WQD 7005 Data Mining

Milestone 5: Communication of insights of data

Name: Ng Kang Wei

Student ID: WQD170068

Video presentation link: https://youtu.be/oLSD_SEunQ0 (please turn on subtitles or caption)

From the correlation analysis, the stocks in focus here are Sime Darby Plantation Berhad (5285) and some stocks that have positive correlation and negative correlation with Sime Darby. The stocks that have positive correlations with Sime Darby includes, Kawan Food Bhd (7216), Caely Holdings Bhd (7154), and Emico Holdings Bhd (9091). The stocks that have negative correlations with Sime Darby includes Hock Heng Stone Industries Bhd (5165), Daibochi Berhad (8125), and FACB Industries Incorporated (2984). In this milestone, a decision tree and logistic regression would be constructed based on the stock data.

Subset and flag

The 7 stocks data are extracted from the overall stock data. Based on this subset data, new columns are added to this data.

The change flag which is depend on the change percentage, if the change percentage is positive, the value of change flag would be 'pos', else if the change percentage is negative, then the value of change flag would be 'neg', otherwise, the change flag would be 'non'.

Other than the change flag, there is another flag added namely the trade flag. This flag would determine should the investors buy the stock or sell the stock or hold the stock depend on the price of the stock. If the price is other than the one mentioned, then the investors should hold the stock. There would be 3 flags namely, Buy, Sell and Hold.

The rules to label the Trade Flag for each stock is as follows:

Stock	Buy	Sell
Sime Darby Plantation Berhad (5285)	<= 5.03	>= 5.10
Caely Holdings Berhad (7216)	<= 1.00	>= 1.50
Emico Holdings Berhad (9091)	<= 0.15	>= 0.17
Hock Heng Stone Industries Berhad (5165)	<= 0.50	>= 0.60
Daibochi Berhad (8125)	<= 1.20	>= 1.50
FACB Indsutries Incorporated (2984)	<= 0.80	>= 1.20

Decision Tree

The subset of the stock data is imported into SAS Enterprise Miner in CSV format. The CSV data is changed into SAS data format. After that, decision tree and logistic regression is performed on the data.

The SAS analysis diagram is shown as follows:

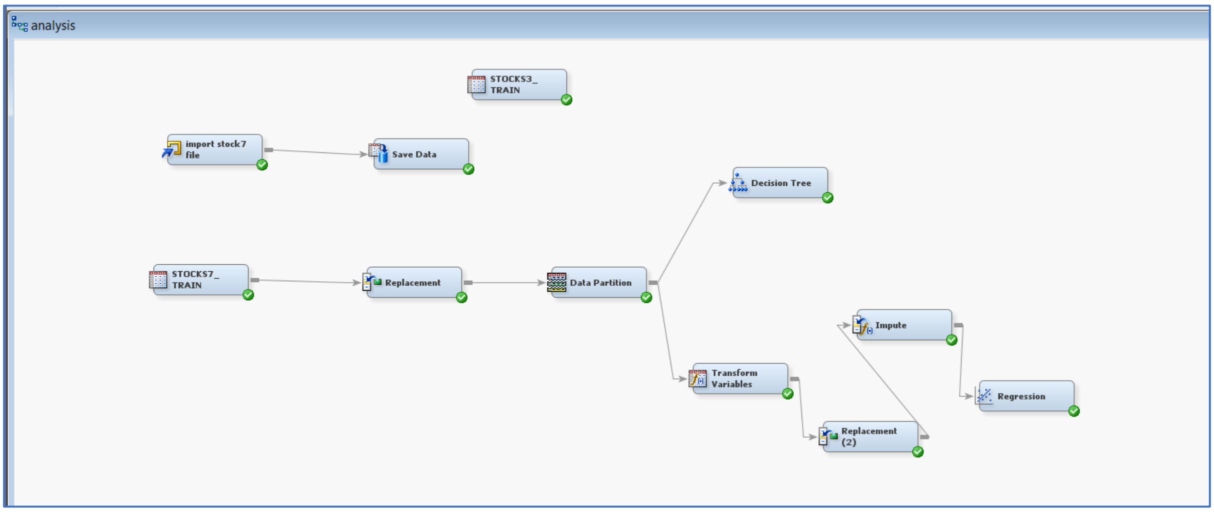
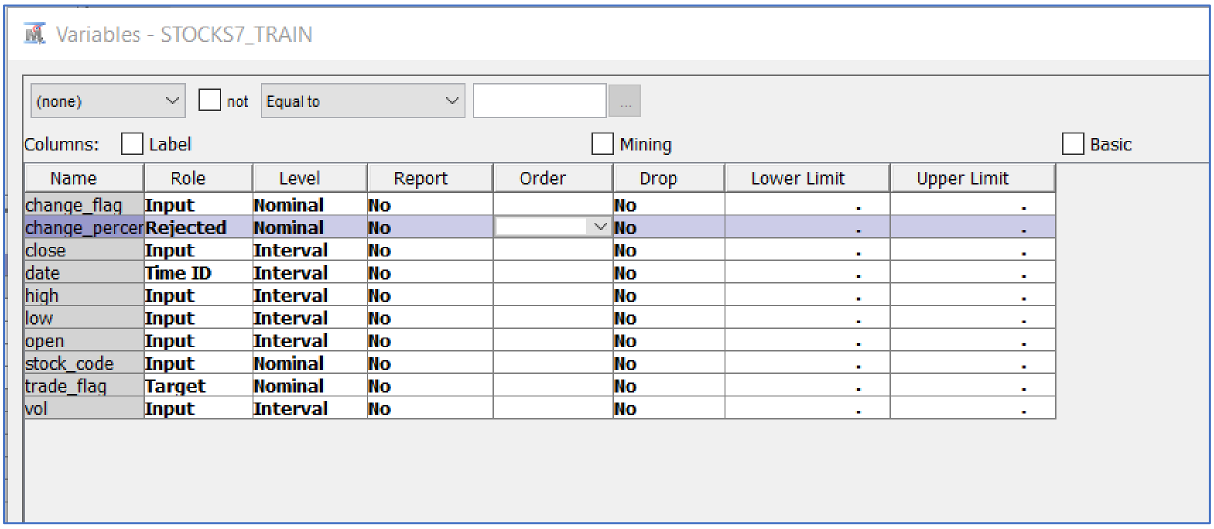


Figure 1: SAS analysis workflow

When the data is imported, the trade flag column is specified as the target column. The decision tree and logistic regression would predict the value of the trade flag based on the training.



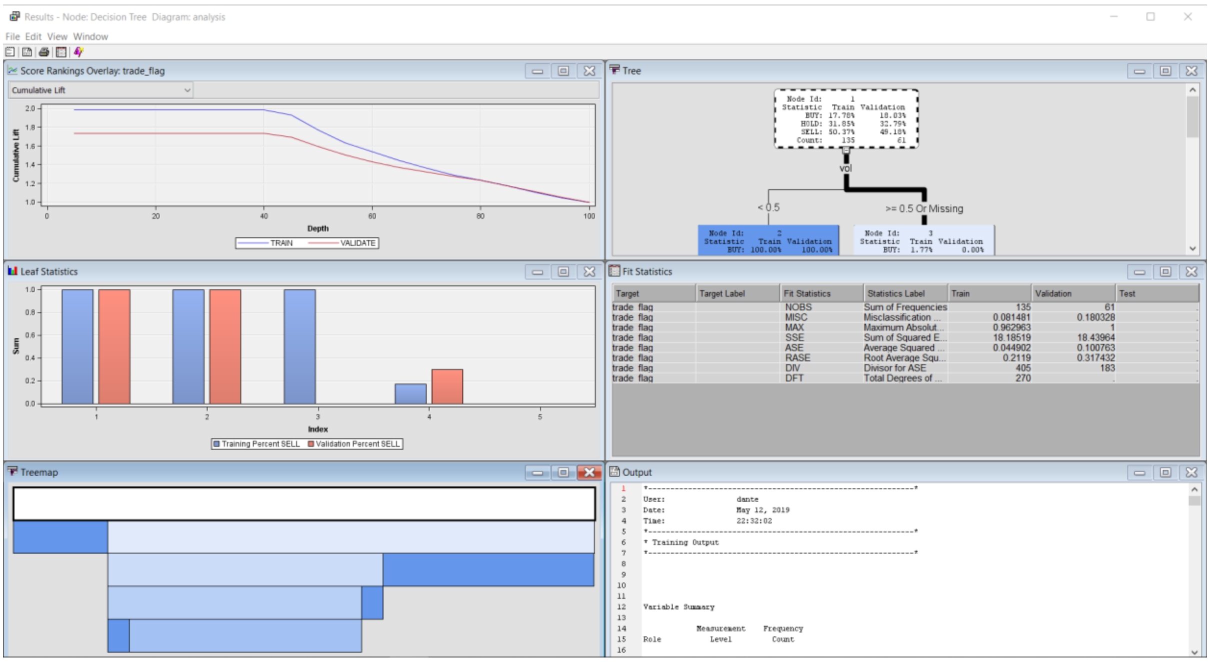
Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
change_flag	Input	Nominal	No		No	.	.
change_percer	Rejected	Nominal	No		No	.	.
close	Input	Interval	No		No	.	.
date	Time ID	Interval	No		No	.	.
high	Input	Interval	No		No	.	.
low	Input	Interval	No		No	.	.
open	Input	Interval	No		No	.	.
stock_code	Input	Nominal	No		No	.	.
trade_flag	Target	Nominal	No		No	.	.
vol	Input	Interval	No		No	.	.

Figure 2: SAS variables role

The data is partitioned into 2 parts, namely training data and validation data. The ratio is 7:3.

In SAS, the method of maximal decision tree is used. The following screenshots show the results of maximal tree.

Results of maximal decision tree



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
trade_flag		NOBS	Sum of Frequencies	135	61	.
trade_flag		MISC	Misclassification ...	0.081481	0.180328	.
trade_flag		MAX	Maximum Absolut...	0.962963	1	.
trade_flag		SSE	Sum of Squared E...	18.18519	18.43964	.
trade_flag		ASE	Average Squared ...	0.044902	0.100763	.
trade_flag		RASE	Root Average Squ...	0.2119	0.317432	.
trade_flag		DIV	Divisor for ASE	405	183	.
trade_flag		DFT	Total Degrees of ...	270	.	.

Figure 3: Results of Maximal Tree

The maximal tree diagram:

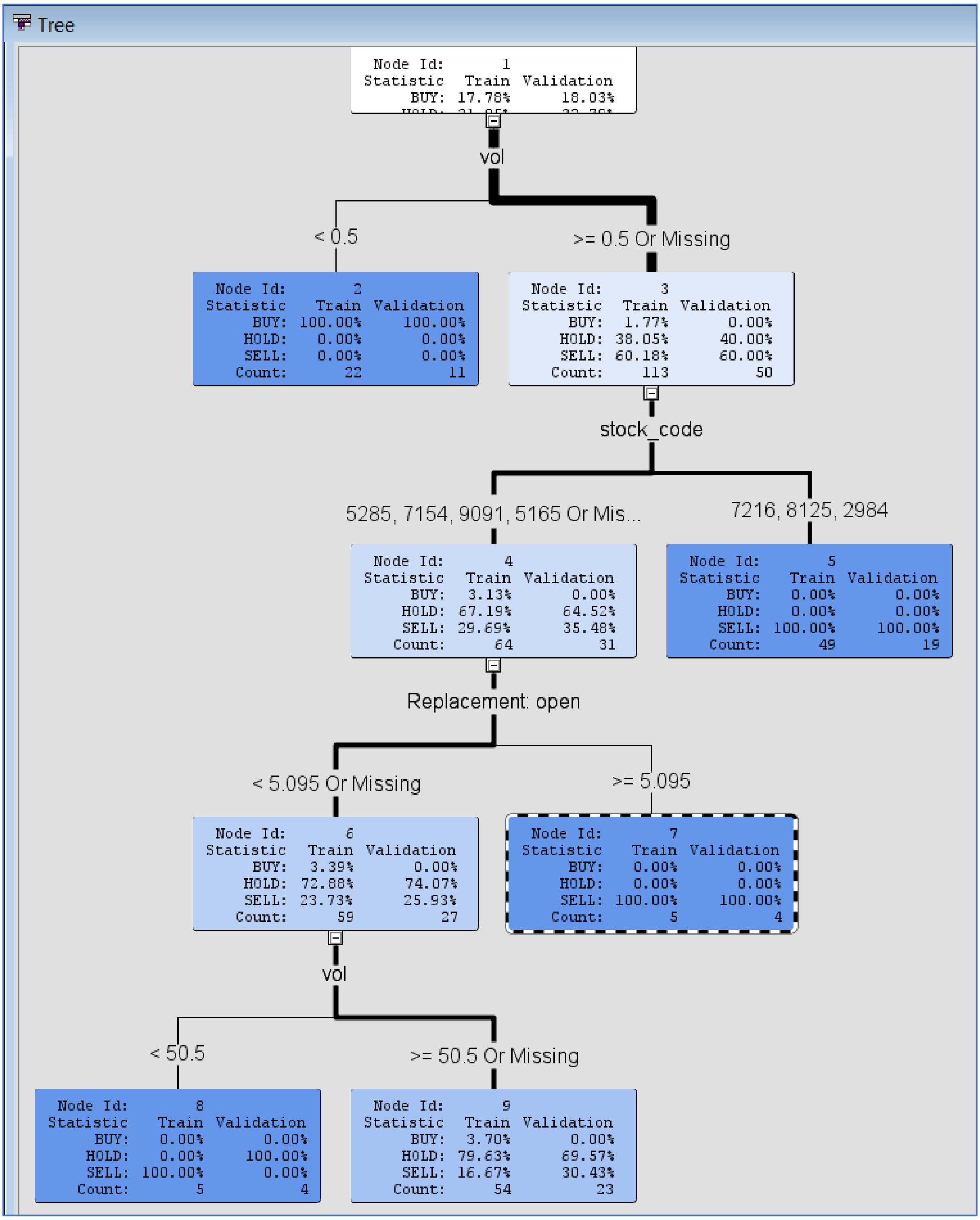


Figure 4: Decision tree constructed

Most of the rules used by the decision tree to split the nodes are dissimilar with the rules defined except one, which is node 7. In Node 7, if the opening price is greater or equal to

5.095 then the trade flag is ‘Sell’, this conform the rules defined for Sime Darby Plantation Bhd.

Assessment plots for the decision tree

Average Square Error

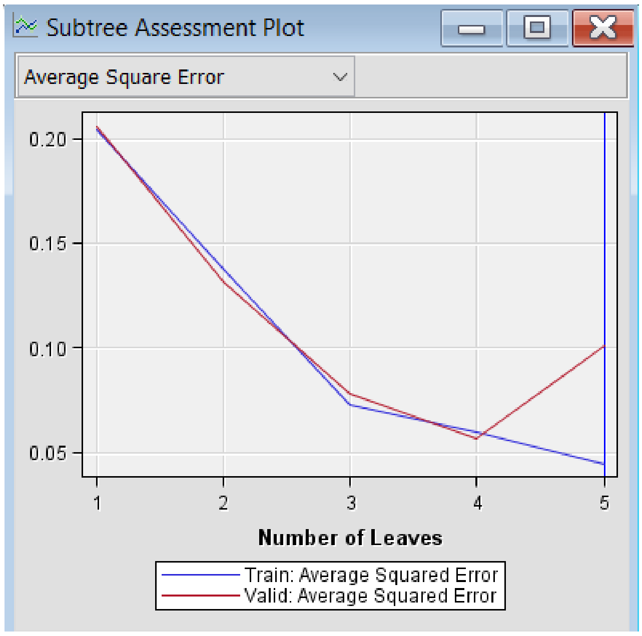


Figure 5: Assessment plot - average square error

The error decreases as the number of leaves increases. The training of the decision tree model might be overfit thus the error in the validation dataset increases after a certain extent.

Misclassification rate

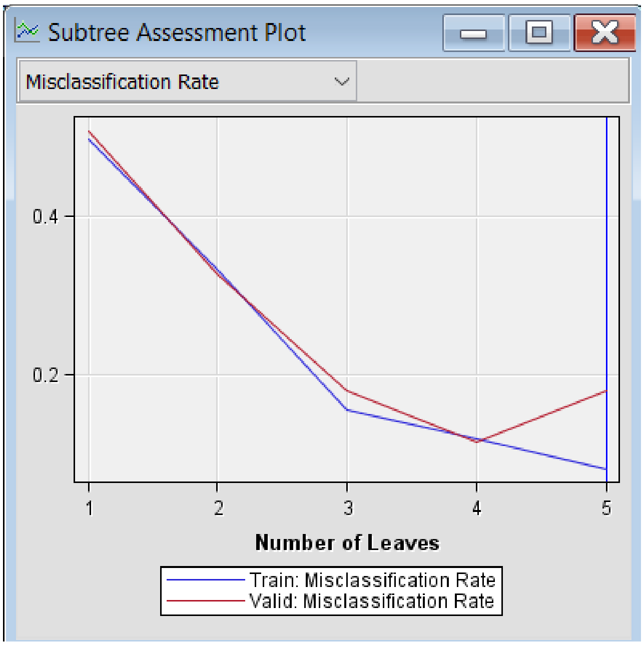


Figure 6: Assessment plot - misclassification rate

Similar with the average square error assessment plot, the misclassification rate decreases as the number of leaves increases. After 5 leaves, the rate spiked.

Logistic Regression

Besides decision tree, logistic regression is also used to predict the trade flag of the stocks. The logistic regression workflow is shown in figure 1.

Results of the logistic regression

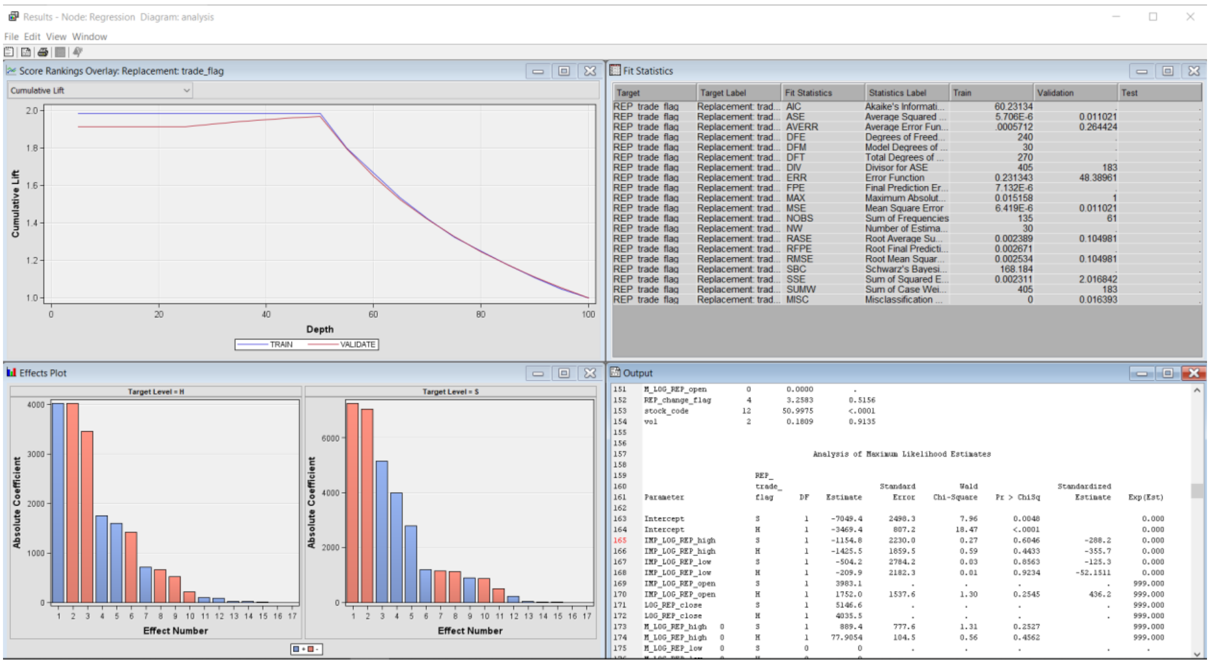


Figure 7: results of logistic regression

The variables used in logistic regression

Variable Summary

Role	Measurement Level	Frequency Count
INPUT	BINARY	3
INPUT	INTERVAL	5
INPUT	NOMINAL	2
REJECTED	INTERVAL	4
REJECTED	NOMINAL	3
TARGET	NOMINAL	1

Figure 8: logistic regression variables summary

10 variables from the dataset are used to predict the target which is the ‘trade flag’. Some of missing values in the variables are imputed and the values underwent a transformation, applying log on the value before it is used to train a logistic regression model.

The model information

```
The DMREG Procedure

                          Model Information

Training Data Set              WORK.EM_DMREG.VIEW
DMDB Catalog                   WORK.REG_DMDB
Target Variable                REP_trade_flag (Replacement: trade_flag)
Target Measurement Level       Nominal
Number of Target Categories    3
Error                          MBernoulli
Link Function                  Logit
Number of Model Parameters     34
Number of Observations         135


          Target Profile

             REP_
 Ordered     trade_        Total
   Value     flag      Frequency

       1     S                68
       2     H                43
       3     B                24
```

Figure 9: logistic regression model information

Fit statistics from logistic regression

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
REP_trade_flag	Replacement: trade_flag	AIC	Akaike's Information Criterion	60.23134	.
REP_trade_flag	Replacement: trade_flag	ASE	Average Squared Error	5.706E-6	0.011021
REP_trade_flag	Replacement: trade_flag	AVERR	Average Error Function	.0005712	0.264424
REP_trade_flag	Replacement: trade_flag	DFE	Degrees of Freedom for Error	240	.
REP_trade_flag	Replacement: trade_flag	DFM	Model Degrees of Freedom	30	.
REP_trade_flag	Replacement: trade_flag	DFT	Total Degrees of Freedom	270	.
REP_trade_flag	Replacement: trade_flag	DIV	Divisor for ASE	405	183
REP_trade_flag	Replacement: trade_flag	ERR	Error Function	0.231343	48.38961
REP_trade_flag	Replacement: trade_flag	FPE	Final Prediction Error	7.132E-6	.
REP_trade_flag	Replacement: trade_flag	MAX	Maximum Absolute Error	0.015158	1
REP_trade_flag	Replacement: trade_flag	MSE	Mean Square Error	6.419E-6	0.011021
REP_trade_flag	Replacement: trade_flag	NOBS	Sum of Frequencies	135	61
REP_trade_flag	Replacement: trade_flag	NW	Number of Estimate Weights	30	.
REP_trade_flag	Replacement: trade_flag	RASE	Root Average Sum of Squares	0.002389	0.104981
REP_trade_flag	Replacement: trade_flag	RFPE	Root Final Prediction Error	0.002671	.
REP_trade_flag	Replacement: trade_flag	RMSE	Root Mean Squared Error	0.002534	0.104981
REP_trade_flag	Replacement: trade_flag	SBC	Schwarz's Bayesian Criterion	168.184	.
REP_trade_flag	Replacement: trade_flag	SSE	Sum of Squared Errors	0.002311	2.016842
REP_trade_flag	Replacement: trade_flag	SUMW	Sum of Case Weights Times Freq	405	183
REP_trade_flag	Replacement: trade_flag	MISC	Misclassification Rate	0	0.016393

Figure 10: fit statistics

Classification table

Classification Table

Data Role=TRAIN Target Variable=REP_trade_flag Target Label=Replacement: trade_flag

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
B	B	100	100	24	17.7778
H	H	100	100	43	31.8519
S	S	100	100	68	50.3704

Data Role=VALIDATE Target Variable=REP_trade_flag Target Label=Replacement: trade_flag

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
B	B	100.000	90.909	10	16.3934
H	H	100.000	100.000	20	32.7869
B	S	3.226	9.091	1	1.6393
S	S	96.774	100.000	30	49.1803

Figure 11: classification table of logistic regression

The classification table show the results of the classification on the validation data. The model achieves 100% accuracy in some scenarios, only 1 mistake when it classifies a target that is supposed to be 'Buy' to 'Sell'.

Logistic regression assessment score

Assessment Score Distribution

Data Role=TRAIN Target Variable=REP_trade_flag Target Label=Replacement: trade_flag

Posterior Probability Range	Number of Events	Number of Nonevents	Mean Posterior Probability	Percentage
0.95-1.00	68	0	0.99951	50.3704
0.00-0.05	0	67	0.00054	49.6296

Data Role=VALIDATE Target Variable=REP_trade_flag Target Label=Replacement: trade_flag

Posterior Probability Range	Number of Events	Number of Nonevents	Mean Posterior Probability	Percentage
0.95-1.00	29	1	0.99916	49.1803
0.90-0.95	1	0	0.92489	1.6393
0.00-0.05	0	30	0.00054	49.1803

Figure 12: logistic regression assessment score

The mean posterior probability is high when the mean posterior probability range is high. This means that the model has a high accuracy rate.

Analysis of maximum likelihood estimates

Analysis of Maximum Likelihood Estimates

Parameter		REP_ trade_ flag	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept		S	1	-7049.4	2498.3	7.96	0.0048		0.000
Intercept		H	1	-3469.4	807.2	18.47	<.0001		0.000
IMP_LOG_REP_high		S	1	-1154.8	2230.0	0.27	0.6046	-288.2	0.000
IMP_LOG_REP_high		H	1	-1425.5	1859.5	0.59	0.4433	-355.7	0.000
IMP_LOG_REP_low		S	1	-504.2	2784.2	0.03	0.8563	-125.3	0.000
IMP_LOG_REP_low		H	1	-209.9	2182.3	0.01	0.9234	-52.1511	0.000
IMP_LOG_REP_open		S	1	3983.1	.	.	.	.	999.000
IMP_LOG_REP_open		H	1	1752.0	1537.6	1.30	0.2545	436.2	999.000
LOG_REP_close		S	1	5146.6	.	.	.	.	999.000
LOG_REP_close		H	1	4035.5	.	.	.	.	999.000
M_LOG_REP_high	0	S	1	889.4	777.6	1.31	0.2527		999.000
M_LOG_REP_high	0	H	1	77.9054	104.5	0.56	0.4562		999.000
M_LOG_REP_low	0	S	0	0	.	.	.	.	.
M_LOG_REP_low	0	H	0	0	.	.	.	.	.
M_LOG_REP_open	0	S	0	0	.	.	.	.	.
M_LOG_REP_open	0	H	0	0	.	.	.	.	.
REP_change_flag	N	S	1	26.5637	36.3284	0.53	0.4646		999.000
REP_change_flag	N	H	1	23.0783	34.3030	0.45	0.5011		999.000
REP_change_flag	O	S	1	18.7851	29.8713	0.40	0.5294		999.000
REP_change_flag	O	H	1	16.2146	28.9787	0.31	0.5758		999.000
stock_code	2984	S	1	228.4	149.0	2.35	0.1253		999.000
stock_code	2984	H	1	101.4	135.7	0.56	0.4547		999.000
stock_code	5165	S	1	2790.8	849.3	10.80	0.0010		999.000
stock_code	5165	H	1	1602.4	336.2	22.71	<.0001		999.000
stock_code	5285	S	1	-7270.4	2030.4	12.82	0.0003		0.000
stock_code	5285	H	1	-4033.5	868.4	21.57	<.0001		0.000
stock_code	7154	S	1	1203.7	345.2	12.16	0.0005		999.000
stock_code	7154	H	1	714.5	166.1	18.51	<.0001		999.000
stock_code	7216	S	1	-1122.4	545.2	4.24	0.0395		0.000
stock_code	7216	H	1	-662.0	476.9	1.93	0.1651		0.000
stock_code	8125	S	1	-876.0	279.9	9.80	0.0017		0.000
stock_code	8125	H	1	-518.3	184.1	7.92	0.0049		0.000
vol		S	1	-0.00050	0.00301	0.03	0.8682	-3.0409	1.000
vol		H	1	-0.00059	0.00299	0.04	0.8424	-3.6132	0.999

Figure 13: analysis of maximum likelihood

In this maximum likelihood estimates, the Pr > ChiSq column show the significance of the variables. If the value is closer to 0, then the variable has more significance in determining the outcome. If the value is closer to 1, then it means the variable is not suitable to be used for prediction.

It is seen that stock codes are an important feature to predict the trade flag.

Classification chart

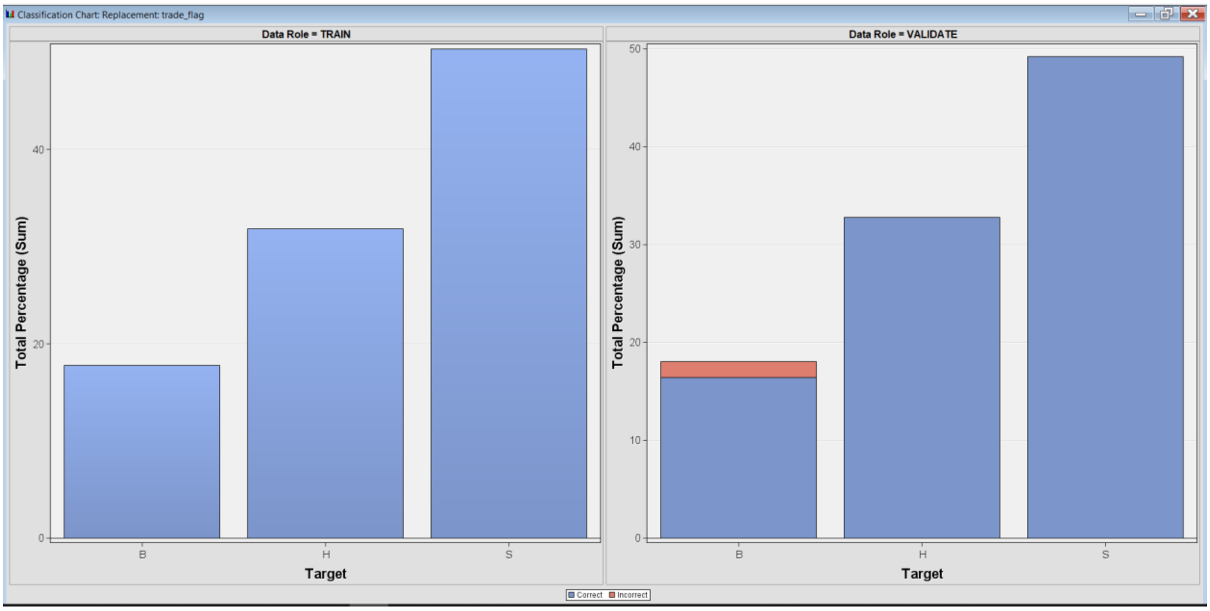


Figure 14: classification chart

The classification chart conveys the same information as the classification table. The model achieved a high accuracy rate on the validation dataset, only 1 mistake. The number of categories are different in the training set and the validation set is because they are consisted of different records.

Rules obtained from the decision tree

1. If the volume is < 0.5 then buy the stock.
2. If the volume is >= 0.5 and the stock code is 7216, 8125, 2984 then sell the stock.
3. If the volume is < 0.5 and the stock code is 5285, 7154, 9091, 5165 and the opening price is >= 5.095 then sell the stock.
4. If the volume is < 0.5 and the stock code is 5285, 7154, 9091, 5165 and the opening price is < 5.095 and the volume is < 50.5 then sell the stock

Conclusion

The third rule obtained from the decision tree is similar with the rules defined for Sime Darby Plantation Berhad to sell the stock if the price is greater than or equal to 5.10.

The other rules obtained from the decision tree are unexpected and could be the hidden insights. These rules would be used to predict the trade flag of the stocks in the future, to check if the features can truly be depended upon to determine the trade flag.