MB 4013 – Multivariate Statistics

FINAL PROJECT REPORT

By:

Cliff Richardo
19020293

Undergraduate Program
School of Business and Management
Institut Teknologi Bandung
2021

Table of Contents

Table of Figures

Table of Tables

1. Background and Problem Definition

1.1. Background.

In this era, most teenagers are familiar with the word “investment”. Investment is the activity of placing capital in the form of money or other valuable assets into an object, institution, or party with the hope that the investor will gain the benefit after a certain period of time. Trading is one of the investments, where the transaction process takes place in the financial market where the system works is to frequently sell and buy assets. In trading, there are several important variables that must be considered in order to achieve the main goal which is profit, starting from knowing the trends, the latest news of a company, the company's future plans, etc. To maximize the output obtained for us, we must first understand the knowledge and strategy of punctuality in making decisions of trading transactions. However, what if we predict the stock value in the future from the value data variable and the increase in stock value.

Scalping is a short term trading method to gain profit from trading volume faster. Scalping is one of the models that is often used for trading. In trading, scalping becomes a priority strategy in making high volumes from small profits.

PT Bank MNC Internasional Tbk is a subsidiary of MNC Financial Services which is engaged in banking. MNC Bank started its commercial operations on January 12, 1990. The bank is part of the MNC Group where the main parent company is PT. MNC Investama Tbk. PT Bank MNC Internasional Tbk has a structured history of stock data and the data held by MNC regarding stock is sufficient for us to carry out our analysis topics. Therefore, we will try to do a scalping model using logistic regression on stock values of PT Bank MNC Internasional Tbk.

1.2. Objectives.

By utilizing using **SMART** goals model, the objectives of the study are as follows:

1. **S**pecific : Create a scalping model using logistic regression that can be understood by traders regarding the stock value of PT Bank MNC Internasional Tbk.
2. **M**easurable : Combine and measure the stock value data variable of PT Bank MNC Internasional Tbk. so as to get a conclusion in the scalping model.
3. **A**ttainable : In order to facilitate decision making and future strategies to trade shares of PT Bank MNC Internasional Tbk. in the short term.
4. **R**elevant : The greater the percentage of data analysis results that can be concluded in, the better the conclusions obtained.
5. **T**ime-Bounded : Get analysis conclusions and predictions for traders in the short term because it is in accordance with the scalping model.

So if we drag all the points, we have a goals to constructing a model of stock values of PT Bank MNC Internasional Tbk using logistic regression (statistical analysis that is often used for predictive modeling) with stock data variables of their company in order to give the reader a suggestion in scalping trade.

1.3. Scope.

- Data open price time of the day in MNCBank from 2014-2021.
- Data change Price time of the day in MNCBank from 2014-2021.
- Data change volume time of the day (using lot) in MNCBank from 2014-2021.
- Data interval high open in MNCBank from 2014-2021.
- Data interval open from the last day in MNCBank from 2014-2021.

2. Methodology

2.1. Data Collection Method.

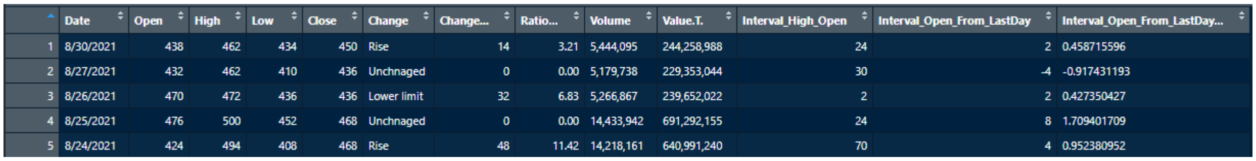
For the purposes of this statistical research, we used secondary data that we found from Kaggle, an online community of data scientists and machine learning practitioners. Kaggle allows users to find and publish datasets, from that we then find a dataset of MNC Bank's stock value from 2014-2021.

Listed below are the known variables and their description,

Table 2.1 Variable of Dataset MNC Bank's stock

Variable	Description
Date	Date of the known stock value
Open	Open price of the stock on the date
High	Highest known price on the date
Low	Lowest known price on the date
Close	Closing price of the stock on the date
Change	Description of the stock price changes that happened the day before the date
Change(%)	Value of the stock price changes that happen within the date
Ratio(%)	Percentage of the stock price changes that happen within the date
Volume	Volume of trades that happens during the date
Value(T)	Changes of value that happens during the date
Interval_High_Open	The interval between the highest known price and the open price within the date
Interval_Open_From_LastDay	The interval between the open price of the day before and the open price within the date
Interval_Open_From_LastDay(%)	The percentage of changes between the open price of the day before and the open price within the date

Listed below are 5 rows from the raw data that we gathered from kaggle.



	Date	Open	High	Low	Close	Change	Change...	Ratio...	Volume	Value.T.	Interval_High_Open	Interval_Open_From_LastDay	Interval_Open_From_LastDay...
1	8/30/2021	438	462	434	450	Rise	14	3.21	5,444,095	244,258,988	24	2	0.458715596
2	8/27/2021	432	462	410	436	Unchnaged	0	0.00	5,179,738	229,353,044	30	-4	-0.917431193
3	8/26/2021	470	472	436	436	Lower limit	32	6.83	5,266,867	239,652,022	2	2	0.427350427
4	8/25/2021	476	500	452	468	Unchnaged	0	0.00	14,433,942	691,292,155	24	8	1.709401709
5	8/24/2021	424	494	408	468	Rise	48	11.42	14,218,161	640,991,240	70	4	0.952380952

Figure 2.1 Example of Raw Data

2.2. Variable Identification.

When we previewed the dataset, we found that there were 13 variables within the collected dataset. We try to identify the independent variables that are collinear to each other through multicollinearity testing, then we select 5 relevant variables in order to pursue whether the price of MNC Bank shares will rise or fall within a day (*scalping trade*). Since we believe that analyzing the opening prices, volumes traded, and the interval of yesterday's price are crucial in order to predict changes in stock prices. Out of 5 relevant variables, we decided that the variable '*Change*' should be the dependent variable (predicted variable) and the rest should be the independent variable (predictor variable) since we want to create a model to find the influence of '*Open*', '*Volume*', '*Interval_High_Open*', and '*Interval_Open_From_LastDay*' to the variable '*Change*'.

To understand more about the variables, we identified the measurement scales for each selected variable. The variable '*Change*' uses a factor type of data that have 5 known levels, which is "Rise", "Unchanged", "Lower Limit", "Upper Limit", and "Fall", the variable '*Open*', '*Interval_Open_From_LastDay*', and '*Interval_High_Open*' uses rupiah/share, the variable '*Volume*' uses quantity of shares traded that day. For our business issue, we only use 2 levels for the predicted variable '*Change*', which is "Rise" and "Fall".

2.3. Statistical Method.

The methodology that we agreed to use for this analysis is the **Binary Logistic Regression Analysis** that could later be used to calculate the probability of rise or fall of the price of MNC Bank shares within a day. Regression analysis presents the association between a response variable and one or more explanatory variables. Since the business issue is whether the model can predict the rise or fall of MNC Bank share price, we decided to use binary logistic regression as the most appropriate method to answer the business issue.

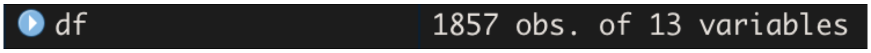
We decided to use Binary Logistic Regression because of the problem that we want to solve. Since we decided to use '*Change*' as the dependent variable and selected 2 main factors to be predicted.

3. Result & Analysis

3.1. Extract, Load, Filter, and Transform.

3.1.1. Extract & Load Data Frame.

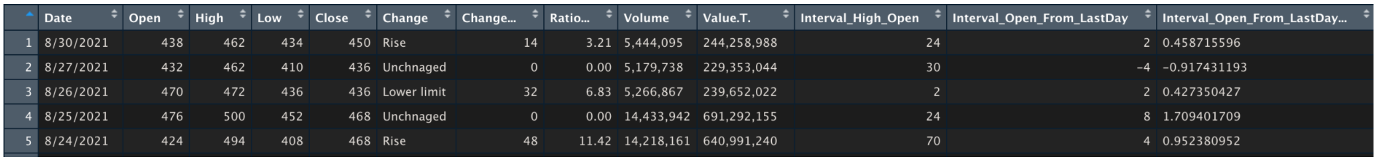
After we set the working directory and set the variable ‘*df*’ to The CSV file, we can see that the data frame consists of 1357 observations and 13 variables.



```
df        1857 obs. of 13 variables
```

Figure 3.1 *df* of 1357 Observations and 13 Variables

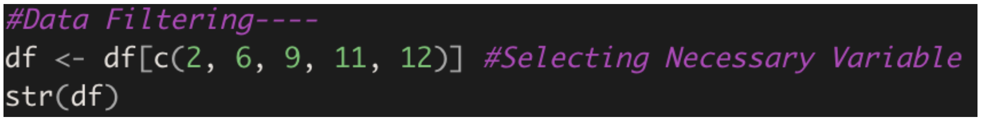
Visualization of raw data frame before filtering and transforming process:



	Date	Open	High	Low	Close	Change	Change...	Ratio...	Volume	Value.T.	Interval_High_Open	Interval_Open_From_LastDay	Interval_Open_From_LastDay...
1	8/30/2021	438	462	434	450	Rise	14	3.21	5,444,095	244,258,988	24	2	0.458715596
2	8/27/2021	432	462	410	436	Unchnaged	0	0.00	5,179,738	229,353,044	30	-4	-0.917431193
3	8/26/2021	470	472	436	436	Lower limit	32	6.83	5,266,867	239,652,022	2	2	0.427350427
4	8/25/2021	476	500	452	468	Unchnaged	0	0.00	14,433,942	691,292,155	24	8	1.709401709
5	8/24/2021	424	494	408	468	Rise	48	11.42	14,218,161	640,991,240	70	4	0.952380952

Figure 3.2 Raw Data Frame Before Filtering and Transforming

3.1.2. Filtering Data Frame.



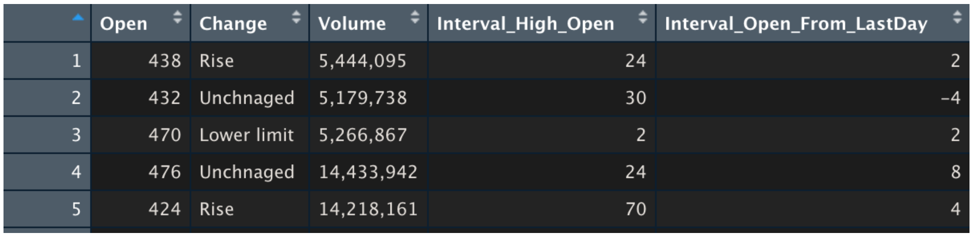
```
#Data Filtering----
df <- df[c(2, 6, 9, 11, 12)] #Selecting Necessary Variable
str(df)
```

Figure 3.3 Data Filtering Process

After we load the data frame and necessary library, we start the filtering process which means we drop all unnecessary variable column, leaving the important one (column 2, 6, 9, 11, and 12) that consists of:

1. *Open*: Opening price of the MNC Bank stock at the moment the stock exchange market opens at 9 A.M. Western Indonesia Time.
2. *Change*: Whether the price of MNC Bank stock rises or falls from their yesterday closing price.
3. *Interval_High_Open*: The interval between all time high prices of the stock and the opening price of the stock that day.
4. *Interval_Open_From_LastDay*: The interval between the present opening price and yesterday's closing price.
5. Volume: Total volume of MNC Bank stock traded that day.

Visualization of raw data frame before transforming process:



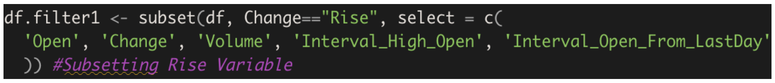
	Open	Change	Volume	Interval_High_Open	Interval_Open_From_LastDay
1	438	Rise	5,444,095	24	2
2	432	Unchnaged	5,179,738	30	-4
3	470	Lower limit	5,266,867	2	2
4	476	Unchnaged	14,433,942	24	8
5	424	Rise	14,218,161	70	4

Figure 3.4 Raw Data Frame Before Transforming

3.1.3. Transforming Data Frame.

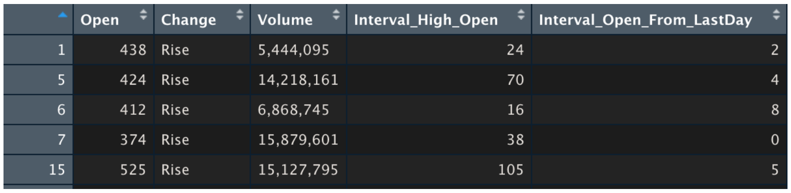
We construct three new variables, ‘*df.filter1*’, ‘*df.filter2*’, ‘*df.FINAL*’ for the transforming process. Since we want to predict whether the price of MNC Bank stock will increase or decrease through the variable ‘*Change*’ and the variable ‘*Change*’ have 5 unique strings (Rise, Fall, Upper Limit, Lower Limit, and Unchange) we only want the condition of that variable to be Rise or Fall only.

‘*df.filter1*’ will subset the variable ‘*Change*’ that includes the character **Rise**. The variable consists of 367 observations.



```
df.filter1 <- subset(df, Change=="Rise", select = c(
  'Open', 'Change', 'Volume', 'Interval_High_Open', 'Interval_Open_From_LastDay'
  )) #Subsetting Rise Variable
```

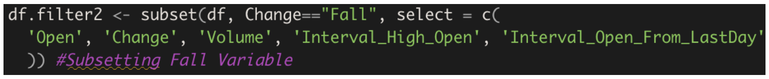
Figure 3.5 Codes of Subsetting Rise Variable



	Open	Change	Volume	Interval_High_Open	Interval_Open_From_LastDay
1	438	Rise	5,444,095	24	2
5	424	Rise	14,218,161	70	4
6	412	Rise	6,868,745	16	8
7	374	Rise	15,879,601	38	0
15	525	Rise	15,127,795	105	5

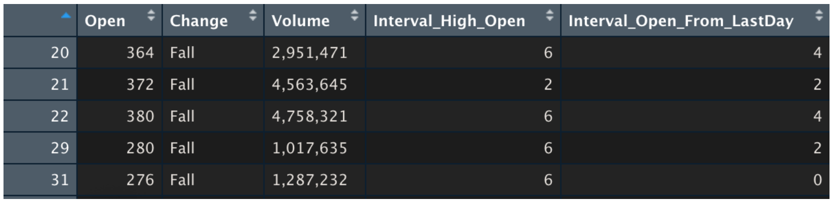
Figure 3.6 The Changes After Subsetting Rise Variable

‘*df.filter2*’ will subset the variable ‘*Change*’ that includes the character **Fall**. The variable consists of 439 observations.



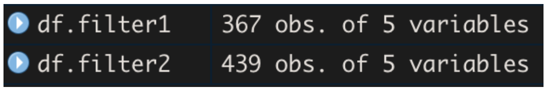
```
df.filter2 <- subset(df, Change=="Fall", select = c(
  'Open', 'Change', 'Volume', 'Interval_High_Open', 'Interval_Open_From_LastDay'
  )) #Subsetting Fall Variable
```

Figure 3.7 Codes of Subsetting Fall Variable



	Open	Change	Volume	Interval_High_Open	Interval_Open_From_LastDay
20	364	Fall	2,951,471	6	4
21	372	Fall	4,563,645	2	2
22	380	Fall	4,758,321	6	4
29	280	Fall	1,017,635	6	2
31	276	Fall	1,287,232	6	0

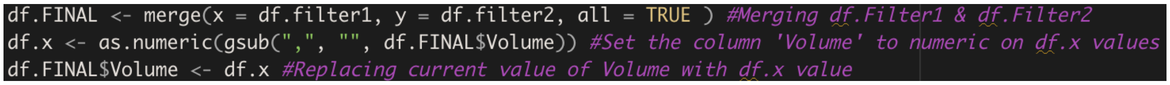
Figure 3.8 The Changes After Subsetting Fall Variable



```
df.filter1    367 obs. of 5 variables
df.filter2    439 obs. of 5 variables
```

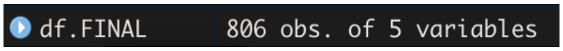
Figure 3.9 Total Variable After Subset

‘*df.FINAL*’ will merge ‘*df.filter1*’ and ‘*df.filter2*’ into one data frame that consists of a total of **806 observations**. After the merge happens, since the ‘*Volume*’ value is a character that is separated by comma (‘ , ’), we have to transform it into a numeric structure and set it into a different variable before replacing it with the original ‘*df.FINAL*’ dataframe values.



```
df.FINAL <- merge(x = df.filter1, y = df.filter2, all = TRUE ) #Merging df.Filter1 & df.Filter2
df.x <- as.numeric(gsub(",", "", df.FINAL$Volume)) #Set the column 'Volume' to numeric on df.x values
df.FINAL$Volume <- df.x #Replacing current value of Volume with df.x value
```

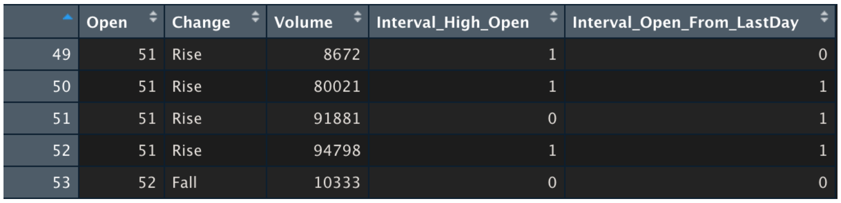
Figure 3.10 Merge Data Frame



```
df.FINAL      806 obs. of 5 variables
```

Figure 3.11 Total Variable of Merge

Visualization of the final and cleaned data frame:

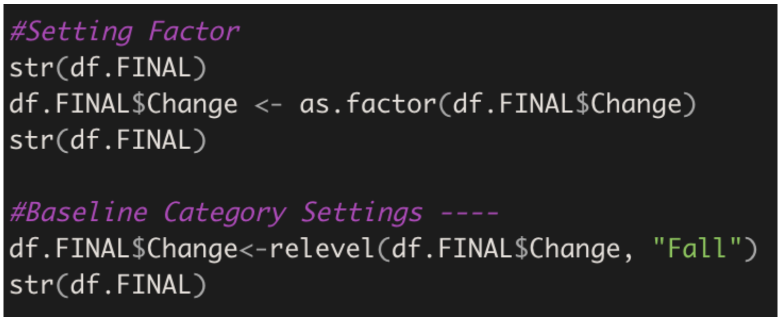


	Open	Change	Volume	Interval_High_Open	Interval_Open_From_LastDay
49	51	Rise	8672	1	0
50	51	Rise	80021	1	1
51	51	Rise	91881	0	1
52	51	Rise	94798	1	1
53	52	Fall	10333	0	0

Figure 3.12 Cleaned Data Frame

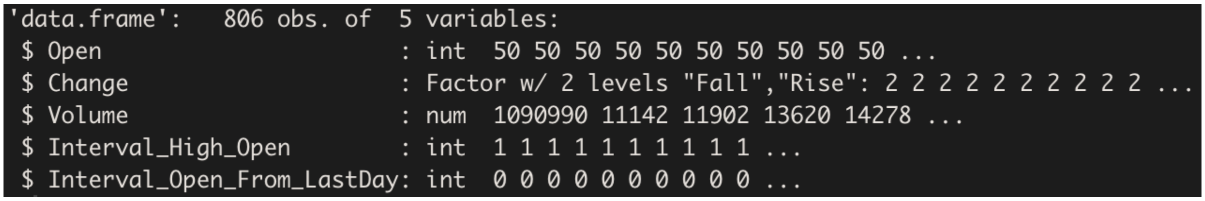
3.2. Constructing the Model.

Before constructing the model, we have to reformat ‘*Change*’ values from character into a factor. After that, we have to set the baseline category of ‘*Change*’ to set the FALL factor into the first factor, since we’re trying to construct a model that tells us whether the stock will RISE or not. We’re also decided to set the threshold values or **alpha(α)** to be **0.05.**



```
#Setting Factor
str(df.FINAL)
df.FINAL$Change <- as.factor(df.FINAL$Change)
str(df.FINAL)

#Baseline Category Settings ----
df.FINAL$Change<-relevel(df.FINAL$Change, "Fall")
str(df.FINAL)
```

Figure 3.13 Codes of Setting Factor and Baseline Category

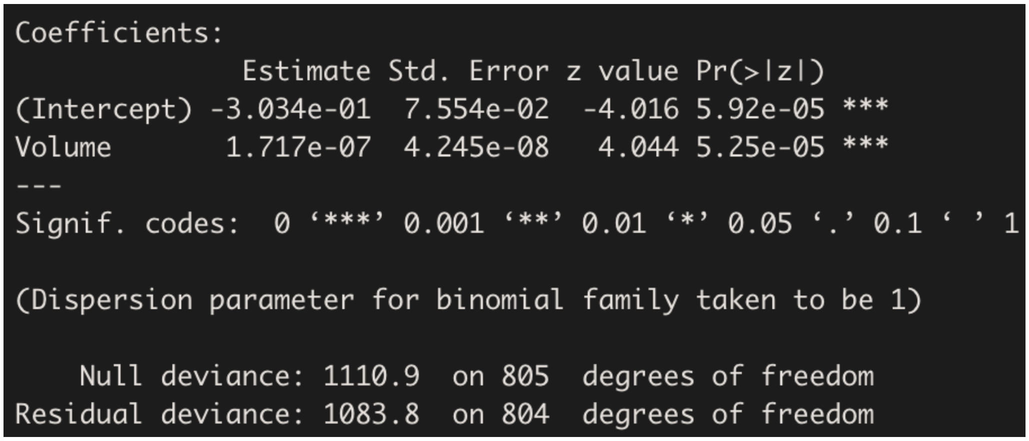


```
'data.frame':   806 obs. of  5 variables:
 $ Open                     : int  50 50 50 50 50 50 50 50 50 50 ...
 $ Change                   : Factor w/ 2 levels "Fall","Rise": 2 2 2 2 2 2 2 2 2 2 ...
 $ Volume                   : num  1090990 11142 11902 13620 14278 ...
 $ Interval_High_Open       : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Interval_Open_From_LastDay: int  0 0 0 0 0 0 0 0 0 0 ...
```

Figure 3.14 Result Setting Factor and Baseline Category

After we set the baseline category, we can construct the model right away. In our cases, we're going to try to construct two models for comparison.

1. **Model 1** → Include '*Change*' as the dependent variable and 'Volume' as the independent variable. The summary of the model are shown in below figure:



```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept) -3.034e-01  7.554e-02  -4.016 5.92e-05 ***
Volume       1.717e-07  4.245e-08   4.044 5.25e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1110.9  on 805  degrees of freedom
Residual deviance: 1083.8  on 804  degrees of freedom
```

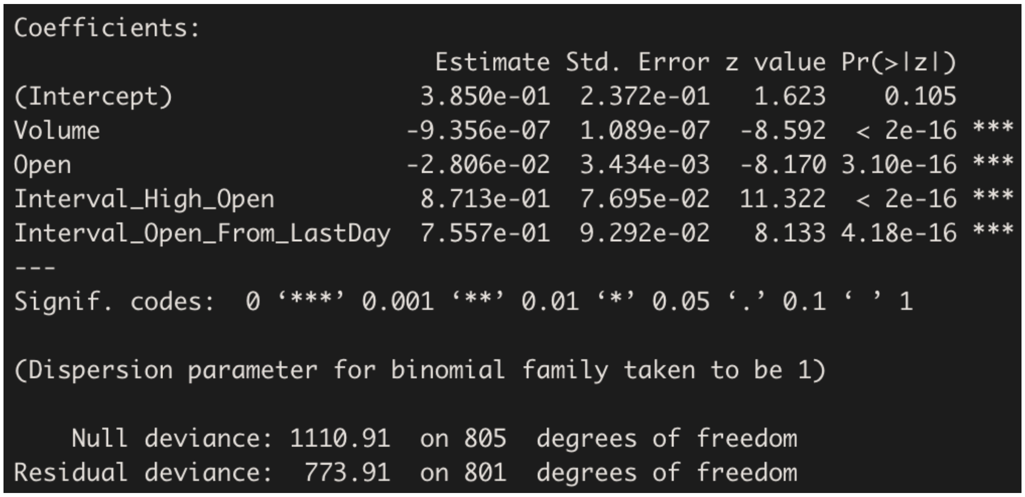
Figure 3.15 Model 1

Both variables of the first model shows that they have a probability of Z or Pr(>|z|) values **under** the threshold values which is 0.05. This means that both variables are statistically significant. **The model representation is**:

$$e^{-0.3034 + 0.0000001\ (X1)}$$

Figure 3.16 Representation of Model 1

2. **Model 2** → Include ‘*Change*’ as the dependent variable and the remaining ‘*df.FINAL*’ data frame variable as an independent variable (‘*Volume*’, ‘*Open*’, ‘*Interval_High_Open*’, and ‘*Interval_Open_From_LastDay*’). The summary of the model are shown in below figure:



```
Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                3.850e-01  2.372e-01   1.623    0.105
Volume                    -9.356e-07  1.089e-07  -8.592  < 2e-16 ***
Open                      -2.806e-02  3.434e-03  -8.170 3.10e-16 ***
Interval_High_Open         8.713e-01  7.695e-02  11.322  < 2e-16 ***
Interval_Open_From_LastDay 7.557e-01  9.292e-02   8.133 4.18e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1110.91  on 805  degrees of freedom
Residual deviance:  773.91  on 801  degrees of freedom
```

Figure 3.17 Model 2

The dependent variable of the second model shows that they have a probability of Z or Pr(>|z|) values **over** the threshold values which is 0.05 meaning that the independent variable data are not sufficient to statistically distinguish the estimated value, thus it is statistically insignificant meaning that we cannot distinguish the probability of the event (given all X=0) from 0.5 (the data are compatible with probabilities larger and lower 0.5). The intercept is simply a constant which shows us how far up the model can move before we begin to understand the effect and gives the expected outcome when all predictors take on the value of zero.

Since the model we develop are predicting whether the price of MNC Bank will rise or fall, and that all of the predictors are impossible to have a value of 0 since it will be meaningless unless the share is being delisted from IDX, we decided to keep the constant and preserve the model. This decision is also backed up by experts from international universities (see references).

Meanwhile all of the independent variables Pr(>|z|) value is **under** the threshold value meaning all of the independent variables are statistically significant. Based on the model 2 coefficients, **the model representation is**:

$$e^{0.385 - 0.0000009(X1) - 0.028(X2) + 0.8713(X3) + 0.7557(X4)}$$

Figure 3.18 Representation of Model 2

3.2.1. Testing Multicollinearity.

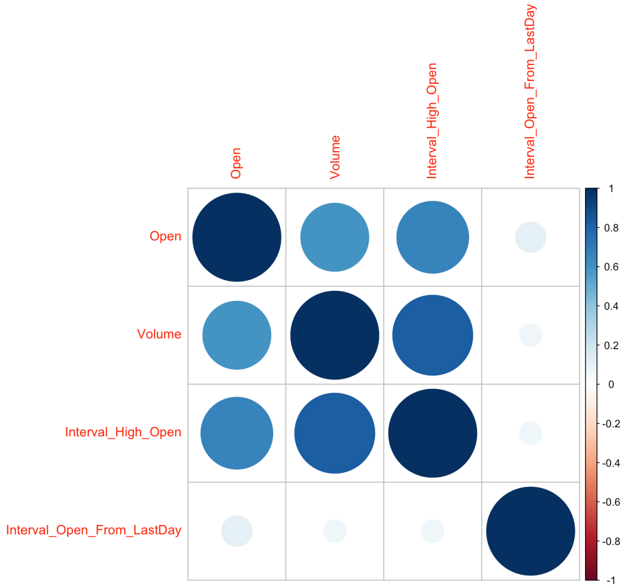
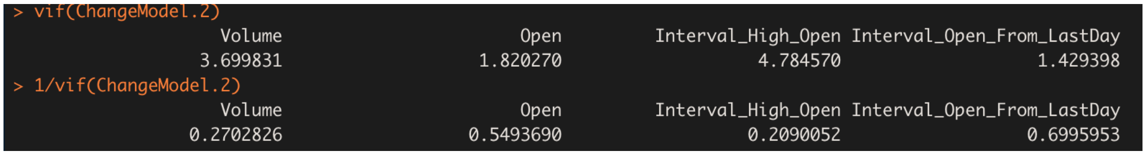


Figure 3.19 Testing Multicollinearity

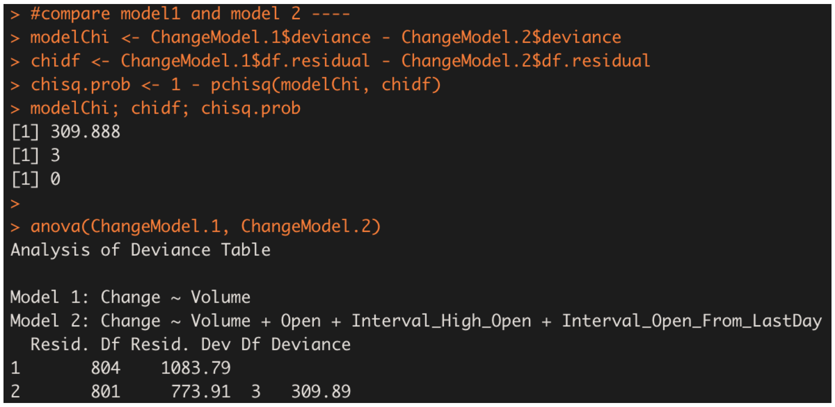


```
> vif(ChangeModel.2)
                    Volume                       Open         Interval_High_Open Interval_Open_From_LastDay 
                  3.699831                   1.820270                   4.784570                   1.429398 
> 1/vif(ChangeModel.2)
                    Volume                       Open         Interval_High_Open Interval_Open_From_LastDay 
                 0.2702826                  0.5493690                  0.2090052                  0.6995953 
```

Figure 3.20 vif and 1/vif of Model 2

Based on the Variance Inflation Factor(VIF), we can conclude that the model **fulfilled** the **assumption** since the largest VIF value is not >10 and there is no VIF tolerance <0.1. And thus, the testing of multicollinearity does **not raise any concern and serious problems**.

3.2.2. Comparing Model 1 and Model 2.



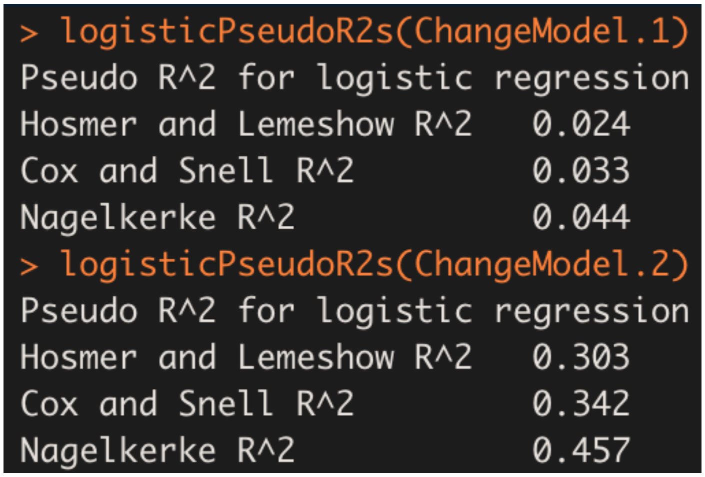
```
> #compare model1 and model 2 ----
> modelChi <- ChangeModel.1$deviance - ChangeModel.2$deviance
> chidf <- ChangeModel.1$df.residual - ChangeModel.2$df.residual
> chisq.prob <- 1 - pchisq(modelChi, chidf)
> modelChi; chidf; chisq.prob
[1] 309.888
[1] 3
[1] 0
> 
> anova(ChangeModel.1, ChangeModel.2)
Analysis of Deviance Table

Model 1: Change ~ Volume
Model 2: Change ~ Volume + Open + Interval_High_Open + Interval_Open_From_LastDay
  Resid. Df Resid. Dev Df Deviance
1       804    1083.79            
2       801     773.91  3   309.89
```

Figure 3.21 Comparison of Model 1 and 2

Since the residual deviance of model 2 shows significant improvement from 1083.79 to 773.91 with 3 degrees of freedom and 0 (or almost 0) P-Value, we can conclude that model 2 shows significant improvement over model 1. Thus we decided to **reject model 1** and **not reject model 2**.

3.2.3. Pseudo R² Comparison.

In order to further prove our decisions in **not rejecting model 2**, we conduct a pseudo R² analysis of both models, the results are shown in the figure below.



```
> logisticPseudoR2s(ChangeModel.1)
Pseudo R^2 for logistic regression
Hosmer and Lemeshow R^2   0.024
Cox and Snell R^2         0.033
Nagelkerke R^2            0.044
> logisticPseudoR2s(ChangeModel.2)
Pseudo R^2 for logistic regression
Hosmer and Lemeshow R^2   0.303
Cox and Snell R^2         0.342
Nagelkerke R^2            0.457
```

Figure 3.22 Logistic Pseudo R² Model 1 and 2

The mean of pseudo R² in **model 1** has a value of **0.0336**, meanwhile the **second model** has a value of **0.3673**. In overall terms, the **second model has a better R²** meaning that **36.73%** of the total variation of changes are explained by the logistic model consisting of opening prices, volume of the stocks traded that day, interval between all time high prices of the stock and the opening price of the stock that day, and interval between the present opening price and yesterday's opening price.

4. Conclusion and Implications.

4.1. Conclusion.

From the model we constructed, we reject the first model while not rejecting the second model since in comparison, the residual deviance of the second model shows significant improvement over the first model. The second model consists of the variable '*Change*' as the dependent variable(predicted variable) and '*Open*', '*Volume*', '*Interval_High_Open*', and '*Interval_Open_From_LastDay*' as the independent variables (predictor variable). The model representation is:

$$e^{0.385 - 0.0000009(X1) - 0.028(X2) + 0.8713(X3) + 0.7557(X4)}$$

Figure 4.1 Model Presentation

In order to prove that our model fulfills the assumption of binary logistic (**no multicollinearity**), we conduct a Variance Inflation Factor (VIF) and also construct a correlation plot for visualization. We found that statistically, **our model does not represent any multicollinearity**, which means that the model is **proper** enough **to be used**. Our analysis of Pseudo R^2 explains that **36.73%** of the total variation of changes can be explained by the four independent variables that we selected.

4.2. Implications.

The forecasting model we develop could allow us to evaluate whether the changes of the prices within a day (*scalping trade*) will result in rises or fall of MNC Bank stock price, in the case that we knew the open price of the day before and that current date, the interval between the highest price point and the opening price of that current date, and the trading volume that happen on the day before. With the application of this model, we may help amateur traders who are interested in *scalping trade* to determine whether or not the current date is feasible to buy or sell the shares of MNC Bank that they currently have. This statistical analysis would also help people who are interested in learning about *scalping trade* but are confused on the calculation that the traders already have. Further technical analysis of shares may be needed in order to analyze the estimation of the all-time high prices of stock and volume shares traded during a day period that are needed in using our model.

References

1. Panji, Aryo. “Saham MNC Bank 2014 - 2021 - Indonesia Stock.” Saham MNC Bank 2014 - 2021 - Indonesia Stock | Kaggle, www.kaggle.com, 2021, https://www.kaggle.com/datasets/aryopanji/saham-mnc-bank-2014-2021-indonesia-stock.
2. Richardo, Cliff. “Research Gate.” Research Gate, 28 July 2022, https://www.researchgate.net/post/What_happens_if_in_binary_logistic_regression_the_interceptb0_is_statistically_not_significant.