MB 4013 - MULTIVARIATE STATISTICS

Final Project

Cliff Richardo - 19020293

Report Overview

TABLE OF CONTENTS

Background and Problem Definition

Background

In this era, most teenagers are familiar with the word “investment”. Trading is one of investments instrument, where the transaction process takes place in the financial market where the system works is to frequently sell and buy assets.

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.

Background

PT Bank MNC Internasional Tbk is a subsidiary of MNC Financial Services which is engaged in banking. 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.

OBJECTIVES


SMART
Specific
Measurable
Attainable
Relevant
Time Based

SCOPE OF THE CASE:

- 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.

Methodology

DATA COLLECTION METHOD

Kaggle as the source of secondary data

VARIABLE IDENTIFICATION

Found 13 variables, selected 5 relevant variables

STATISTICAL METHOD

We chose Binary Logistic Regression Analysis

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

Result & Analysis

Extract & Load Data Frame

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	239,652,822	2	2	0.427158427
4	8/25/2021	426	500	412	468	Unchnaged	0	0.00	14,413,942	651,292,155	24	8	1.709401709
5	8/24/2021	424	494	408	468	Rise	48	11.42	14,218,161	640,993,240	70	4	0.952380952

Figure 3.2 Raw Data Frame Before Filtering and Transforming

Filtering Data Frame

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

Figure 3.3 Data Filtering Process

- 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

Transforming Data Frame

- Subsetting "RISE"

```
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

Transforming Data Frame

- Subsetting "FALL"

```
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

Transforming Data Frame

- Merging data frame

```
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

	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

Constructing the Model

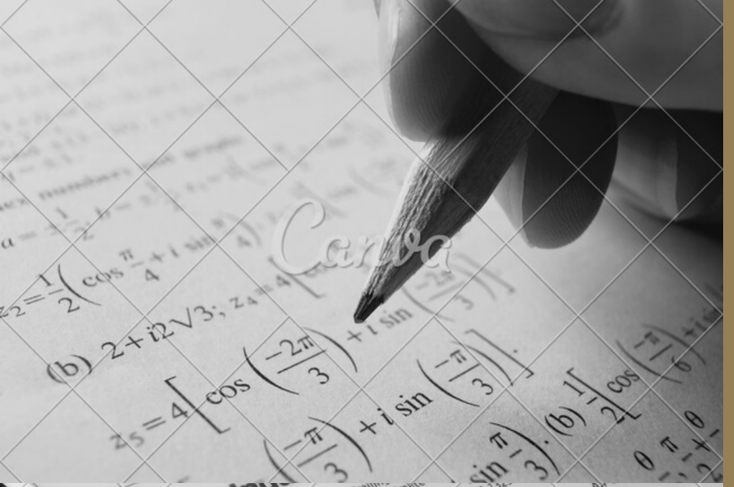
```
#Setting Factor
str(df.FINAL)
df.FINAL$Change <- as.factor(df.FINAL$Change)
str(df.FINAL)

#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



Constructing the Model

```
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

- Model Representation:

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

Constructing the Model

```
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

- Model Representation:

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

Testing Multicollinearity.

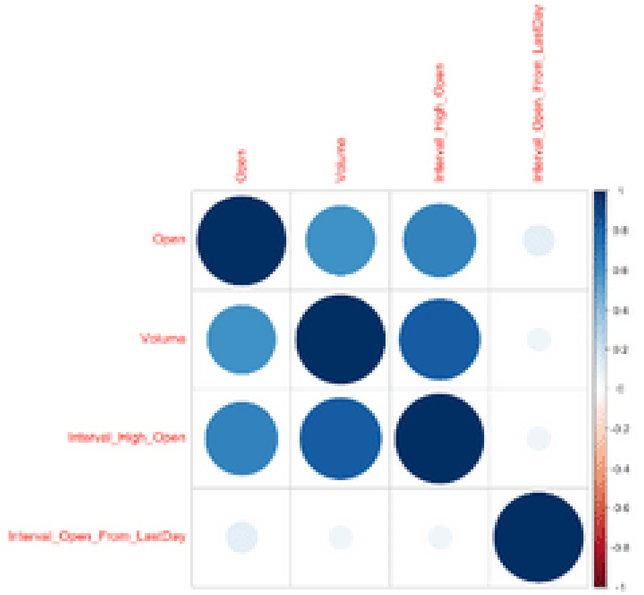


Figure 3.19 Testing Multicollinearity

```
> 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

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 Comparation of Model 1 and 2

Pseudo R^2 Comparison

```
> 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 R^2 Model 1 and 2

Conclusion and Implications

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:

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

Conclusion

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.

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

PANJI, ARYO. “SAHAM MNC BANK 2014 - 2021 - INDONESIA STOCK.” SAHAM MNC BANK 2014 - 2021 - INDONESIA STOCK | KAGGLE,

https://www.kaggle.com/datasets/aryopanji/saham-mnc-bank-2014-2021-indonesia-stock.

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.

THANK YOU !