Using Data to Make Investment Decisions

Final Report

ST362-B   Spring 2022

George Mandl

200655330

mand5330@mylaurier.ca

Introduction

The data set that I will be using for analysis is one that I have created myself using Microsoft Excel and Python in the Juypter Notebook environment. The data set is called "StockPricePredictionDataSet.csv" and it will be referred to as "the master data" or "the data set" in this report. The data set has 25 columns and 162,115 rows. It contains data representing all 500 companies in the Standard and Poor's index: the S&P500. The S&P 500 is very well known and tracks large companies whose equities are traded in high quantity which allows their trends to be clear and easily calculated.

The data set contain some standard columns that accompany investment data such as Symbol (the company's ticker symbol on the stock market), Date, High (the highest price of the day), Low (the lowest price of the day), Open (the price of the equity at market open), Close (the price of the equity at market close), and Volume (the quantity of shares traded in the day). For the remainder of the columns, I have implemented various technical indicators in Python. Investment analysis can be broken into two categories: fundamental analysis and technical analysis. Fundamental analysis is concerned with the intrinsic value of stocks and technical analysis is concerned with the future price of the stock and predicts it using charts or numerical calculations. Technical indicators can identify whether a stock is overvalued and will drop in price or undervalued and will rise in price. Technical indicators are based on arithmetic and statistics in correspondence with the equity's current and past prices. Fifteen of the columns in the master data are indicators and are to be used as predictor variables to predict the future price of an equity (See appendix for an explanation of each technical indicator used). There are two columns that are transformed from other columns to be used as response variables: FuturePrice and % Change (X..Change). The distance in the future for these features is 10 trading days which is approximately 14 days since Saturdays and Sundays are not counted as trading days because stock markets are not open. Being able to accurately predict the future price of an equity is highly significant because you are then able to compare it to the current price and make investment decisions based on the comparison. Each row of the data set represents the daily information for a given stock.

I have already built a trading algorithm that uses the technical indicators which I implemented into Python code, predicts the future price of the equity, and compares the future prices to report back which stocks I should invest in. The purpose of this project is to evaluate and improve the linear regression model that I currently use to predict the stock prices. I hope that this results in not only an influx of profits but also a mitigating of risk.

Some questions that can be used to address the correctness of the algorithm's regression model and improve its accuracy are:

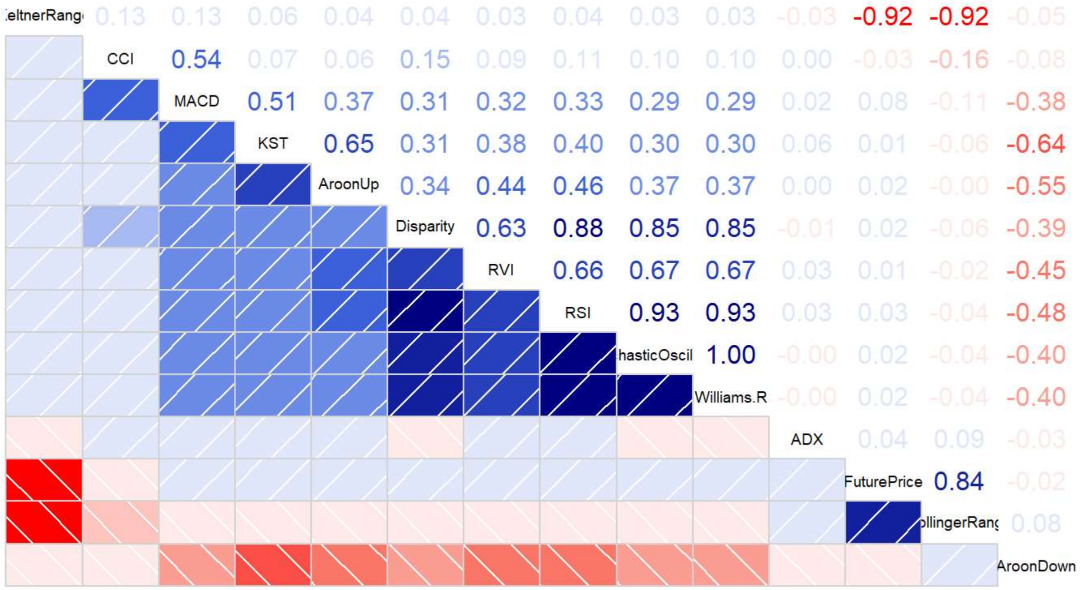
- How strong is the multicollinearity between technical indicators?
- Should technical indicators with high multicollinearity be removed from the model?
- Which of the response variable results in the more accurate model?
- Should the regression be done across the entire data set or automated for individual equities?
- Does standardizing the predictor variables improve the model's accuracy?

How strong is the multicollinearity between technical indicators?

Multicollinearity occurs when using a multiple linear regression model and is the idea that the predictor variables of the model have some correlation. No perfect multicollinearity is one of the assumptions of the OLS multiple linear regression model. If this assumption is violated, the model cannot be estimated accurately. For example, imagine the model is Y = b0 + b1X1 + b2X2 and X1 and X2 have perfect multicollinearity. Then b1 cannot be interpreted as the change in Y resulting from a one unit change in X1 holding all other factors constant because you cannot hold all other factors constant when X1 and X2 are multicollinear. A one unit change in X1 directly results in a one unit change in X2.

Since all the technical indicators are calculated using the equity's close, high, low, and open variables it is highly likely that some indicators have a high correlation. The goal of the following tests is to find which variables display high multicollinearity. These variables will be removed to produce a less biased and more accurate model.

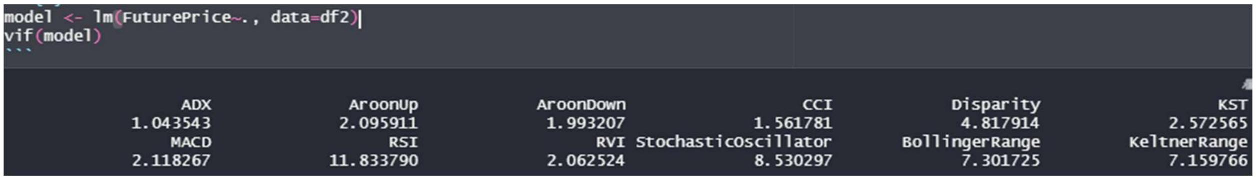
The elementary analysis for discovering multicollinearity is creating a correlation matrix between the response variables. The goal of the analysis is to determine which variables have perfect or high multicollinearity and potentially eliminate them. Perfect multicollinearity in this context is a pair of indicators whose correlation is either -1 or 1. The threshold for high multicollinearity will be 0.8, any pair of indicators whose correlation is greater than |0.8| are considered highly multicollinear. The following image shows the correlation matrix of regression parameters.



Correlation Matrix of Regression Parameters

The correlation matrix shows one perfect multicollinear relationship which is between the Williams % R indicator and the Stochastic Oscillator indicator. We must remove one of these indicators from our predictor variable set otherwise our multiple linear regression model will not function accurately. Going forward, I have chosen to remove the Williams % R indicator. There is also high multicollinearity amongst several indicators' pairs according to the correlation matrix. Stochastic Oscillator, RSI, and Disparity are pairwise highly multicollinear and the Bollinger Band Range is highly correlated to the Keltner Channel Range. None of these columns will be dropped from regression model immediately since the use of the correlation matrix is fairly elementary but they will be under careful consideration during the next tests for multicollinearity.

The other analysis we will do for multicollinearity is analyzing the variance inflation factors (VIF) of the indicators. VIF regresses each predictor variable on the other predictor variables and measures how much of the variance of the regressed predictor is observed from the other predictors. Higher values of VIF indicate multicollinearity meaning a lot of the variance is detected from other predictors. There is no universal cut-off for multicollinearity indication but values of 5 and 10 are often used. For this project, a cut-off of 5 will be used. After the removal of the Williams % R indicator, the following are the VIF values of the remaining indicator variables.



```
model <- lm(FuturePrice~., data=df2)
vif(model)
```

```
           ADX            AroonUp          AroonDown                CCI          Disparity                KST
      1.043543           2.095911           1.993207           1.561781           4.817914           2.572565
          MACD                RSI                RVI StochasticOscillator     BollingerRange        KeltnerRange
      2.118267          11.833790           2.062524           8.530297           7.301725           7.159766
```

Remember from the correlation matrix analysis that the Disparity, RSI, Stochastic Oscillator, Bollinger Band Range, and Keltner Channel Range variables were involved in highly multicollinear relationship. Based on the VIF values the same variables are showing multicollinearity.

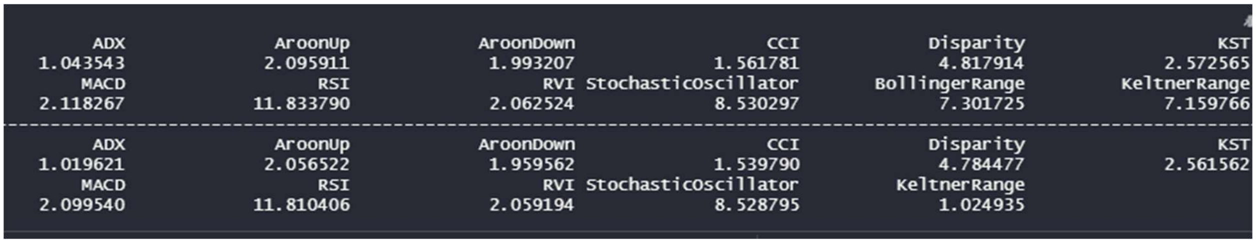
Both the variance inflation factors, and the correlation matrix indicate that there is a high correlation between the Bollinger Band Range and Keltner Channel Range indicators, and the Disparity, RSI, and Stochastic Oscillator indicators. Should the Bollinger Band Range indicator be removed from the regression parameters? What is effect of removing the Disparity and Stochastic Oscillator indicators from the regression parameters? These are the next questions that will be answered by this report.

Should the Bollinger Band Range indicator be removed from the regression parameters?

From the previous question it was discovered that there is high multicollinearity between the Bollinger Band Range indicator and Keltner Channel Range indicator. We will analyze the result of removing the

Bollinger Band Range indicator from the model through the VIF values to determine the effect on multicollinearity and through Extra Sum of Squares to evaluate the impact on model performance.

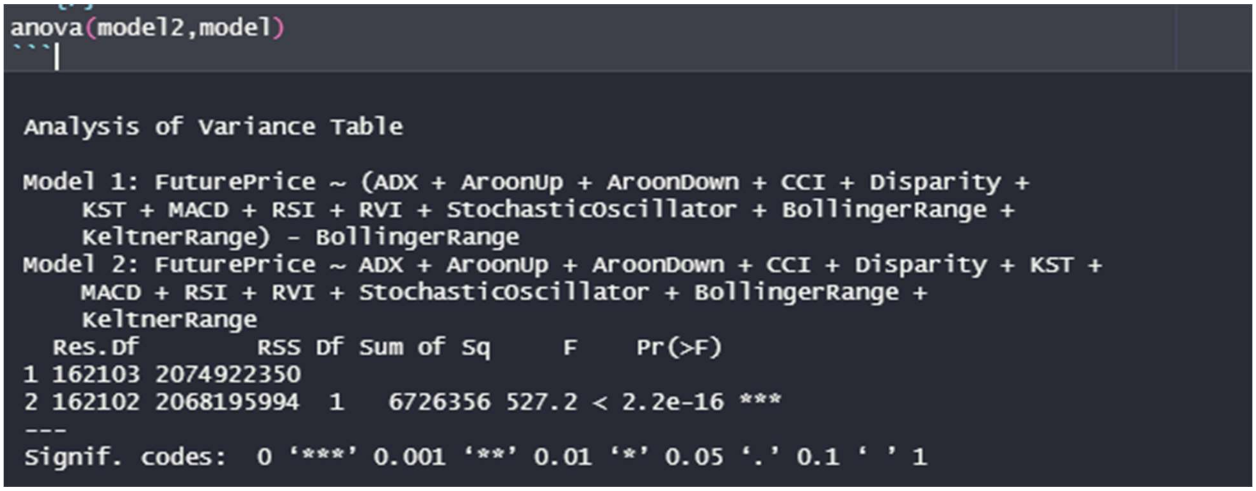
First, multicollinearity will be analyzed by observing the change in VIF values when the Bollinger Band Range indicator is removed from the model.



```
                 ADX              AroonUp            AroonDown                  CCI            Disparity                  KST
            1.043543             2.095911             1.993207             1.561781             4.817914             2.572565
                MACD                  RSI                  RVI StochasticOscillator       BollingerRange         KeltnerRange
            2.118267            11.833790             2.062524             8.530297             7.301725             7.159766
--------------------------------------------------------------------------------------------------------------------------------
                 ADX              AroonUp            AroonDown                  CCI            Disparity                  KST
            1.019621             2.056522             1.959562             1.539790             4.784477             2.561562
                MACD                  RSI                  RVI StochasticOscillator         KeltnerRange
            2.099540            11.810406             2.059194             8.528795             1.024935
```

The VIF values above the line are the same values from the previous question, when the Bollinger Band Range variable was still included in the model. Below the line are the VIF values after the Bollinger Band Range variable was removed from the model. Notice two things. First, the VIF of the Keltner Channel Range indicator plummeted and the variable is no longer multicollinear. Secondly, magnitude aside, all the variable's VIF values decreased. The conclusion is that the removal of the Bollinger Band Range indicator reduces the multicollinearity in the model significantly.

Next, we will the significance of removing the Bollinger Band Range indicator on the model's performance using Extra Sum of Squares (ESS). ESS is used when one model is nested as a subset of another. ESS compares the models with a partial F-test which determines statistical significance of the variables that differentiate the two models. In our case we have the model without the Bollinger Band Range variable nested within the model that contains the Bollinger Band Range variable. The null hypothesis of our partial F-test will be that the Bollinger Band Range variable is linearly insignificant to the regression model.



```
anova(model2,model)
```

```
Analysis of Variance Table

Model 1: FuturePrice ~ (ADX + AroonUp + AroonDown + CCI + Disparity + 
    KST + MACD + RSI + RVI + StochasticOscillator + BollingerRange + 
    KeltnerRange) - BollingerRange
Model 2: FuturePrice ~ ADX + AroonUp + AroonDown + CCI + Disparity + KST + 
    MACD + RSI + RVI + StochasticOscillator + BollingerRange + 
    KeltnerRange
  Res.Df        RSS Df Sum of Sq     F    Pr(>F)    
1 162103 2074922350                                 
2 162102 2068195994  1   6726356 527.2 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

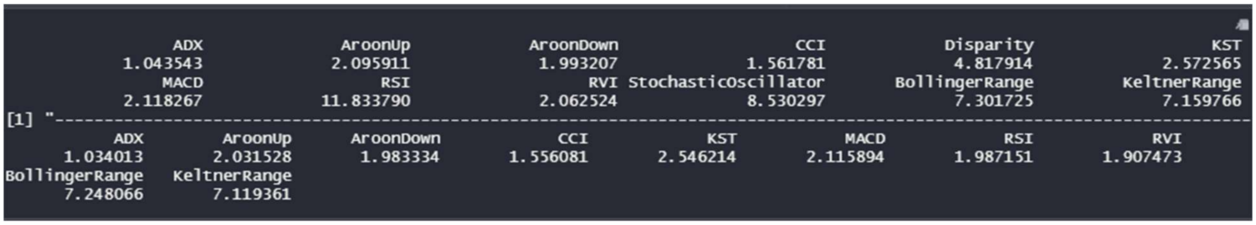
The F-value from the partial F-test is 527.2 and the p-value is extremely low so there is no statistically significant evidence for removing the Bollinger Band Range variable.

Although removing the Bollinger Band Range indicator from the regression parameters reduces the multicollinearity, there is no statistically significant evidence that supports the removal of the indicator. Therefore, the Bollinger Band Range indicator should not be removed from the regression parameters.

Should the Disparity and Stochastic Oscillator indicators be removed from the regression parameters?

From the earlier question regarding multicollinearity, we observed that the RSI, Disparity and Stochastic Oscillator indicators were pairwise highly correlated. The result of removing the Disparity and Stochastic Oscillator indicators from the model will again analyzed through the VIF values to determine the effect on multicollinearity and through Extra Sum of Squares to evaluate the impact on model performance.

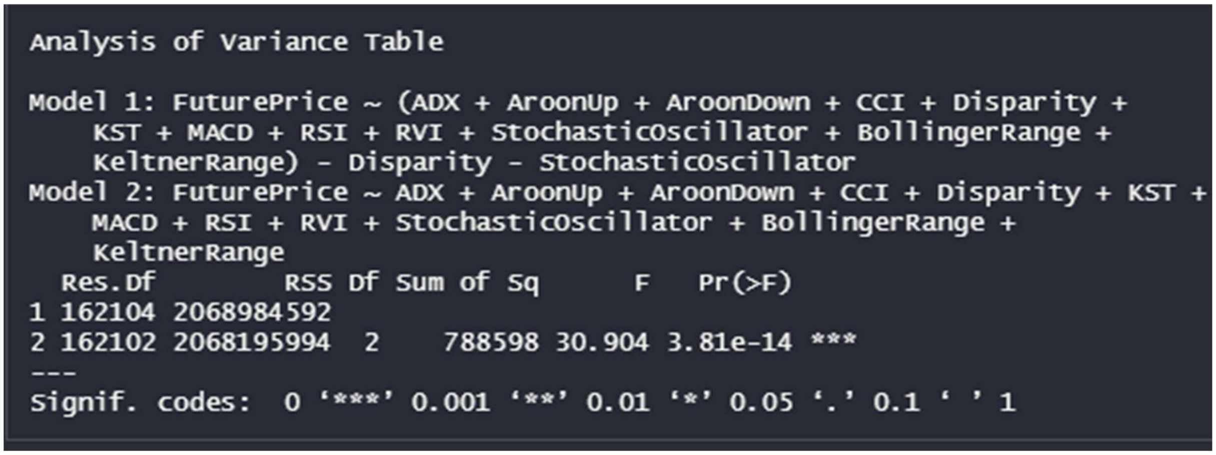
First, multicollinearity will be analyzed by observing the change in VIF values when the Stochastic Oscillator and Disparity indicators are removed from the model.



```
                 ADX              AroonUp            AroonDown                  CCI            Disparity                  KST 
            1.043543             2.095911             1.993207             1.561781             4.817914             2.572565 
                MACD                  RSI                  RVI StochasticOscillator       BollingerRange         KeltnerRange 
            2.118267            11.833790             2.062524             8.530297             7.301725             7.159766 
[1] "-----------------------------------------------------------------------------------------------------------------------
           ADX        AroonUp      AroonDown            CCI            KST           MACD            RSI            RVI 
      1.034013       2.031528       1.983334       1.556081       2.546214       2.115894       1.987151       1.907473 
BollingerRange   KeltnerRange 
      7.248066       7.119361 
```

The VIF values above the line are the same values from the previous question when the Disparity and Stochastic Oscillator variables were still included in the model. Below the line are the VIF values after the Disparity and Stochastic Oscillator variables were removed from the model. Notice two things. First, the VIF of the RSI indicator plummeted and the variable is no longer multicollinear. Secondly, magnitude aside, all the variable's VIF values decreased. The conclusion is that the removal of the Disparity and Stochastic Oscillator indicators reduces the multicollinearity in the model significantly.

Next, we will the significance of removing the indicators on the model's performance using Extra Sum of Squares (ESS). ESS is relevant in this case also since we have the model without the Stochastic Oscillator and Disparity variables nested within the model that contains the Stochastic Oscillator and Disparity variables. The null hypothesis of our partial F-test will be that the Stochastic Oscillator and Disparity variables are linearly insignificant to the regression model.



```
Analysis of Variance Table

Model 1: FuturePrice ~ (ADX + AroonUp + AroonDown + CCI + Disparity +
    KST + MACD + RSI + RVI + StochasticOscillator + BollingerRange +
    KeltnerRange) - Disparity - StochasticOscillator
Model 2: FuturePrice ~ ADX + AroonUp + AroonDown + CCI + Disparity + KST +
    MACD + RSI + RVI + StochasticOscillator + BollingerRange +
    KeltnerRange
  Res.Df        RSS Df Sum of Sq      F   Pr(>F)
1 162104 2068984592
2 162102 2068195994  2    788598 30.904 3.81e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

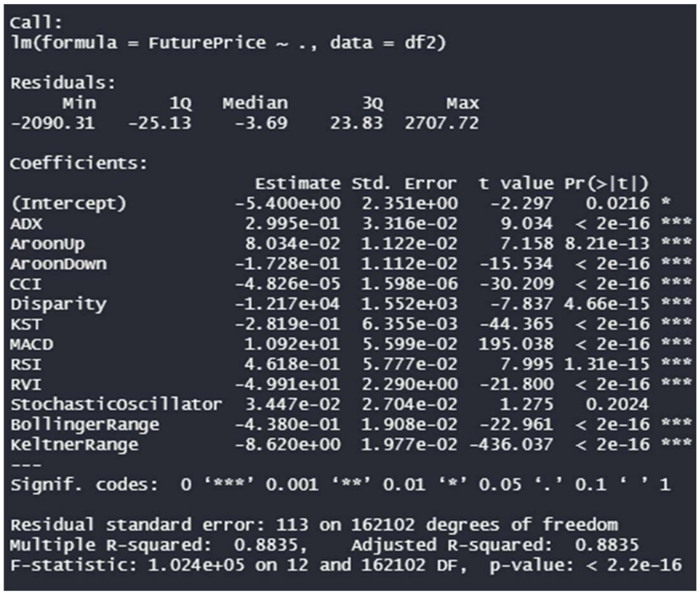
The F-value from the partial F-test is 30.904 and the p-value is extremely low so there is no statistically significant evidence for removing the Stochastic Oscillator and Disparity variables.

Although removing the Stochastic Oscillator and Disparity indicators from the regression parameters reduces the multicollinearity, there is no statistically significant evidence that supports the removal of the indicators. Therefore, the Stochastic Oscillator and Disparity indicators should not be removed from the regression parameters.

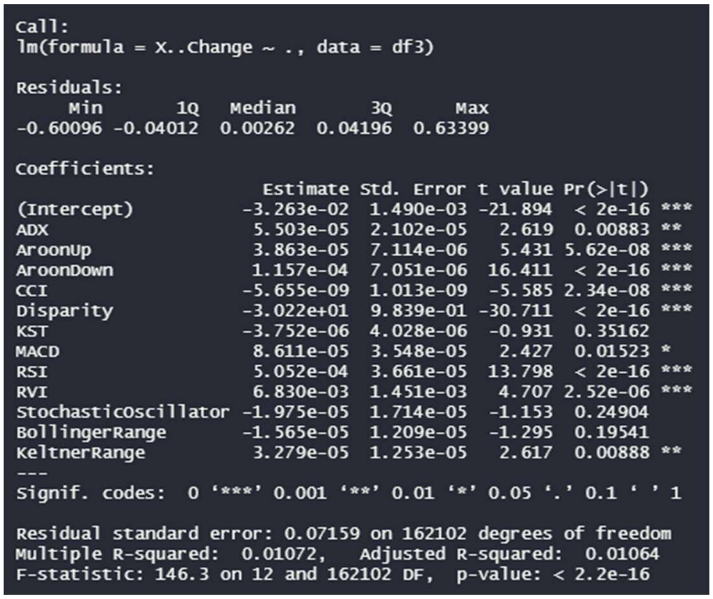
Which response variable results in the most accurate model?

Now that we have cleaned our predictor variables, we will determine which response variable to use for the rest of our data set analysis. The data set contains two possible response variables. The first is FuturePrice which is the Close column shifted backwards 2 weeks. The second is X..change which is the % Change in Future Price. This variable is calculated by subtracting the current price (Close) from the future price (FuturePrice) and dividing by the current price.

I will assess which response variable is more accurate by regressing both response variables to our predictor variables. Then I will conduct an F-test on both models. The resulting F-statistic will determine which response variable is more accurate and will be used in all future models. The reason the F-statistic is being used to compare and not the residual sum of squares or the mean sum of squares is because the variables have vastly different units of measurement. FuturePrice is measured in dollars and ranges from 10 to 5000. The % Change in Future Price is measured as a percentage and is ranges from approximately –1 to 1. Additionally, the R^2 is an appropriate determinant since it is unitless and it measures the amount of variance in the response that is observed by the predictors.



```
Call:
lm(formula = FuturePrice ~ ., data = df2)

Residuals:
     Min       1Q   Median       3Q      Max 
-2090.31   -25.13    -3.69    23.83  2707.72 

Coefficients:
                       Estimate Std. Error  t value Pr(>|t|)    
(Intercept)          -5.400e+00  2.351e+00   -2.297   0.0216 *  
ADX                   2.995e-01  3.316e-02    9.034  < 2e-16 ***
AroonUp               8.034e-02  1.122e-02    7.158 8.21e-13 ***
AroonDown            -1.728e-01  1.112e-02  -15.534  < 2e-16 ***
CCI                  -4.826e-05  1.598e-06  -30.209  < 2e-16 ***
Disparity            -1.217e+04  1.552e+03   -7.837 4.66e-15 ***
KST                  -2.819e-01  6.355e-03  -44.365  < 2e-16 ***
MACD                  1.092e+01  5.599e-02  195.038  < 2e-16 ***
RSI                   4.618e-01  5.777e-02    7.995 1.31e-15 ***
RVI                  -4.991e+01  2.290e+00  -21.800  < 2e-16 ***
StochasticOscillator  3.447e-02  2.704e-02    1.275   0.2024    
BollingerRange       -4.380e-01  1.908e-02  -22.961  < 2e-16 ***
KeltnerRange         -8.620e+00  1.977e-02 -436.037  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 113 on 162102 degrees of freedom
Multiple R-squared:  0.8835,	Adjusted R-squared:  0.8835 
F-statistic: 1.024e+05 on 12 and 162102 DF,  p-value: < 2.2e-16
```



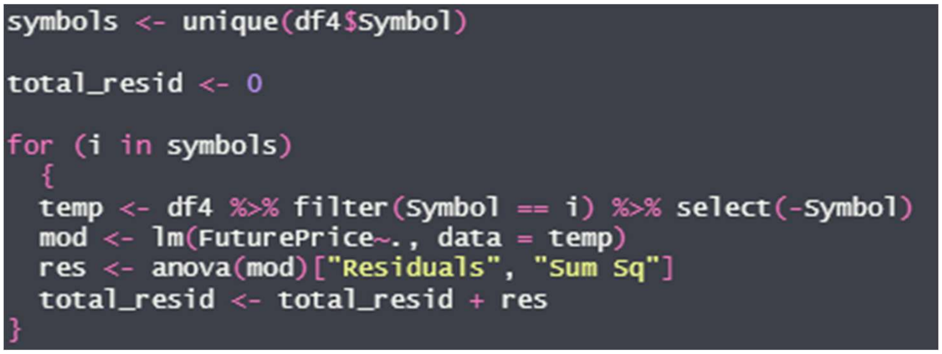
```
Call:
lm(formula = X..Change ~ ., data = df3)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.60096 -0.04012  0.00262  0.04196  0.63399 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)          -3.263e-02  1.490e-03 -21.894  < 2e-16 ***
ADX                   5.503e-05  2.102e-05   2.619  0.00883 ** 
AroonUp               3.863e-05  7.114e-06   5.431 5.62e-08 ***
AroonDown             1.157e-04  7.051e-06  16.411  < 2e-16 ***
CCI                  -5.655e-09  1.013e-09  -5.585 2.34e-08 ***
Disparity            -3.022e+01  9.839e-01 -30.711  < 2e-16 ***
KST                  -3.752e-06  4.028e-06  -0.931  0.35162    
MACD                  8.611e-05  3.548e-05   2.427  0.01523 *  
RSI                   5.052e-04  3.661e-05  13.798  < 2e-16 ***
RVI                   6.830e-03  1.451e-03   4.707 2.52e-06 ***
StochasticOscillator -1.975e-05  1.714e-05  -1.153  0.24904    
BollingerRange       -1.565e-05  1.209e-05  -1.295  0.19541    
KeltnerRange          3.279e-05  1.253e-05   2.617  0.00888 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07159 on 162102 degrees of freedom
Multiple R-squared:  0.01072,	Adjusted R-squared:  0.01064 
F-statistic: 146.3 on 12 and 162102 DF,  p-value: < 2.2e-16
```

The left image is the summary of the regression model using Future Price as the response variable and the right image is the summary of the regression model using % Change in Future Price as the response variable. Observe the two results of interest: F-statistic and R² at the bottom of both images. The F-statistic of the future price is slightly higher but the R² value is what pops out. The R² value of the future price model is significantly higher than the R² value of the % Change model. Due to this the FuturePrice variable will used for all the remaining models and to answer the question: Future Price is the response variable resulting in the more accurate model.

Is it more accurate to predict future prices on an entire dataset or individually along each equity?

The stocks that are included in the data set are all quite different. Their companies are all in different sectors, markets, and price ranges. The price range is what is of most significance in our case. This entire time we have been running our regression models across an entire data set of different stocks. Some of those stocks trade in the range of $25-$50 whereas some stocks trade in the range of $3,000-$5,000. One would assume that this causes a problem in the estimations of the coefficients of the prediction formula in the regression model. The coefficients that maybe provide optimal weights to some stocks will be very inaccurate for other stocks. One would assume that the predictions for the higher priced stocks will be underestimated and the predictions for the lower priced stocks will be overestimated.

There is interest in seeing if it is more accurate to create a regression model on an entire dataset of stocks and predict the future stock prices using this model. The alternative would be to automate the creation of regression models and prediction of future prices for individual stocks. The latter, is exactly what I have done through the following code:



```
symbols <- unique(df4$Symbol)

total_resid <- 0

for (i in symbols)
  {
  temp <- df4 %>% filter(Symbol == i) %>% select(-Symbol)
  mod <- lm(FuturePrice~., data = temp)
  res <- anova(mod)["Residuals", "Sum Sq"]
  total_resid <- total_resid + res
}
```

The code first creates a list of all the unique stocks in the data set. Then creates a model for each individual stock with data that is filtered to only contain the stock of interest.

In the code, I incorporated the summation of the residual sum of squares of across all the individual stocks. I will compare this total residual variable to the residual sum of squares of the regression on the entire data set to determine which is more accurate for use in my predictive algorithm. This is possible because the formula for the linear models across the stocks is the same formula that is used across the data set. The linear models both have the same response and predictors except one is run just once on a large data set and the other is ran frequently on small datasets.

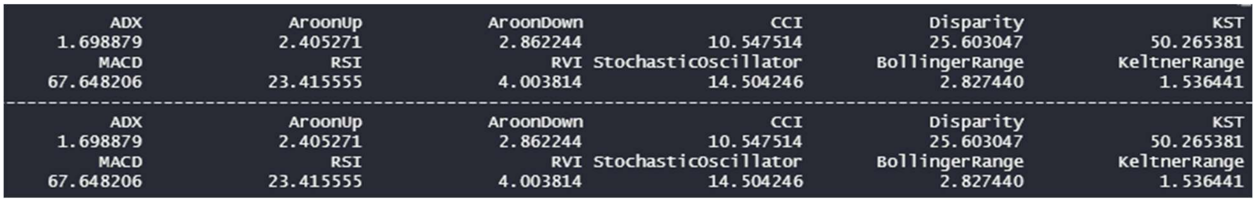
The model that ran across the entire data set resulted in a residual sum of squares of 2,068,984,592 and the model that ran across each stock individually and then summed all the residuals had a residual sum of squares of 137,585,492. Therefore, running the regression across each stock is significantly more accurate and it is worth automating the prediction of future prices to run on individual equities. For future regression analysis, models that contain only one individual stock will be used.

Does standardizing the variables increase model accuracy?

Standardizing a variable means to subtract it from its mean and divide it by its standard deviation. The resulting variable then has a mean of 0 and a standard deviation of 1. This scales the variables and allows them to be easily compared and combined with other variables. Scaling can be very effective in reducing model error and increasing model accuracy. On the other hand, the technical indicators become scaled directly in their calculations. There is interest to see if the standardization of the technical indicators impacts the model.

The model that will be used for this analysis will contain the stock 3M (Ticker Symbol MMM) only. First, a model will be created using the regular, unstandardized variables. Then, the variables will be standardized, and a model will be created and compared to the first.

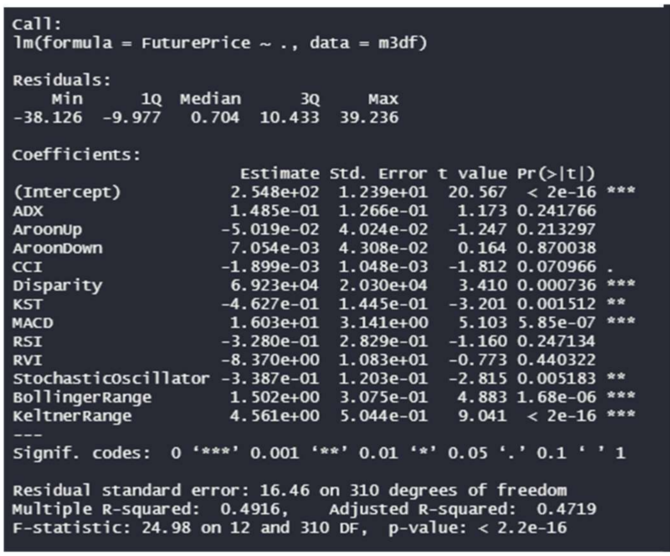
Standardizing reduces multicollinearity so the first analysis will be done using VIF. The VIF values will first be calculated using the unstandardized data and then using the standardized data.



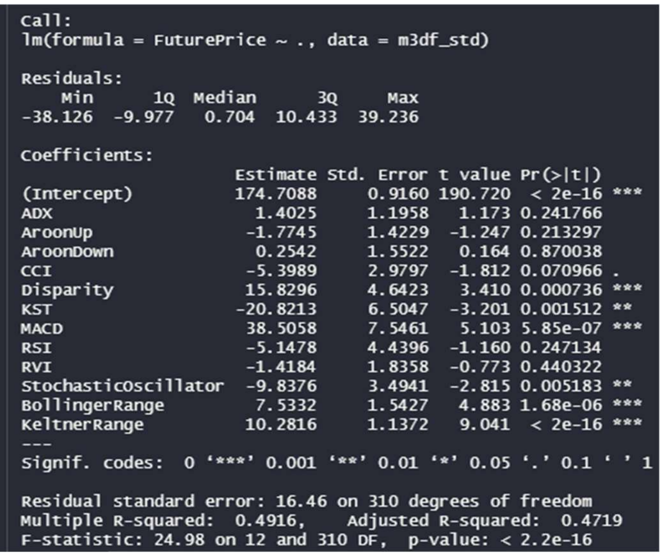
```
                ADX             AroonUp           AroonDown                 CCI           Disparity                 KST 
           1.698879            2.405271            2.862244           10.547514           25.603047           50.265381 
               MACD                 RSI                 RVI StochasticOscillator      BollingerRange         KeltnerRange 
          67.648206           23.415555            4.003814           14.504246            2.827440            1.536441 
---------------------------------------------------------------------------------------------------------------------------
                ADX             AroonUp           AroonDown                 CCI           Disparity                 KST 
           1.698879            2.405271            2.862244           10.547514           25.603047           50.265381 
               MACD                 RSI                 RVI StochasticOscillator      BollingerRange         KeltnerRange 
          67.648206           23.415555            4.003814           14.504246            2.827440            1.536441 
```

Notice that the VIF results for both models are identical. This means that standardizing the variables in our data has no effect on the multicollinearity, for better nor for worse.

Next will be the analysis of standardization on the model's performance. First a model will be fitted and analyzed on the unstandardized data and then the same will be done on the standardized data.



```
Call:
lm(formula = FuturePrice ~ ., data = m3df)

Residuals:
    Min      1Q  Median      3Q     Max 
-38.126  -9.977   0.704  10.433  39.236 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)           2.548e+02  1.239e+01  20.567  < 2e-16 ***
ADX                   1.485e-01  1.266e-01   1.173 0.241766    
AroonUp              -5.019e-02  4.024e-02  -1.247 0.213297    
AroonDown             7.054e-03  4.308e-02   0.164 0.870038    
CCI                  -1.899e-03  1.048e-03  -1.812 0.070966 .  
Disparity             6.923e+04  2.030e+04   3.410 0.000736 ***
KST                  -4.627e-01  1.445e-01  -3.201 0.001512 ** 
MACD                  1.603e+01  3.141e+00   5.103 5.85e-07 ***
RSI                  -3.280e-01  2.829e-01  -1.160 0.247134    
RVI                  -8.370e+00  1.083e+01  -0.773 0.440322    
StochasticOscillator -3.387e-01  1.203e-01  -2.815 0.005183 ** 
BollingerRange        1.502e+00  3.075e-01   4.883 1.68e-06 ***
KeltnerRange          4.561e+00  5.044e-01   9.041  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.46 on 310 degrees of freedom
Multiple R-squared:  0.4916,	Adjusted R-squared:  0.4719 
F-statistic: 24.98 on 12 and 310 DF,  p-value: < 2.2e-16
```



```
Call:
lm(formula = FuturePrice ~ ., data = m3df_std)

Residuals:
    Min      1Q  Median      3Q     Max 
-38.126  -9.977   0.704  10.433  39.236 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)          174.7088     0.9160 190.720  < 2e-16 ***
ADX                    1.4025     1.1958   1.173 0.241766    
AroonUp               -1.7745     1.4229  -1.247 0.213297    
AroonDown              0.2542     1.5522   0.164 0.870038    
CCI                   -5.3989     2.9797  -1.812 0.070966 .  
Disparity             15.8296     4.6423   3.410 0.000736 ***
KST                  -20.8213     6.5047  -3.201 0.001512 ** 
MACD                  38.5058     7.5461   5.103 5.85e-07 ***
RSI                   -5.1478     4.4396  -1.160 0.247134    
RVI                   -1.4184     1.8358  -0.773 0.440322    
StochasticOscillator  -9.8376     3.4941  -2.815 0.005183 ** 
BollingerRange         7.5332     1.5427   4.883 1.68e-06 ***
KeltnerRange          10.2816     1.1372   9.041  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.46 on 310 degrees of freedom
Multiple R-squared:  0.4916,	Adjusted R-squared:  0.4719 
F-statistic: 24.98 on 12 and 310 DF,  p-value: < 2.2e-16
```

The image on the left shows the results of the unstandardized model and the image on the right shows the results of the standardized model. The individual variables performance changes since they are transformed in the standardization, but the underlying model results seen at the bottom of both images are the same.

These results confirm my hypothesis that adequate scaling is done in the calculation of the indicators and standardization provides no benefit to the model. To answer the question: standardization of the predictor variables does not increase model accuracy.

Conclusion

The general conclusion to be made from this project is that a regression model is effective at predicting future prices of the stock market. My hypothesis that fitting models to stocks individually rather than the entire data set was supported by the results from this research and I can now go forward with implementing the automation of regression models within my algorithm. The statistical analysis of the algorithm has also produced useful insights that will allow the algorithm to be more accurate and effective. Parameter optimization can be conducted through support of response variable and predictor variable analysis. The exploration of multicollinearity allows for the reduction of bias in the model through elimination of counterproductive variables. Overall, the project has been eye-opening on the possibilities and use of regression in real-world topics, and it is exciting to see the applications of the course's contents in the field.

Appendix

General Financial Terminology

Equity

An equity is the stake which shareholders of a company hold. Equity's which are listed to the public can be traded one exchanges such as the Toronto Stock Exchange (TSX) and the NASDAQ stock market. For the purpose of this report the term equity is synonymous with the terms stock, share, security, and asset.

Short

On a high level, a short is a way for an investor to make money betting on the stock falling. It can be thought of as first selling an equity and ten buying it back at a future date. If you think the price will fall then the price on the future date will be lower than what you shorted it for, allowing the investor to profit while the value of the equity falls.

Overbought

A condition that indicates that the equity is trading at a higher price than it is worth. If an equity is overbought, then it is a signal to sell or short it and can indicate a potential price pullback.

Oversold

A condition that indicates that the equity is trading at a lower price than it is worth. If an equity is oversold, then it is a buy signal and can indicate a potential price bounce.

Overvalued

A condition that indicates that the equity is trading at a higher price than it is worth. If an equity is overvalued, then it is a signal to sell or short it and can indicate a potential price pullback.

Undervalued

A condition that indicates that the equity is trading at a lower price than it is worth. If an equity is undervalued, then it is a buy signal because you are purchasing it for a bargain. An undervalued condition could also indicate a potential price bounce.

Bullish

A condition that indicates that the equity is trading at a lower price than it is worth. A bullish signal would lead an investor to believe that the equity will go higher and indicate that they should purchase the equity.

Bearish

A condition that indicates that the equity is trading at a higher price than it is worth. A bearish signal would lead an investor to believe that the equity will go lower and indicate that they should sell or short the equity.

Note: For the purpose of this project the terms *Bullish, Undervalued,* and *Oversold* are synonymous and the terms *Bearish, Overvalued,* and *Overbought* are synonymous.

Moving Average (MA)

A moving average is a calculation done over a time series and is used in investing to smooth out the price of an equity. It updates the average price while it moves along the time series, hence the name moving average. A rising moving average indicates that an equity is in an uptrend and a declining moving average indicates than an equity is a downtrend.

Simple Moving Average (SMA)

The simplest way of calculating a moving average is to take the arithmetic mean of the stock prices over the lookback period.

$$\text{SMA} = \text{Price } 1 + \text{Price } 2 + \ldots + \text{Price } n$$

Price is the price from i days ago from i equals 0 to n-1 n is the length of the lookback period

Exponential Moving Average (EMA)

The EMA gives more weight to recent prices. It multiplies the previous day's EMA by its weight and adds it to the price of the equity that day.

EMA = Price * Smoothing Factor1 + Lookback Length + Previous Day EMA * [1 - Smoothing Factor1 + Lookback Length]

Typical Price

The typical price takes the average of an equity's closing price, daily high, and daily low to appropriately represent the equity's price.

Typical Price = Close + High + Low3

Oscillator

An oscillator is a tool in technical analysis of equities used to measure overbought and oversold conditions. It is set up to be bounded between two extreme values and fluctuates between the bounds. As the oscillator trends towards the upper extreme, it indicates that the equity is overbought and as it trends towards the lower extreme, it is indicating that the equity is oversold.

Technical Indicators

Average Directional Index (ADX) - Trend Indicator

The ADX is a technical indicator that measures the strength of a trend using moving averages of the directional movements usually over 14 days. It is a popular indicator because investors reduce risk and increase profit potential by trading the direction of a strong trend. The values of ADX range from 0 to 100.

0 - 25	Absent/Weak Trend
25 - 50	Strong Trend

50 - 75	Very Strong Trend
75 - 100	Extremely Strong Trend

Relative Strength Index (RSI) - Momentum Indicator

RSI is a momentum indicator that measures the magnitude of an equity's recent price change to evaluate oversold or overbought conditions. RSI is an oscillator that fluctuates between 0 and 100. An RSI value over 70 indicates an overbought condition and can indicate that the equity is primed for a pullback. An RSI under 30 indicates an oversold condition.

RSI is calculated in two steps. The first step uses the average percentage gain and loss over the period in which you are looking back. The standard lookback period is 14 days. Days where the stock gained value are counted as 0 in the average loss calculation and days where the stock lost value are counted as 0 in the average gain calculation. The formula for the first step is:

$$RSI_1 = 100 - \frac{100}{(1 + \frac{Average\ Gain}{Average\ Loss})}$$

The second calculation smooths the results over the lookback period similar to a moving average.

$$RSI_2 = 100 - \frac{100}{(1 + \frac{13 * Previous\ Gain + Current\ Gain}{13 * Previous\ Loss + Current\ Loss})}$$

Moving Average Convergence Divergence (MACD) - Momentum Indicator

The MACD is a trend-following momentum indicator that shows the relationship between two EMAs. Generally, the MACD is equal to the 26-day EMA subtracted from the 12-day EMA. For the purpose of trade signals, the MACD is plotted against its 9-day EMA which we call the signal line. The buy signal is when the MACD crosses above the signal line and the sell signal is when the MACD crosses below the signal line. This strategy is called MACD crossover.

Williams %R - Momentum Indicator

The Williams %R is a momentum indicator that oscillates between 0 and -100. It measures the overbought and oversold levels of the equity and is used to find entry and exit points in the market. It compares a stock's closing price to the high low range over a specific period.

$$Williams\ \%R = \frac{Highest\ High - Close}{Highest\ High - Lowest\ Low}$$

A value of -20 or higher indicates the equity is overbought and a value of -80 or lower indicates the equity is oversold.

Stochastic Oscillator - Momentum Indicator

The stochastic oscillator is a momentum indicator which compares the closing price to a range of prices over a period of time. It is bounded between 0 and 100 and is used to generate overbought or oversold signals.

$$Stochastic\ Oscillator = \frac{Close - Lowest\ Low}{Highest\ High - Lowest\ Low}$$

A value greater than 80 is an indication of overbought and a value less than 20 is an indication of oversold.

Relative Vigor Index (RVI) - Momentum Indicator

The RVI is a momentum indicator that measures the strength of a trend by comparing the equity's close to its trading range while using a SMA to smooth results. The RVI is calculated in two steps. First the numerator and denominator are calculated separately and then their SMA is taken.

$$Numerator\ = \frac{a + 2b + 2c + d}{6}$$

$$a\ =\ close\ -\ open$$

$$b\ =\ close\ -\ open\ one\ bar\ prior\ to\ a$$

$$c \;=\; close \;-\; open\ one\ bar\ prior\ to\ b$$

$$d \;=\; close \;-\; open\ one\ bar\ prior\ to\ c$$

$$Denimonator = \frac{e + 2f + 2g + h}{6}$$

$$e \;=\; high \;-\; low\ of\ bar\ a$$

$$f \;=\; high \;-\; low\ of\ bar\ b$$

$$g \;=\; high \;-\; low\ of\ bar\ c$$

$$h \;=\; high \;-\; low\ of\ bar\ d$$

$$RVI \;=\; \frac{SMA\ of\ numerator}{SMA\ of\ denominator}$$

The RVI is also plotted against a signal line and then uses crossovers as a trading signal. If the RVI crosses above the signal line it is a buy signal and if the RVI crosses below its signal line, then it is a sell signal.

$$Signal\ Line \;=\; \frac{RVI \;+\; 2i \;+ 2j \;+ k}{6}$$

$$i \;=\; RVI\ one\ bar\ prior$$

$$j \;=\; RVI\ one\ bar\ prior\ to\ i$$

$$k \;=\; RVI\ one\ bar\ prior\ to\ j$$

Another trading signal is RVI-price divergence. Divergence between the RVI and price suggests there will be a near-term change in the trend in the direction of the RVI’s trend. For example, if the equity is rising and RVI is falling, then the equity will correct itself in the near term.

Know Sure Thing Indicator (KST) - Momentum Indicator

KST is a momentum indicator that takes the SMA of four rate-of-change (ROC) periods and adds them together. KST is then compared against its signal line to indicate overbought and oversold trends.

$$KST \ = \ 10day\ SMA\ of\ 10day\ ROC \ + \ 2\ (\ 10day\ SMA\ of\ 15day\ ROC\ ) \ +$$

$$3\ (\ 10day\ SMA\ of\ 20day\ ROC\ ) \ + \ 4\ (\ 15day\ SMA\ of\ 30day\ ROC\ )$$

KST fluctuates above and below the 0 line. A positive KST indicates a bullish trend and a negative KST indicates a bearish trend. KST is also plotted against its signal line which is the 9-day SMA of the KST. When KST crosses over the signal line it is a buy signal and when KST crosses below the signal line it is a sell signal.

Disparity Index - Momentum Indicator

The disparity index is a momentum indicator that measures the relative position of the closing price to a selected moving average.

$$Disparity \ = \frac{Close \ - \ SMA}{SMA * 100}$$

When disparity is greater than zero the equity has upward momentum, when the disparity is less than zero the equity has downward momentum.

Commodity Channel Indicator (CCI) - Momentum Indicator

CCI is a momentum-based oscillator used to signal overbought and oversold trends by assessing the equity’s direction and strength.

$$CCI \ = \frac{Typical\ Price \ - \ SMA\ of\ Typical\ Price}{0.015 \ * \ |\,Typical\ Price \ - \ SMA\ of\ Typical\ Price\,|}$$

Overbought and oversold levels are not fixed since the indicator is unbounded. Once the signal levels are established a crossover technique is used for trading. When the CCI crosses above the overbought level it is a sell/short signal and when the CCI crosses below the oversold it is a buy signal.

Bollinger Band - Volatility Indicator

Bollinger Bands are a set of trendlines plotted two standard deviations away from the equity's SMA. The bands are typically calculated by adding and subtracting 2 standard deviations of the typical price with the SMA of the equity's typical price.

$$Upper\ Band\ = SMA(Typical\ Price\ ,n)\ +\ 2 * (Typical\ Price, n)$$

$$Lower\ Band\ = SMA(Typical\ Price\ ,n)\ -\ 2 * (Typical\ Price, n)$$

If the equity's price moves close to the upper band, then it indicates overbought levels and if the equity's price moves close to the lower band, then it indicates oversold levels.

Keltner Channel - Volatility Indicator

Keltner Channel uses volatility-based bands that are placed on either side of the equity's price. It indicates the direction of the equity's trend using the average true range (ATR). The true range takes the max of the high minus the low, the absolute value of the high minus the closing price, and the absolute value of the loss minus the closing price. The Keltner Channel is equal to the EMA of the equity's price plus/minus twice the ATR.

$$Upper\ Band\ = EMA\ of\ price\ +\ 2\ *\ ATR$$

$$Lower\ Band\ = EMA\ of\ price\ -\ 2\ *\ ATR$$

Price reaching the upper band indicates a bullish trend and price reaching the lower band indicates a bearish trend.

Aroon Indicator - Volume Indicator

The Aroon indicator is a technical indicator that is used to identify change in the equity's price and the strength of that trend. It measures the time between when highs are formed and when lows are formed. A strong uptrend will have regular new highs and a strong downtrend sees regular lows. The Aroon indicator has two components - Aroon Up which measures the strength of the uptrend, and Aroon Down which measures the strength of the downtrend. The general time frame of the Aroon indicator is 25 days.

$$Aroon\ Up\ = \frac{25\ -\ periods\ since\ 25\ period\ High}{25}$$

$$Aroon\ Down = \frac{25\ -\ periods\ since\ 25\ period\ Low}{25}$$

The trade signal for the Aroon indicator is the crossover between the Aroon Up and the Aroon Down. When the Aroon Up line crosses above the Aroon Down line it is a buy signal and when the Aroon Down line crosses above the Aroon Up line it is a sell/short signal.