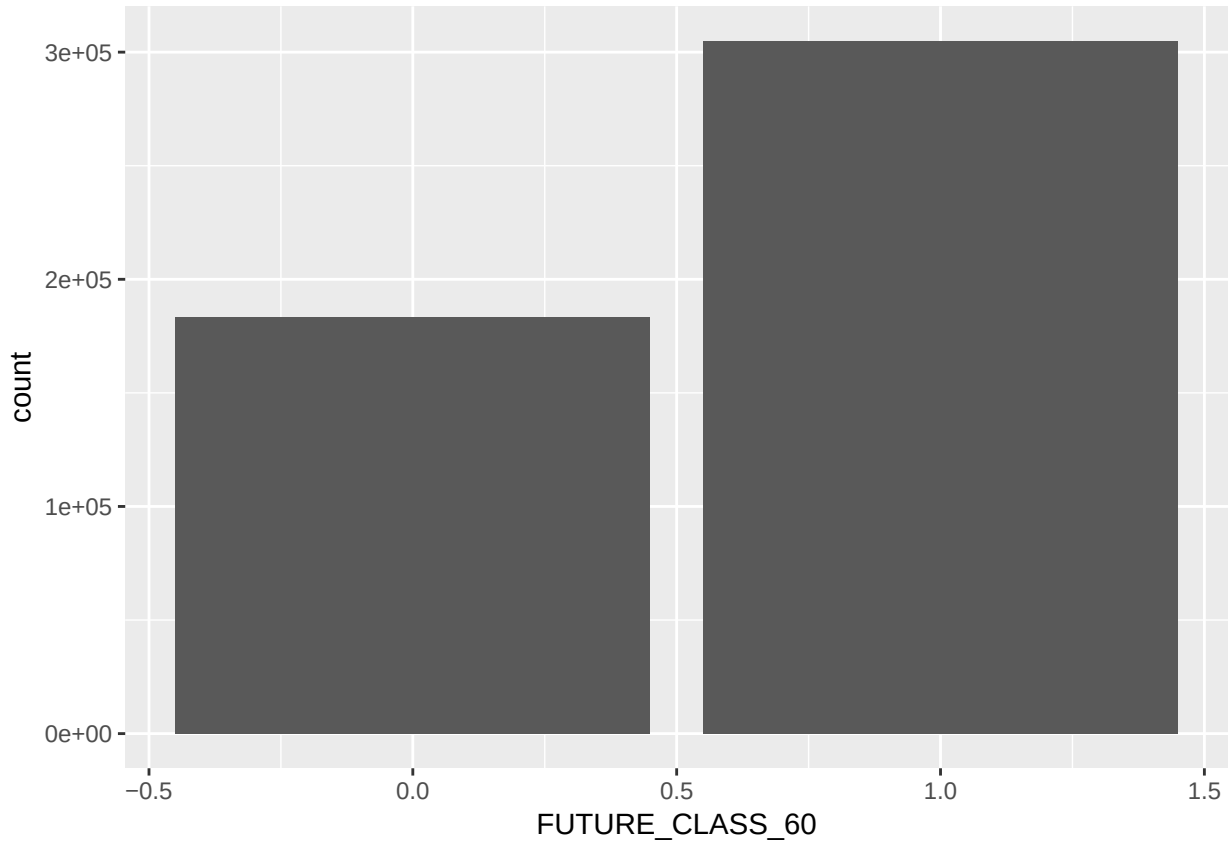
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction      0      1
##          0     23     42
##          1  46036 100433
##
##                Accuracy : 0.6855
##                  95% CI : (0.6832, 0.6879)
##     No Information Rate : 0.6857
##     P-Value [Acc > NIR] : 0.5438
##
##                   Kappa : 1e-04
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.0004994
##             Specificity : 0.9995820
##          Pos Pred Value : 0.3538462
##          Neg Pred Value : 0.6856946
##              Prevalence : 0.3143230
##          Detection Rate : 0.0001570
##    Detection Prevalence : 0.0004436
##       Balanced Accuracy : 0.5000407
##
##        'Positive' Class : 0
##
```

```
# Model almost always predicts buy
# Accuracy of .6856
```

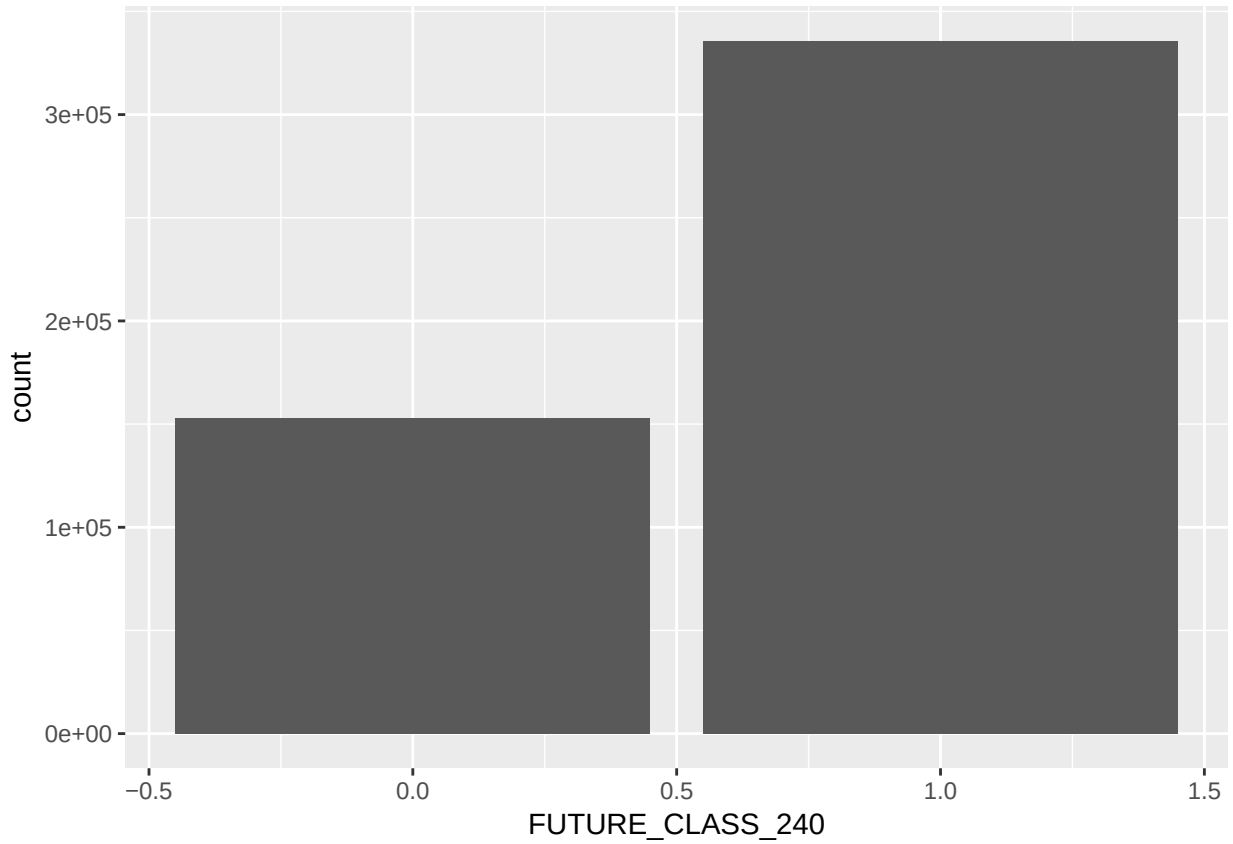
```
# Distribution of binary response variables
# Note that as time frame gets longer, more stocks are "buys" due to bull market
ggplot(master_dataset) + geom_bar(aes(x=FUTURE_CLASS_20))
```




```
ggplot(master_dataset) + geom_bar(aes(x=FUTURE_CLASS_60))
```

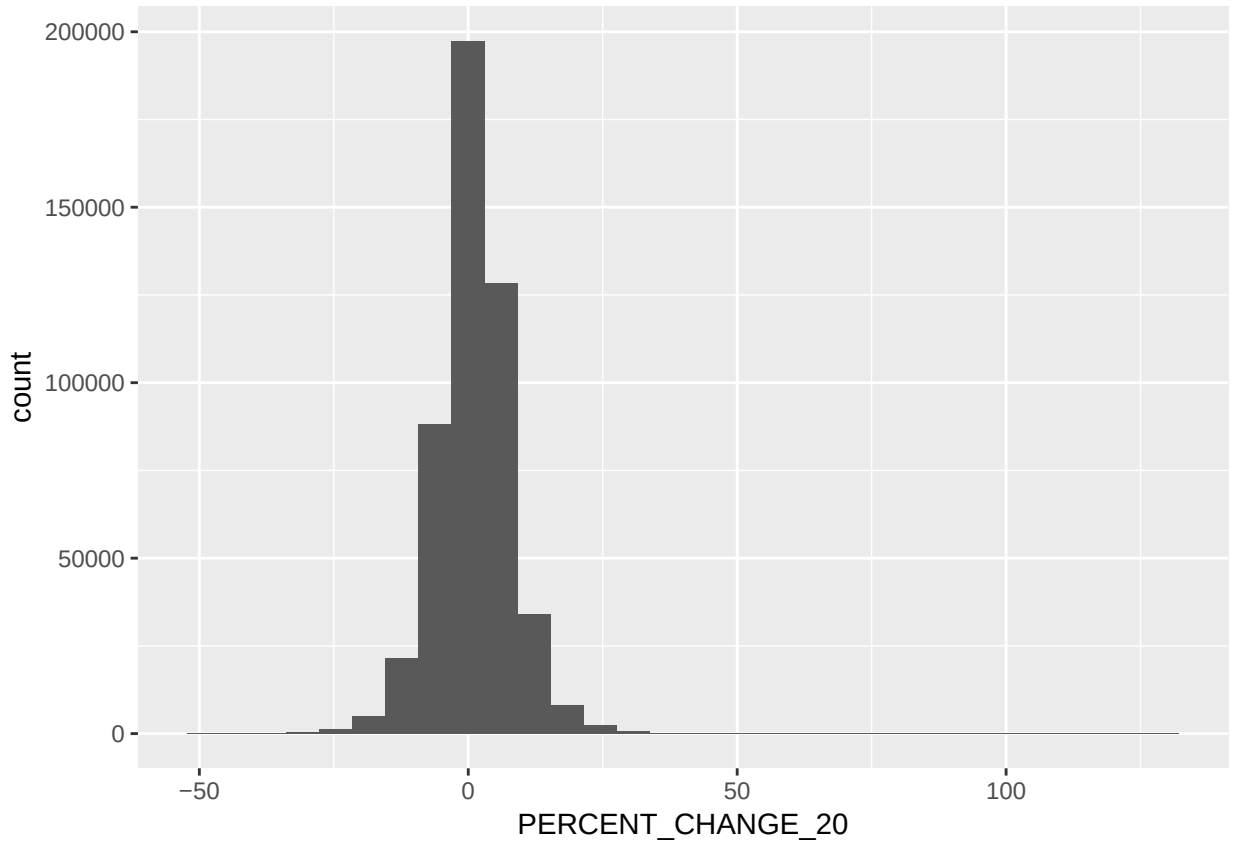



```
ggplot(master_dataset) + geom_bar(aes(x=FUTURE_CLASS_240))
```



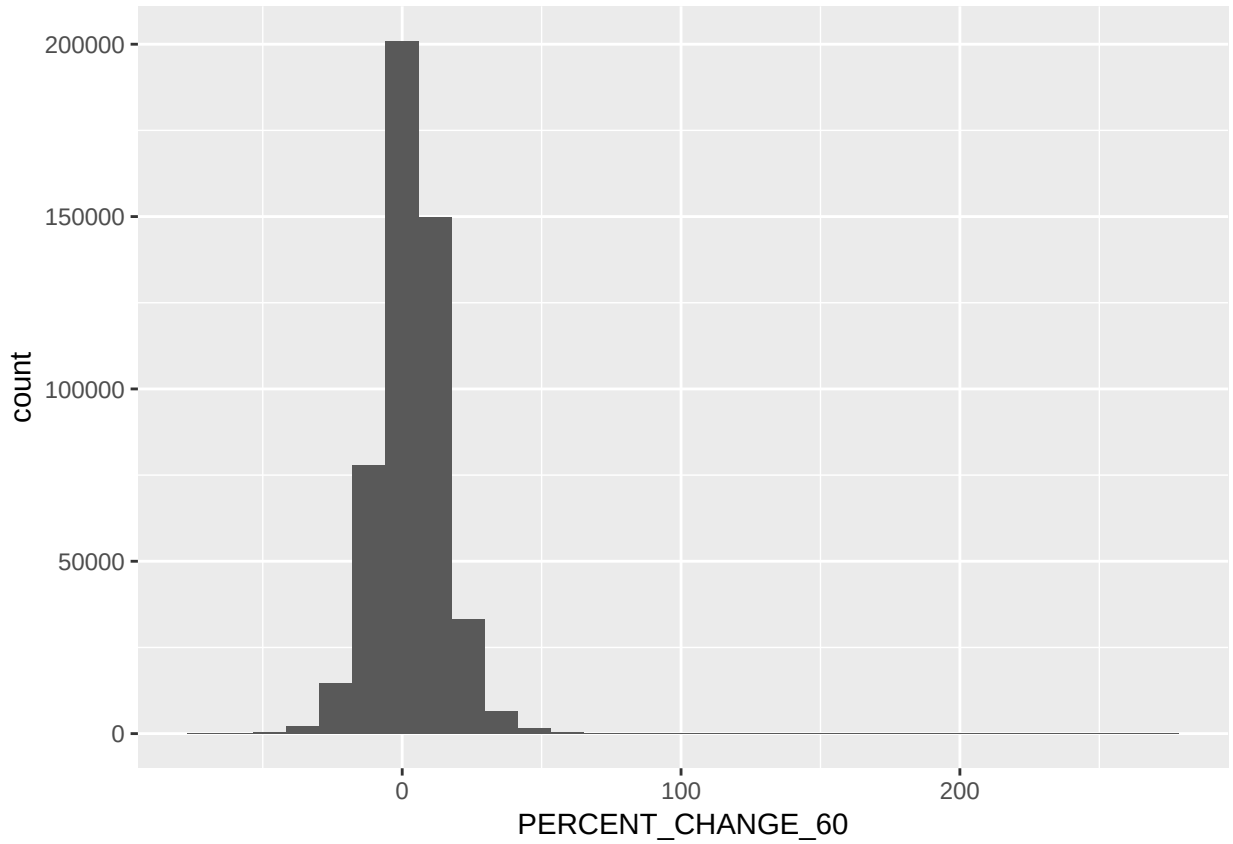

```
# distribution of percent_change response variables
# tends to normal distribution, cannot go below -100% change
ggplot(master_dataset) + geom_histogram(aes(x=PERCENT_CHANGE_20))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



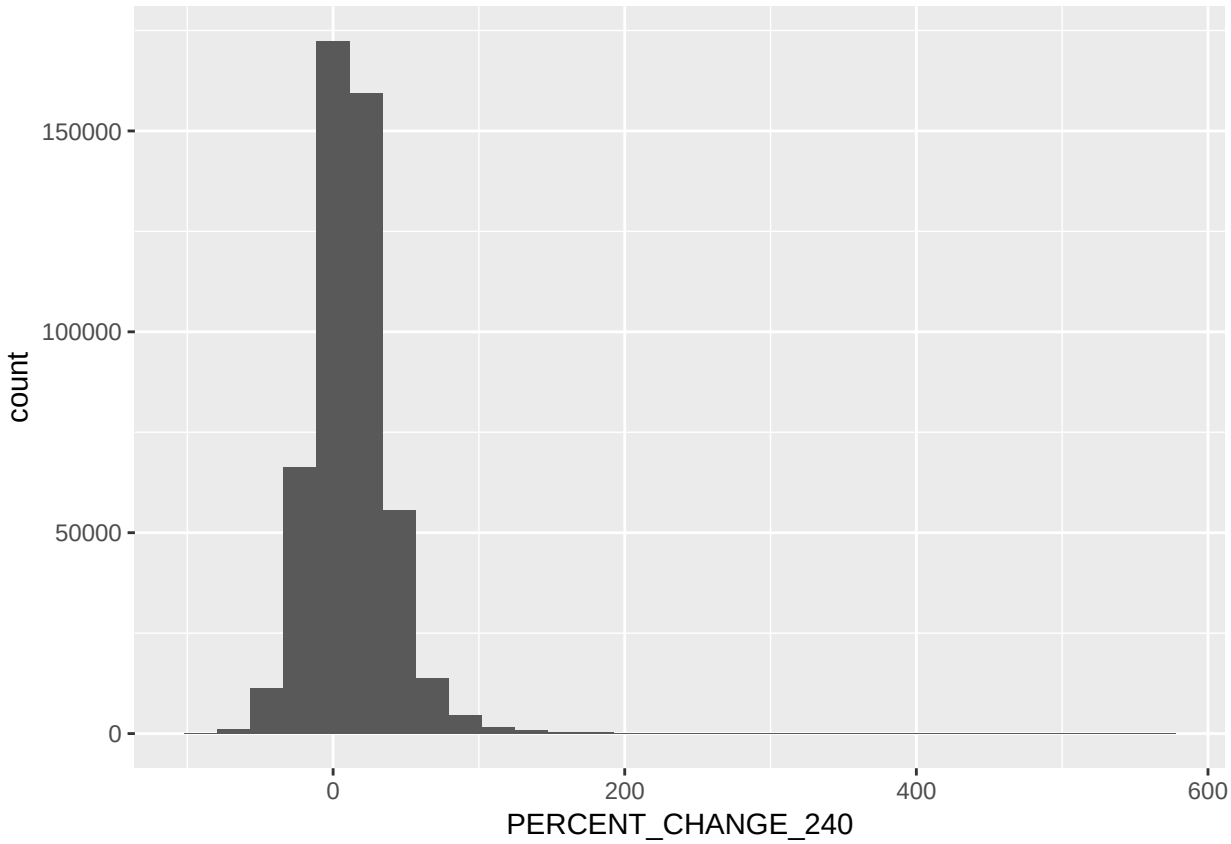

```
ggplot(master_dataset) + geom_histogram(aes(x=PERCENT_CHANGE_60))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(master_dataset) + geom_histogram(aes(x=PERCENT_CHANGE_240))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


150000
100000
50000
0
count
0
200
400
600
PERCENT_CHANGE_240