SEMESTER II, 2018/2019 SESSION

WQD7005

DATA MINING

PREPARED BY

PARVITRA SOMAHSUNDRAM WQD180045

LECTURER **:** Prof. Dr. The Ying Wah

SUBMISSION DATE **:** 2ND June 2019

Table of Contents

1.0 Data Acquisition

The stock data is crawled from https://www.thestar.com.my/business/marketwatch/stocks/?qcounter= to retrieve all the stock price of all the company from A-Z with the following attributes: -

- Company Name
- Company Code
- Open Price
- High Price
- Low Price
- Last Price
- Change Price
- Volume
- Buy rate
- Selling Rate

The following is the dataset which crawled and export in CSV file without organised date (no attributed fields).

	A	B	C
1	UCHI TECHNOLOGIES BHD	7100	['2.720', '2.730', '2.700', '2.720', '0.000', '0.00', '1,592', '2.710 / 400', '2.720 / 3,543']
2			
3	UCHITEC-CD: CW UCHI TECHNOLOGIES BERHAD (AM)	7100CD	['-', '-', '-', '0.005', '0.000', '0.00', '0', '0.000 / 0', '0.000 / 0']
4			
5	UCHITEC-CE: CW UCHI TECHNOLOGIES BERHAD (RHB)	7100CE	['-', '-', '-', '0.025', '0.000', '0.00', '0', '0.000 / 0', '0.000 / 0']
6			
7	UCREST BERHAD	5	['0.255', '0.270', '0.255', '0.260', '0.005', '1.96', '105,198', '0.255 / 29,657', '0.260 / 6,115']
8			
9	UEM SUNRISE BERHAD	5148	['0.850', '0.860', '0.835', '0.840', '-0.010', '-1.18', '10,411', '0.835 / 416', '0.840 / 701']
10			
11	UEMS-C64: CW UEM SUNRISE BERHAD (MIBB)	514864	['-', '-', '-', '0.040', '0.000', '0.00', '0', '0.000 / 0', '0.000 / 0']
12			
13	UEMS-C65: CW UEM SUNRISE BERHAD (MACQ)	514865	['0.085', '0.085', '0.085', '0.085', '0.000', '0.00', '732', '0.000 / 0', '0.000 / 0']
14			
15	UEMS-C66: CW UEM SUNRISE BERHAD (KIBB)	514866	['0.070', '0.070', '0.070', '0.070', '0.000', '0.00', '974', '0.000 / 0', '0.000 / 0']
16			
17	UNITED U-LI CORPORATION BHD	7133	['0.440', '0.455', '0.435', '0.450', '0.010', '2.27', '8,102', '0.435 / 525', '0.450 / 1,849']
18			
19	UNITED MALACCA BHD	2593	['5.550', '5.550', '5.450', '5.500', '-0.010', '-0.18', '591', '5.500 / 93', '5.560 / 20']

2.0 Management of data

To store the data crawled in 1.0 section; we had managed the data to SQL database. In section 1.0; we had crawled the data and save in CSV format without having respective attributes.

- Create database with attributes

```sql
use StockCrawl;
create table StockData (
ID int Primary Key auto_increment,
CompanyName varchar(50) Not NULL,
CompanyStockCode varchar(30),
OpenPrice varchar(30),
HighPrice varchar(30),
LowPrice varchar (30),
LastPrice varchar(30),
Chg varchar(30),
ChgPercentage varchar(30),
Volume varchar(30),
BuyVolume varchar(30),
SellVolume varchar(30),
CreatedAt Timestamp default current_timestamp
);
```

- Export the crawl data to the database

ID	CompanyName	CompanyStockCode	OpenPrice	HighPrice	LowPrice	LastPrice	Chg	ChgPercentage	Volume	BuyVolume	SellVolume	CreatedAt
4	ZECON BHD	7028	0.285	0.295	0.275	0.285	0.000	0.00	14,154	0.280 / 1,180	0.285 / 96	2019-03-21 10:24:23
5	ZELAN BHD	2283	0.090	0.095	0.090	0.095	0.005	5.56	28,483	0.090 / 8,317	0.095 / 24,838	2019-03-21 10:24:23
6	ZHULIAN CORPORATION BHD	5131	1.360	1.360	1.360	1.360	-0.010	-0.73	49	1.360 / 31	1.370 / 100	2019-03-21 10:24:23
7	ZECON BHD	7028	0.285	0.295	0.275	0.285	0.000	0.00	14,154	0.280 / 1,180	0.285 / 96	2019-03-21 10:35:50
8	ZELAN BHD	2283	0.090	0.095	0.090	0.095	0.005	5.56	28,483	0.090 / 8,317	0.095 / 24,838	2019-03-21 10:35:50
9	ZHULIAN CORPORATION BHD	5131	1.360	1.360	1.360	1.360	-0.010	-0.73	49	1.360 / 31	1.370 / 100	2019-03-21 10:35:50
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL

2.1 Star Schema

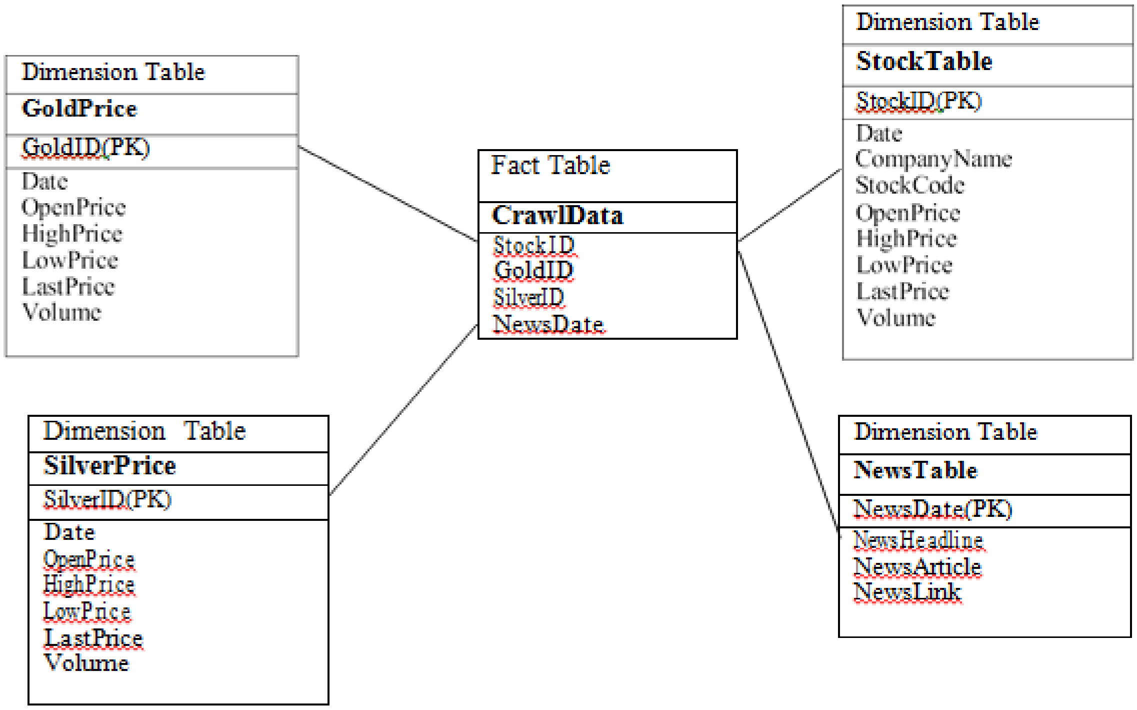


Figure 1 Star Schema

The importance of dimension is to allow us to drill up and drill down on the data.

The ID attributes contain the date with the time information, this unique key the table. This attributes and the date attributes will be differ in term of time; which used to merge with three table data and to crawl news data with the ID.

Through StockTable dimension, we can get the information of all the stock company's stock price in varies of their open price, low price, high price, last price and volume of the stock of the particular date.; with this dimension, we can compare the stock price day to day, month to month, year to year and able to get a pattern from it.

NewsTable dimension able to provide an insight on whether the particular company has good or bad news by daily basic and monthly basic; this will give an overview whether news will impact the price of the stock.

On the SilverPrice dimension, we can retrieve the silver price details in varies of their open price, low price, high price, last price and volume of the silver of the particular date; this apply as well to the GoldPrice dimension.

3.0 Processing of data

In this milestone, we had crawled news data to shows the dependency of the stock price with the news.

News Data Crawl with attributes of stock date, company name, news header, news info and news link from https://www.theedgemarkets.com .

	StockDate	CompanyName	NewsHeader	NewsInfo	NewsLink
0	April 30, 2019 17:00 pm +08	HLBANK	Are banks worth a second look as half of them ...	HALF of the banks listed on Bursa Malaysia hav...	https://www.theedgemarkets.com/article/are-ban...
1	April 30, 2019 09:10 am +08	HLBANK	KLCI drifts lower in line with region, key Chi...	KUALA LUMPUR (April 30): The FBM KLCI drifted ...	https://www.theedgemarkets.com/article/klci-dr...
2	April 26, 2019 10:18 am +08	HLBANK	KLCI retreats in line with subdued regional ma...	KUALA LUMPUR (April 26): The FBM KLCI retreate...	https://www.theedgemarkets.com/article/klci-re...
3	April 18, 2019 13:10 pm +08	HLBANK	KLCI falls 0.59% as Public Bank, Maxis weigh	KUALA LUMPUR (April 18): The FBM KLCI fell 0.5...	https://www.theedgemarkets.com/article/klci-fa...
4	April 18, 2019 10:21 am +08	HLBANK	KLCI stays in negative zone on worries of pote...	KUALA LUMPUR (April 18): The FBM KLCI remained...	https://www.theedgemarkets.com/article/klci-st...

Covariance Software used: Jupiter

The following ate the covariance between stock price and news polarity for Mah Sing and HL Bank.

For HL Bank it shows positive result as the news had impact the stock price whereas for Mah Sing the stock price had no dependency with news it shows negative covariance.

In [155]:
```
cov_mat = HLBANK.cov()
cov_mat
```
Out[155]:

	stockprice	newspolarity
stockprice	0.022323	0.008053
newspolarity	0.008053	0.040963

In [156]:
```
corr_mat= HLBANK.corr()
corr_mat
```
Out[156]:

	stockprice	newspolarity
stockprice	1.00000	0.26632
newspolarity	0.26632	1.00000

Figure 2 HL Bank Covariance

In [72]:
```
cov_mat = MAHSING.cov()
cov_mat
```
Out[72]:

	stockprice	newspolarity
stockprice	1.179654e-04	-2.934750e-34
newspolarity	-2.934750e-34	3.228225e-33

In [74]:
```
corr_mat= MAHSING.corr()
corr_mat
```
Out[74]:

	stockprice	newspolarity
stockprice	1.000000e+00	-4.755673e-16
newspolarity	-4.755673e-16	1.000000e+00

Figure 3 Mah Sing Covariance

4.0 Interpretation of data

The crawl data been interpreted to visualise as Heat Map and Pair Plot for visualization by using stock price and news polarity variable.

4.1 HeatMap

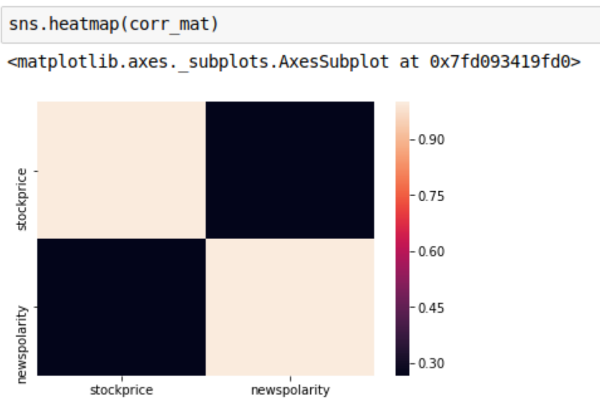


Figure 4 HeatMap of HL Bank

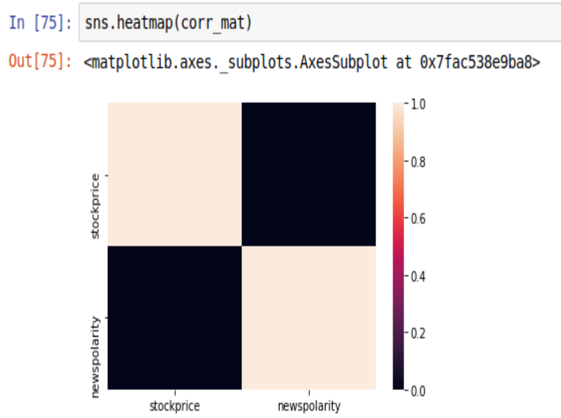


Figure 5 HeatMap of Mah Sing

The lighter the colour the stronger the relationship between the variables and the darker the colour in heat map it show there is no relationship between the variable.

4.2 Pair Plot

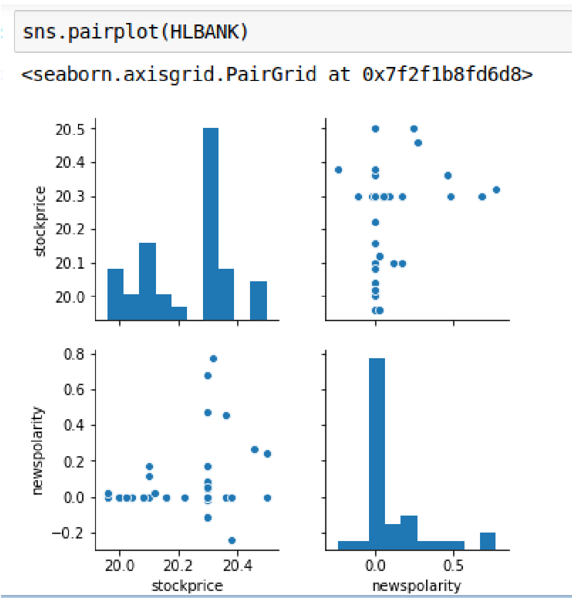


Figure 6 PairPlot of HL Bank

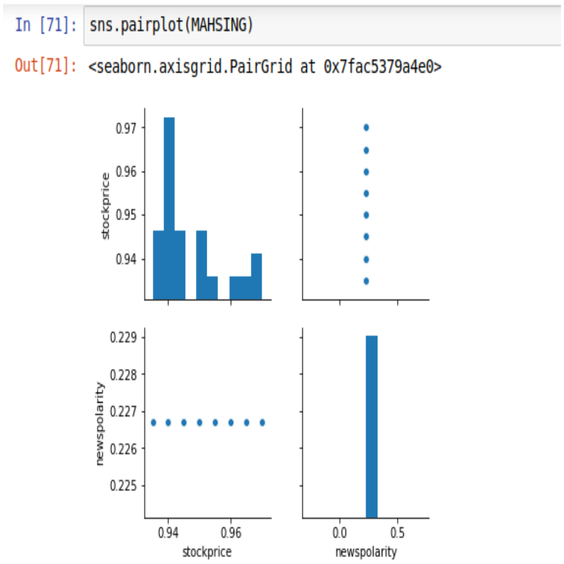


Figure 7 PairPlot of Mah Sing

Mah Sing pair plot shows negative correlation which mean there is no dependency on news polarity with the Mah sing stock price where Hong Leong bank shows positive correlation which mean the news gives impact to the stock price of the company.

5.0 Communication of insights of data

Decision Tree and Regression Model has plotted to visualise the data to view:-

5.1 Decision tree

Creating Training and Validation Data

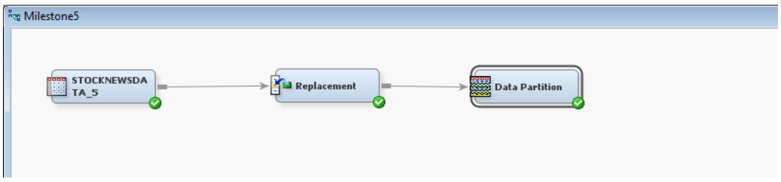
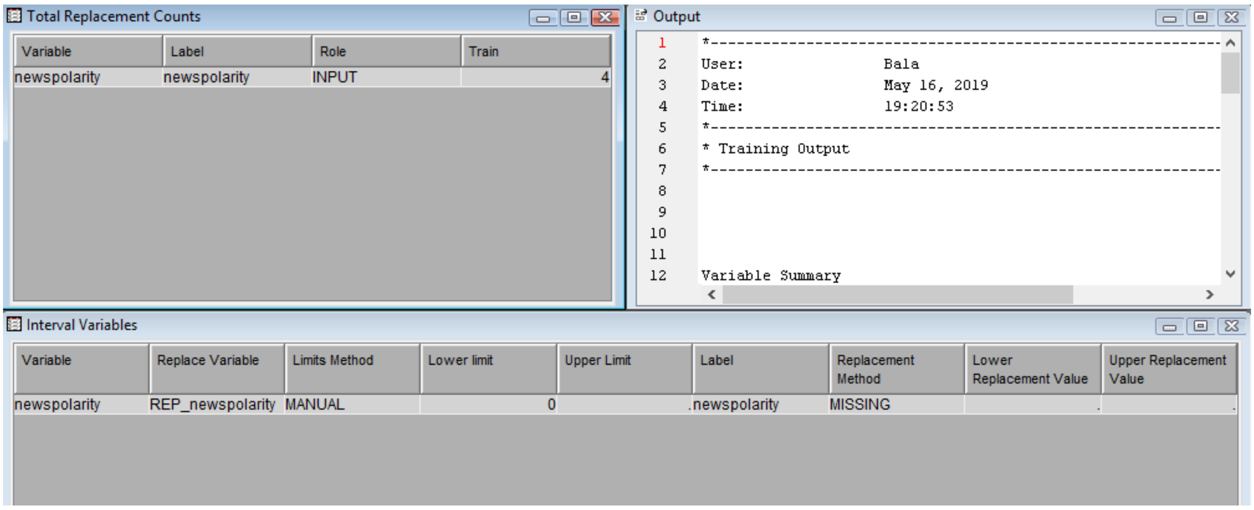


Figure 8 Data Diagram of Decision Tree

Replacement

Any newspolarity value that fall below the lower limit of 0 are set to missing .
All other values of this variable will not impacted.
*It will list all the less than 0 (negative value) as missing value.
Total 4 missing value.



Variable	Label	Role	Train
newspolarity	newspolarity	INPUT	4

Variable	Replace Variable	Limits Method	Lower limit	Upper Limit	Label	Replacement Method	Lower Replacement Value	Upper Replacement Value
newspolarity	REP_newspolarity	MANUAL	0	.	newspolarity	MISSING	.	.

Figure 9 Output of Replacement Execution

A new column is added to the analysis data: Replacement: newspolarity. The replace value is shown by a dot(.) which indicated a missing value.

Obs #	Date	companyname	stockprice	newspolarity	Replacement: newspolarity
24	03/18/2019	HLBANK	20.3	0.057218337	0.057218
25	03/19/2019	HLBANK	20.3	0.051419984	0.05142
26	03/20/2019	HLBANK	20.3	0.045454545	0.045455
27	03/21/2019	HLBANK	20.3	0.171373457	0.171373
28	03/22/2019	HLBANK	20.3	0.170526094	0.170526
29	03/25/2019	HLBANK	20.3	-0.018025078	.
30	03/26/2019	HLBANK	20.3	0.084722222	0.084722
31	03/27/2019	HLBANK	20.3	0.084313603	0.084314
32	03/28/2019	HLBANK	20.3	0	0
33	03/29/2019	HLBANK	20.3	-0.110567841	.
34	04/01/2019	HLBANK	20.3	0.474296947	0.474297
35	04/01/2019	HLBANK	20.3	0.474296947	0.474297
36	04/02/2019	HLBANK	20.3	0.265410084	0.26541
37	04/02/2019	HLBANK	20.46	0.265410084	0.26541
38	04/03/2019	HLBANK	20.46	0.459323938	0.459324
39	04/03/2019	HLBANK	20.36	0.459323938	0.459324
40	04/04/2019	HLBANK	20.36	0.055348485	0.055348
41	04/04/2019	HLBANK	20.3	0.055348485	0.055348
42	04/05/2019	HLBANK	20.3	0	0
43	04/07/2019	HLBANK	20.22	0	0
44	04/08/2019	HLBANK	20.22	0.771217439	0.771217
45	04/08/2019	HLBANK	20.32	0.771217439	0.771217
46	04/09/2019	HLBANK	20.32	0.173360528	0.173361
47	04/09/2019	HLBANK	20.1	0.173360528	0.173361
48	04/10/2019	HLBANK	20.1	0	0
49	04/10/2019	HLBANK	20.08	0	0
50	04/11/2019	HLBANK	20.08	0	0
51	04/11/2019	HLBANK	20.16	0	0
52	04/12/2019	HLBANK	20.16	0	0
53	04/14/2019	HLBANK	20.36	0	0
54	04/15/2019	HLBANK	20.36	0.113318312	0.113318
55	04/15/2019	HLBANK	20.1	0.113318312	0.113318
56	04/16/2019	HLBANK	20.1	0	0
57	04/16/2019	HLBANK	20	0	0
58	04/17/2019	HLBANK	20	0	0
59	04/17/2019	HLBANK	20	0	0
60	04/18/2019	HLBANK	20	-0.241016002	.
61	04/18/2019	HLBANK	20.38	-0.241016002	.

Figure 10 Data Set of Replaced Missing Value

Data Partition

With smaller raw datasets, model stability can become an important issue.
Thus, increasing the number of cases devoted to the training partition is a reasonable course of action.

Property	Value
General	
Node ID	Part
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	50.0
Validation	50.0
Test	0.0
Report	

Figure 11 Dataset Allocation Setting

```
Variable Summary

            Measurement    Frequency
Role           Level         Count

INPUT         INTERVAL         2
INPUT         NOMINAL          1
REJECTED      INTERVAL         1
TIMEID        INTERVAL         1



Partition Summary

                                         Number of
Type              Data Set              Observations

DATA        EMWS1.Repl_TRAIN                76
TRAIN       EMWS1.Part_TRAIN                38
VALIDATE    EMWS1.Part_VALIDATE             38
```

Figure 12 Summary of Data Partition

Constructing Decision Tree

- Created the maximal tree

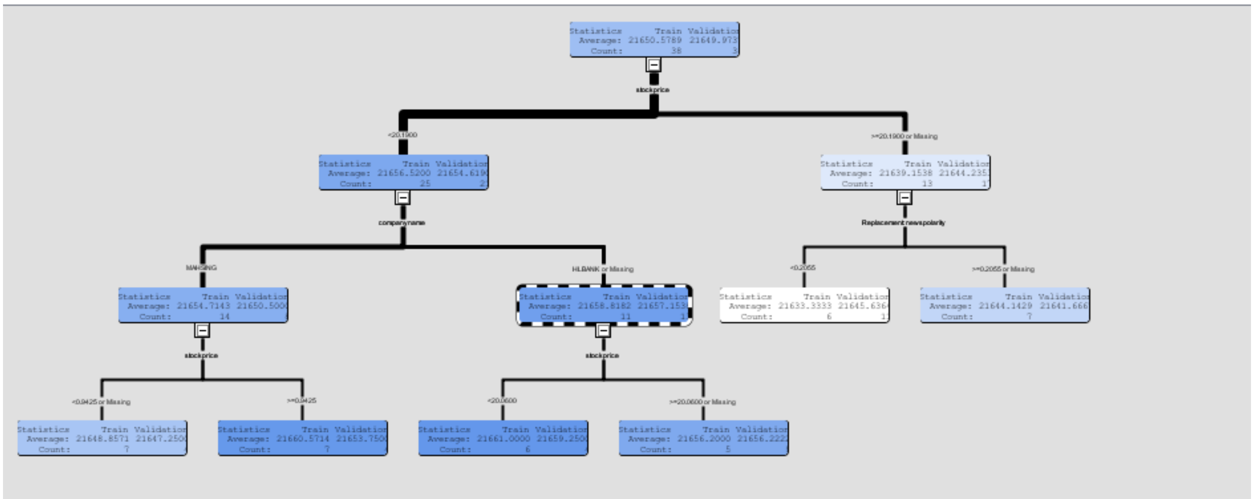


Figure 13 Maximal Tree

It shows the predictive model assigns one of 6 predicated targets for train and validation data.

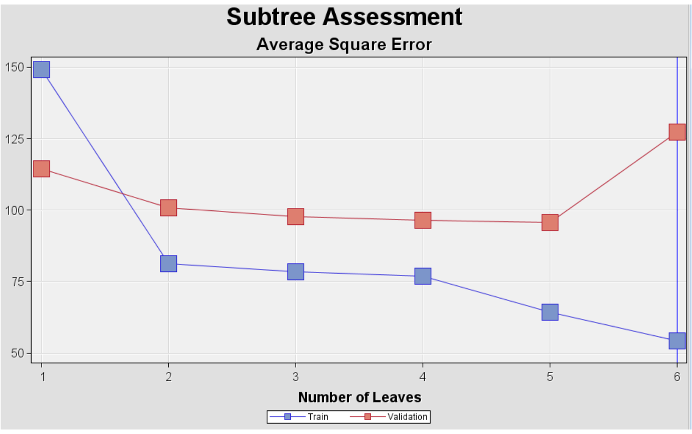


Figure 14 Subtree Assessment

Assessing a Decision Tree

*Frozen tree property prevents the maximal tree from being changed by other property settings when the flow is run.

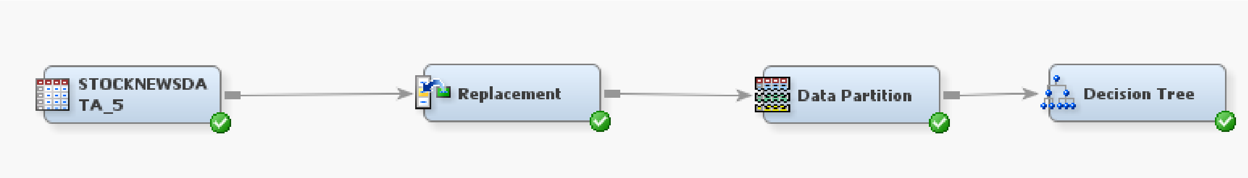


Figure 15 Data Node of Decision tree

The result window contains a variety of diagnostic plots and tables, including a mean predicted chart, a tree map, and a table of fit statistics. The diagnostic tools shown in the results vary with the measurement level of the target variable.

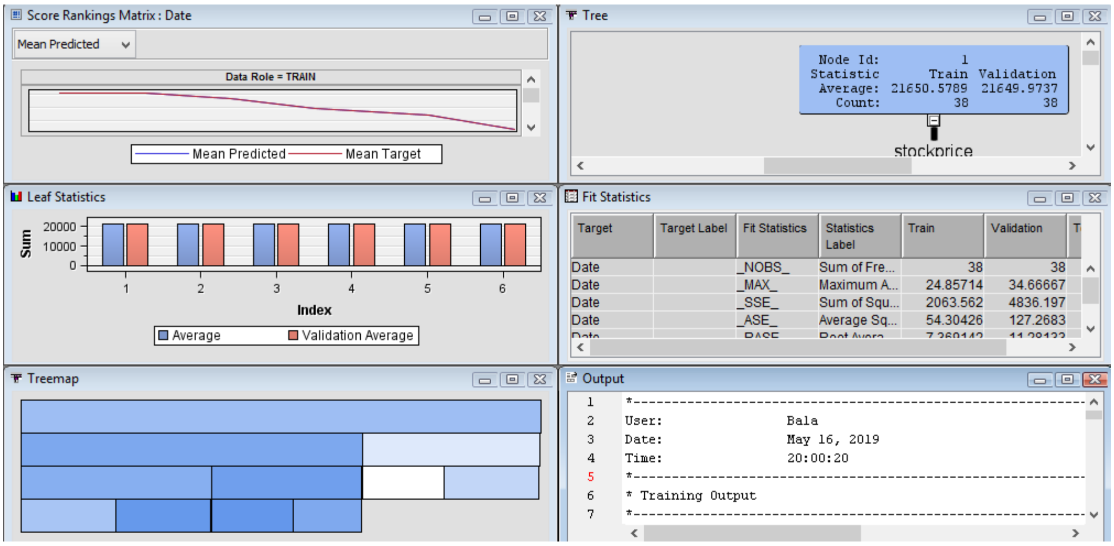


Figure 16 Diagnostic plots and table

Figure 17 Plot of Average Square Error

This plot shows the Average Square Error corresponding to each sub tree as the data is sequentially split. It is similar to the one generated with the Interactive Decision Tree tool, and it confirms suspicions about the optimality of 6 leaf tree. The performance on the training sample becomes monotonically better as the tree becomes more complex. However, the performance on the validation sample only improves up to a tree of, approximately, four or five leaves and the diminishes as model complexity increases.

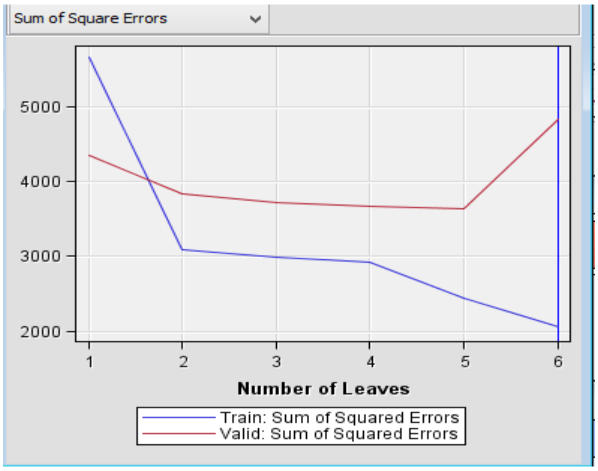


Figure 18 Plot of Sum of Square Error

This plot is similar to the performance under Average square error. But this will be inaccurate for all cases that are not in the assigned class.

To evaluate the number of leaves between maximal tree , decision tree, probability tree

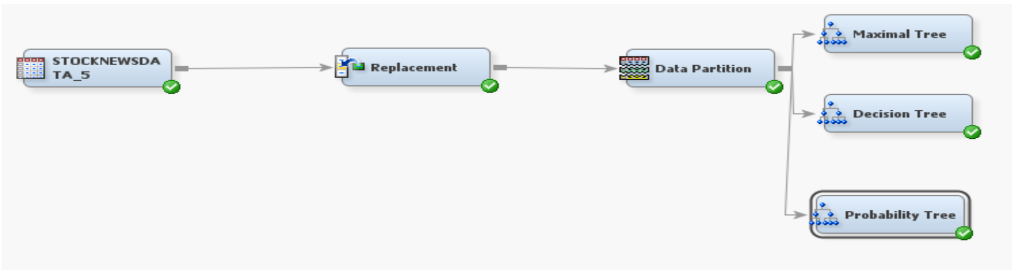


Figure 19 Data node

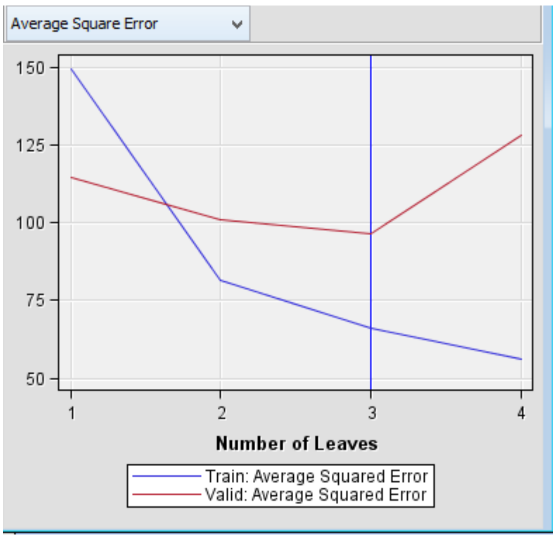


Figure 20 Plot of Average Square Error

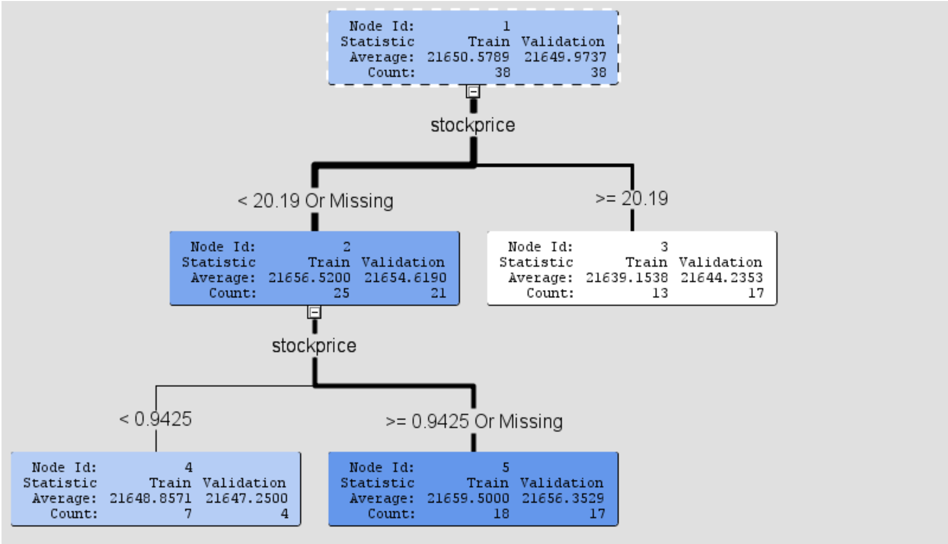


Figure 21 Probability Tree Diagram

The output for probability tree and decision tree give 4 leaves.

5.2 Linear Regression

<u>Managing Missing Values</u>

There are several inputs with a noticeable frequency of missing values, for example , news polarity.

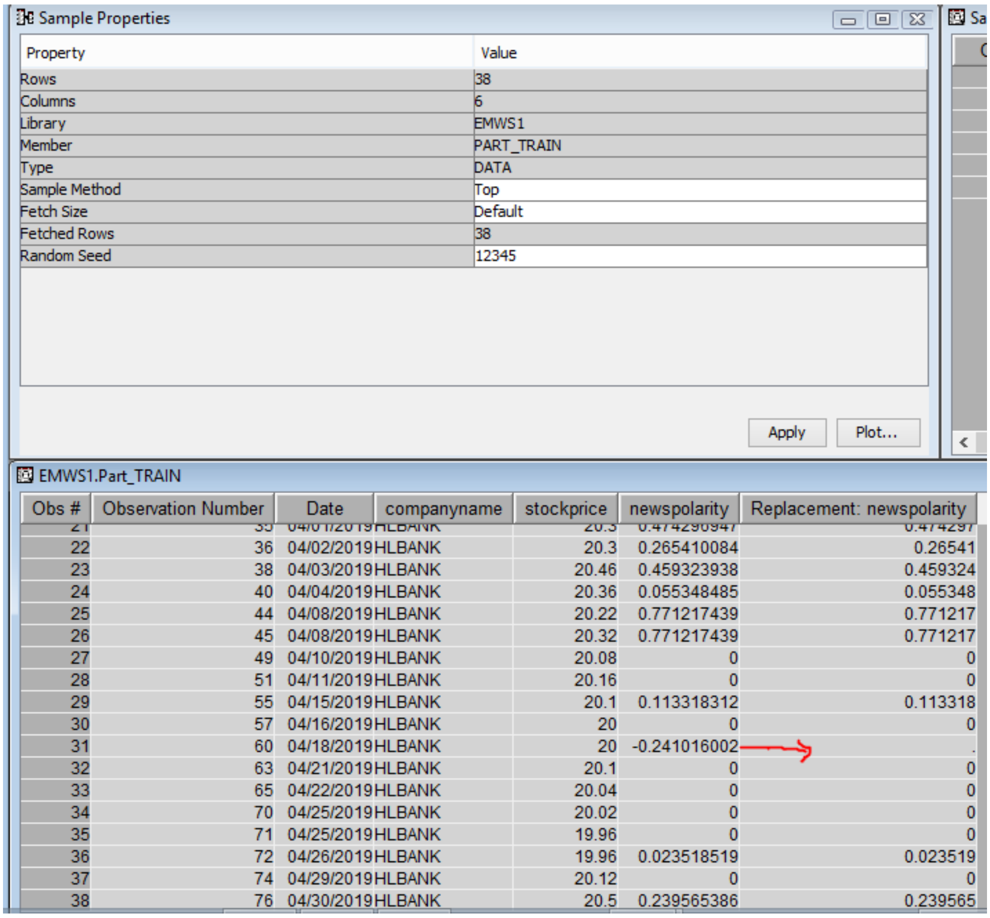


Figure 22 Missing Value

1 input has missing value:-

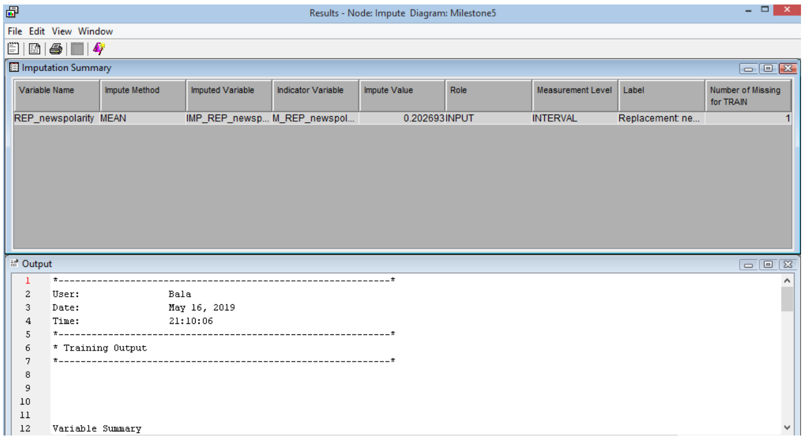


Figure 23 Imputation Result

Running the regression node

Regression Diagram below:-

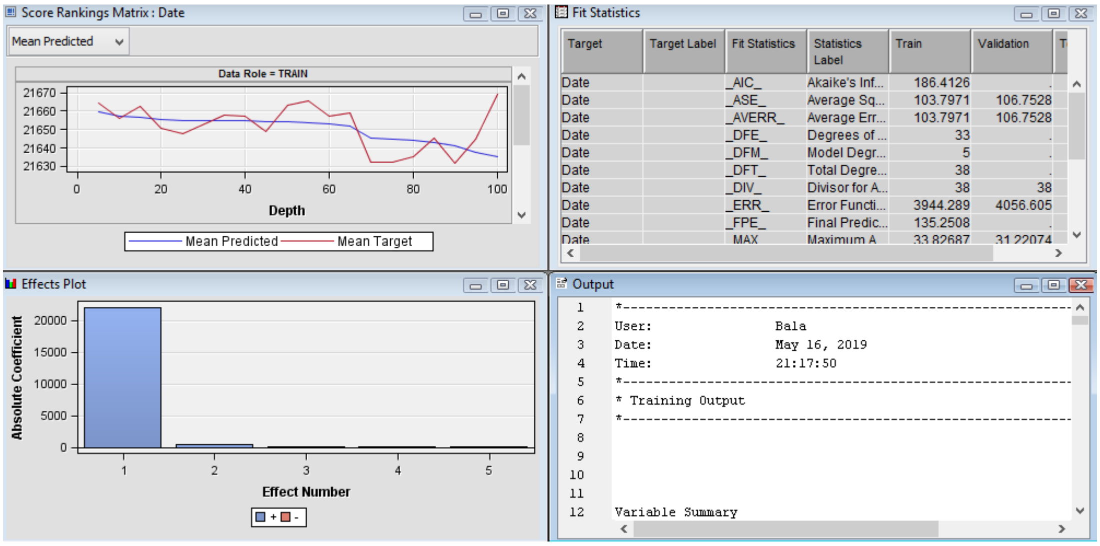


Figure 24 Regression Diagram

The initial lines of the output window summarize the roles of variables used (or not) by the Regression node. The fit model has 4 inputs that predict a binary target.

```
Variable Summary

             Measurement     Frequency
Role            Level          Count

INPUT         BINARY             1
INPUT         INTERVAL           2
INPUT         NOMINAL            1
REJECTED      INTERVAL           1
TARGET        INTERVAL           1
```

Figure 25 Summary of Variables

The next lines give more information about the model, including the training data set name, target variable name, number of target categories and most importantly, the number of model parameter.

```
              Model Information

Training Data Set              WORK.EM_DMREG.VIEW
DMDB Catalog                   WORK.REG_DMDB
Target Variable                Date
Target Measurement Level       Interval
Error                          Normal
Link Function                  Identity
Number of Model Parameters     5
Number of Observations         38
```

Figure 26 Summary of Model Information

The type 3 analysis tests the statistical significance of adding the indicated input to a model that already contains other listed inputs. A value near 0 in Pr > F column approximately indicates a significant input; a value near 1 indicates an extraneous input.

Type 3 Analysis of Effects

Effect	DF	Sum of Squares	F Value	Pr > F
IMP_REP_newspolarity	1	50.3219	0.42	0.5209
M_REP_newspolarity	1	0.2420	0.00	0.9644
companyname	1	698.5254	5.84	0.0213
stockprice	1	712.0180	5.96	0.0202

Figure 27 Summary of Type 3 Analysis of Effects

The statistical shows the significant measure a range from <0.0001(highly significant) to 0.9593(highly dubious). Results such as this suggest that certain inputs can be dropped without affecting the predictive prowess of model.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Date		_AIC_	Akaike's Inf...	186.4126	.	.
Date		_ASE_	Average Sq...	103.7971	106.7528	.
Date		_AVERR_	Average Err...	103.7971	106.7528	.
Date		_DFE_	Degrees of ...	33	.	.
Date		_DFM_	Model Degr...	5	.	.
Date		_DFT_	Total Degre...	38	.	.
Date		_DIV_	Divisor for A...	38	38	.
Date		_ERR_	Error Functi...	3944.289	4056.605	.
Date		_FPE_	Final Predic...	135.2508	.	.
Date		_MAX_	Maximum A...	33.82687	31.22074	.
Date		_MSE_	Mean Squa...	119.5239	106.7528	.
Date		_NOBS_	Sum of Fre...	38	38	.
Date		_NW_	Number of ...	5	.	.
Date		_RASE_	Root Avera...	10.18809	10.33212	.
Date		_RFPE_	Root Final ...	11.62974	.	.
Date		_RMSE_	Root Mean ...	10.9327	10.33212	.
Date		_SBC_	Schwarz's ...	194.6006	.	.
Date		_SSE_	Sum of Squ...	3944.289	4056.605	.
Date		_SUMW_	Sum of Cas...	38	38	.

Figure 28 Fit Statistics

Selecting Input

The stepwise procedure starts with Step 0, an intercept-only regression model. The value of the intercept parameter is chosen so that the model predicts the overall target mean for every case.

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Date		_AIC_	Akaike's Inf...	192.1853	.	.
Date		_ASE_	Average Sq...	149.1385	114.2867	.
Date		_AVERR_	Average Err...	149.1385	114.2867	.
Date		_DFE_	Degrees of ...	37	.	.
Date		_DFM_	Model Degr...	1	.	.
Date		_DFT_	Total Degre...	38	.	.
Date		_DIV_	Divisor for A...	38	38	.
Date		_ERR_	Error Functi...	5667.263	4342.895	.
Date		_FPE_	Final Predic...	157.2	.	.
Date		_MAX_	Maximum A...	27.57895	24.57895	.
Date		_MSE_	Mean Squa...	153.1693	114.2867	.
Date		_NOBS_	Sum of Fre...	38	38	.
Date		_NW_	Number of ...	1	.	.
Date		_RASE_	Root Avera...	12.21223	10.6905	.
Date		_RFPE_	Root Final ...	12.53794	.	.
Date		_RMSE_	Root Mean ...	12.37616	10.6905	.
Date		_SBC_	Schwarz's ...	193.8229	.	.
Date		_SSE_	Sum of Squ...	5667.263	4342.895	.
Date		_SUMW_	Sum of Cas...	38	38	.

Figure 29 Fit Statistics

Stepwise Selection Procedure

Step 0: Intercept entered.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	0	0	.	.	.
Error	37	5667.263159	153.169275		
Corrected Total	37	5667.263159			

Model Fit Statistics

R-Square	0.0000	Adj R-Sq	0.0000
AIC	192.1853	BIC	193.7093
SBC	193.8229	C(p)	11.4153

Figure 30 Stepwise Selection

Optimizing Regression Complexity

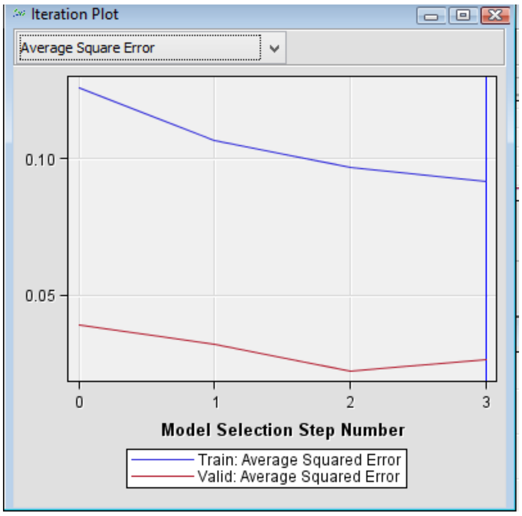


Figure 31 Average Sqaure Error Plot

The iteration plot shows the smallest validation average square error occurs at step 2. The vertical blue line shows the model with the optimal validation error (step 3).

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	0	0	.	.	.
Error	37	5667.263159	153.169275		
Corrected Total	37	5667.263159			

Model Fit Statistics

R-Square	0.0000	Adj R-Sq	0.0000
AIC	192.1853	BIC	193.7093
SBC	193.8229	C(p)	11.4153

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	21650.6	2.0077	10783.9	<.0001

Figure 32 Analysis of Variable

Transforming Inputs

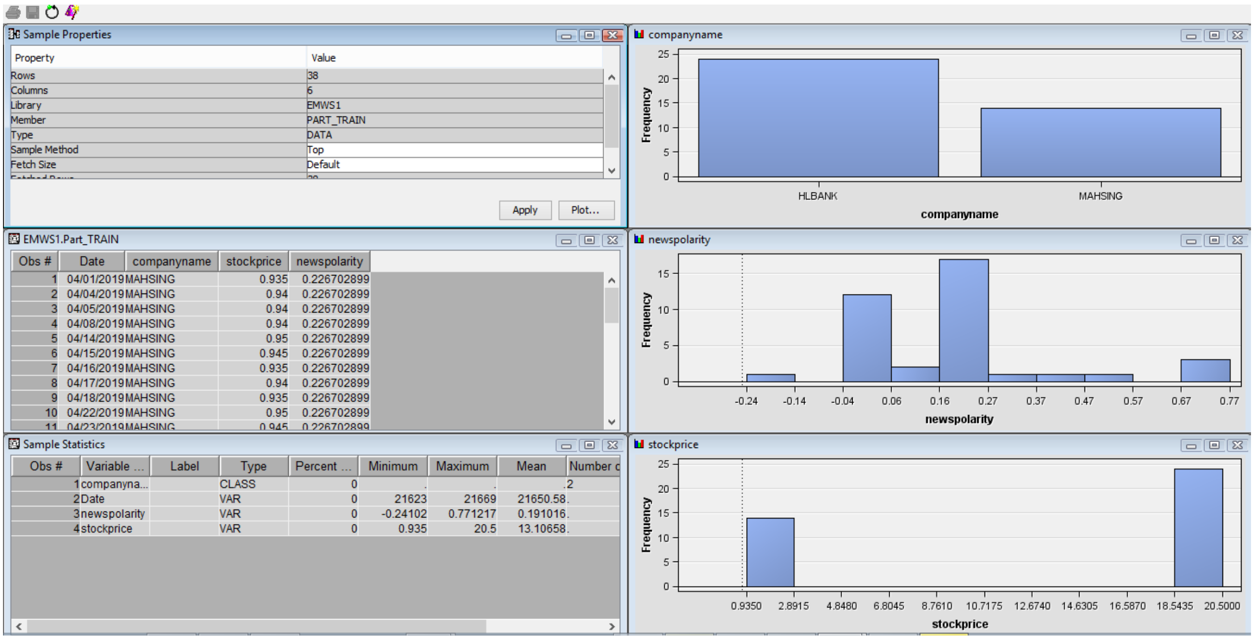


Figure 33 Plot of Transformed Inputs

Selecting log method for following variables:-

Name	Method	Number of Bins	Role	Level
Date	**Default**	**4**	Target	Interval
REP_newspolarit	**Default**	**4**	Input	Interval
companyname	**Default**	**4**	Input	Nominal
newspolarity	**Log**	**4**	Rejected	Interval
stockprice	**Log**	**4**	Input	Interval

Figure 34 Variable Roles

```
Computed Transformations
(maximum 500 observations printed)

                          Input
 Input Name     Role      Level          Name          Level      Formula

newspolarity    REJECTED  INTERVAL  LOG_newspolarity   INTERVAL   log(newspolarity  + 1.241016002)
stockprice      INPUT     INTERVAL  LOG_stockprice     INTERVAL   log(stockprice  + 1)
```

Figure 35 Summary of Computed Transformation

Notice the formula column, while a log transformation was specified, the actual transformation used was log (input+1). This default action of the transform variable tools avoids problems with 0-values of the underlying inputs.

Categorical Input

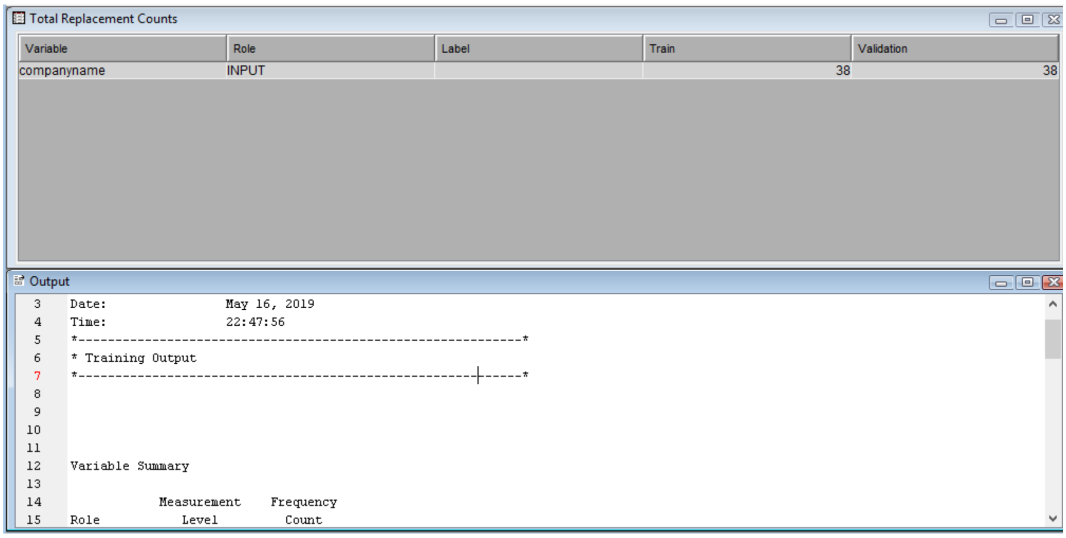
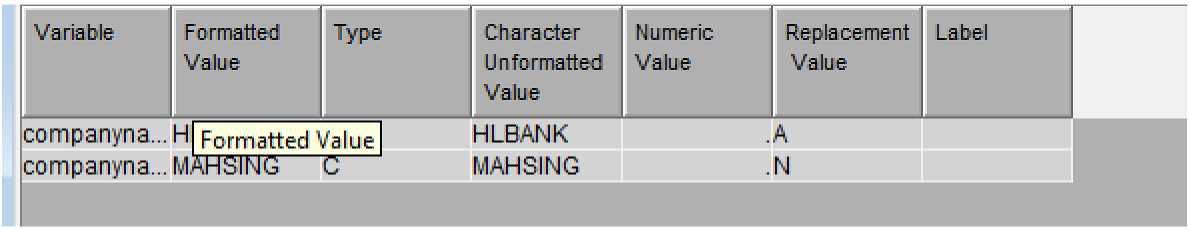


Figure 36 Total Replacement Counts Output

The total replacement count window shows the number of replacement that occurs in the training and validation data.



Variable	Formatted Value	Type	Character Unformatted Value	Numeric Value	Replacement Value	Label
companyna...	H Formatted Value		HLBANK		.A	
companyna...	MAHSING	C	MAHSING		.N	

Figure 37 Replacement value Variable

The replaced level values are consistent with expectation.

6.0 Recommendation

Time-series forecasting to show the investor the next 3 months stock price(High Price) of Mah Sing stock. A time series is a sequence of measurements recorded at equally-spaced intervals (hourly, weekly, monthly, etc.). As the name suggests, time series are inherently temporal. They often exhibit trends or seasonal patterns.

For example, the following plot shows Mah Sing stock price from 15/03/2019 till 30/04/2019:

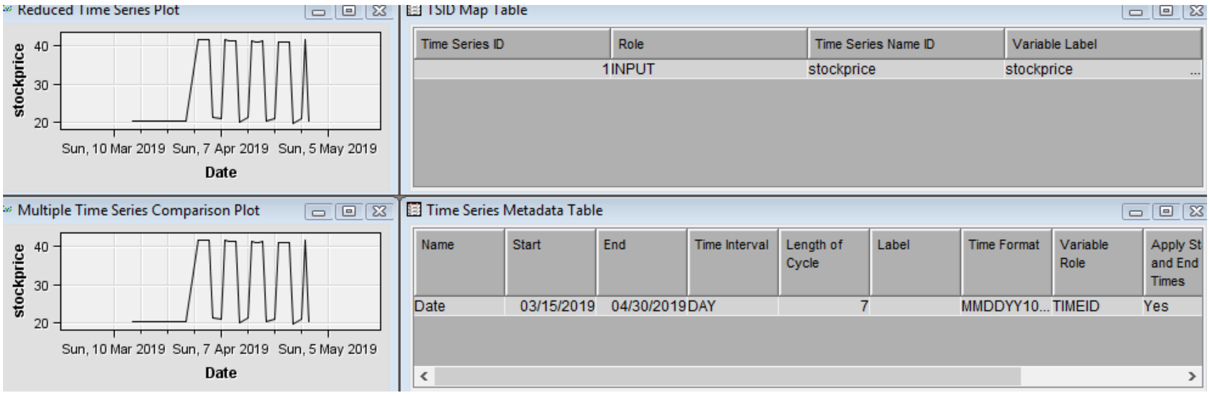


Figure 38 Plot of Mah Sing Stock Price 15/03-30/04

Generating Forecast

We can prepare the data and generate forecasts in Enterprise Miner with a simple three-node flow:

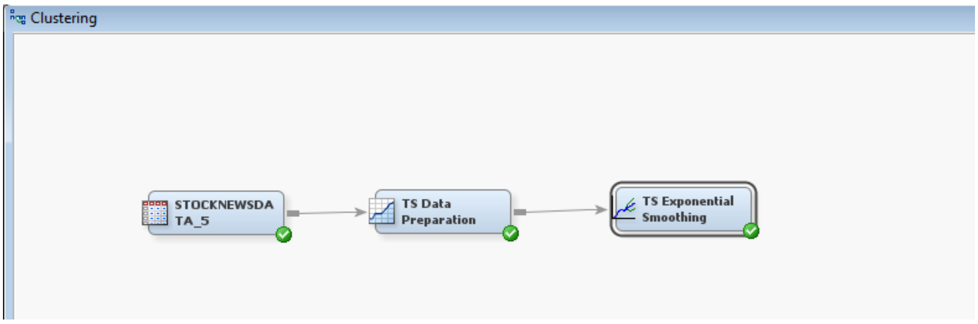


Figure 39 The Clustering Data Diagram

Data Source node

Here is a snippet of our historical data&colon

Obs #	Time Series ID	Time Series Name ID	Date	Time Series Value
1	1	stockprice	15Mar2019	20.3
2	1	stockprice	16Mar2019	.
3	1	stockprice	17Mar2019	.
4	1	stockprice	18Mar2019	20.3
5	1	stockprice	19Mar2019	20.3
6	1	stockprice	20Mar2019	20.3
7	1	stockprice	21Mar2019	20.3
8	1	stockprice	22Mar2019	20.3
9	1	stockprice	23Mar2019	.
10	1	stockprice	24Mar2019	.
11	1	stockprice	25Mar2019	20.3
12	1	stockprice	26Mar2019	20.3
13	1	stockprice	27Mar2019	20.3
14	1	stockprice	28Mar2019	20.3
15	1	stockprice	29Mar2019	20.3
16	1	stockprice	30Mar2019	.
17	1	stockprice	31Mar2019	.
18	1	stockprice	01Apr2019	41.54
19	1	stockprice	02Apr2019	41.7
20	1	stockprice	03Apr2019	41.76
21	1	stockprice	04Apr2019	41.6
22	1	stockprice	05Apr2019	21.24
23	1	stockprice	06Apr2019	.
24	1	stockprice	07Apr2019	21.16
25	1	stockprice	08Apr2019	41.48
26	1	stockprice	09Apr2019	41.37
27	1	stockprice	10Apr2019	41.14
28	1	stockprice	11Apr2019	41.19
29	1	stockprice	12Apr2019	20.16
30	1	stockprice	13Apr2019	.
31	1	stockprice	14Apr2019	21.31
32	1	stockprice	15Apr2019	41.41
33	1	stockprice	16Apr2019	41.04
34	1	stockprice	17Apr2019	40.94
35	1	stockprice	18Apr2019	41.32
36	1	stockprice	19Apr2019	20.38
37	1	stockprice	20Apr2019	.
38	1	stockprice	21Apr2019	21.04
39	1	stockprice	22Apr2019	41.09
40	1	stockprice	23Apr2019	41.07
41	1	stockprice	24Apr2019	41.06
42	1	stockprice	25Apr2019	40.95
43	1	stockprice	26Apr2019	19.96
44	1	stockprice	27Apr2019	.
45	1	stockprice	28Apr2019	21.09
46	1	stockprice	29Apr2019	41.59
47	1	stockprice	30Apr2019	20.5

Figure 40 Dataset of Mah Sing StockPrice 15/03-30/04

Each row represents different date of the stock price. There are some missing values.

The Data Source node defines modelling roles for the 2 columns:

Name	Use	Role	Level
Date	**Default**	Time ID	Interval
stockprice	**Default**	Target	Interval

Figure 41 Variable of Data Source

Date has assigned as the TimeID and stock price as the target.

TS Data Preparation node

The TS Data Preparation node transforms our data into a monthly series.

The time interval configured as Day and set Accumulation to Total. This will aggregate Mah Sing stock price by day.

Time Interval	
Specify an Interval	Day
Seasonal Cycle Selection	Default
Length of Cycle	2
Start and End Time	User Specified
Date Time Selector	...
Accumulation	Total

Figure 42 Time Interval Properties

To restrict the analysis to the time window of interest, ignoring early data points, had specified a custom Start and End times for the series using the Date Time Selector control:

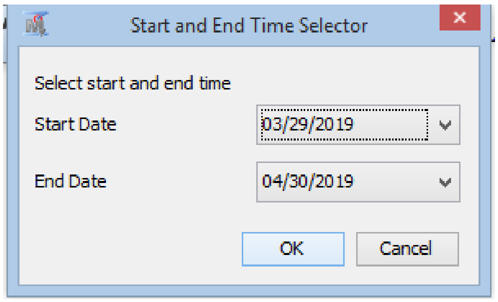


Figure 43 Start and End Time Selector

When this node is run, the TS Data Preparation node exports the transformed dataset:

Obs #	Time Series ID	Time Series Name ID	Date	Time Series Value
1	1	stockprice	29Mar2019	20.3
2	1	stockprice	30Mar2019	.
3	1	stockprice	31Mar2019	.
4	1	stockprice	01Apr2019	41.54
5	1	stockprice	02Apr2019	41.7
6	1	stockprice	03Apr2019	41.76
7	1	stockprice	04Apr2019	41.6
8	1	stockprice	05Apr2019	21.24
9	1	stockprice	06Apr2019	.
10	1	stockprice	07Apr2019	21.16
11	1	stockprice	08Apr2019	41.48
12	1	stockprice	09Apr2019	41.37
13	1	stockprice	10Apr2019	41.14
14	1	stockprice	11Apr2019	41.19
15	1	stockprice	12Apr2019	20.16
16	1	stockprice	13Apr2019	.
17	1	stockprice	14Apr2019	21.31
18	1	stockprice	15Apr2019	41.41
19	1	stockprice	16Apr2019	41.04
20	1	stockprice	17Apr2019	40.94
21	1	stockprice	18Apr2019	41.32
22	1	stockprice	19Apr2019	20.38
23	1	stockprice	20Apr2019	.
24	1	stockprice	21Apr2019	21.04
25	1	stockprice	22Apr2019	41.09
26	1	stockprice	23Apr2019	41.07
27	1	stockprice	24Apr2019	41.06
28	1	stockprice	25Apr2019	40.95
29	1	stockprice	26Apr2019	19.96
30	1	stockprice	27Apr2019	.
31	1	stockprice	28Apr2019	21.09
32	1	stockprice	29Apr2019	41.59
33	1	stockprice	30Apr2019	20.5

Figure 44 Dataset of Mah Sing StockPrice after Date selector

TS Exponential Smoothing node

In Enterprise Miner, forecasting is supported using Exponential Smoothing, a very popular and useful method for producing forecasts. This technique computes a new series of fitted values and forecasts, where each value in the series is the weighted average of the values observed at prior time points. It is called "Exponential smoothing" because the weights decrease exponentially to ensure that recent observations carry more weight than older ones. Exponential Smoothing tends to work best for short-term forecasts, i.e., forecasting a few time periods into the future.

In node properties, I've accepted the default Forecasting Method (Automatic), which tells Enterprise Miner to try various exponential smoothing methods and choose the one that has the best fit to the observed series.

The Forecast Lead set to 90, which will give us forecasts for the next 90 days.

Train	
Variables	...
Specify an Interval	Day
Accumulation	Total
Seasonality	Default
Forecasting Method	Best
Forecast Lead	90
Forecast Back	0
Forecast Sum Start	1
Significance Level	0.05

Figure 45 TSEM Node Properties

Results

Here is the plot of fitted values and forecasts that the TSEM node produces:

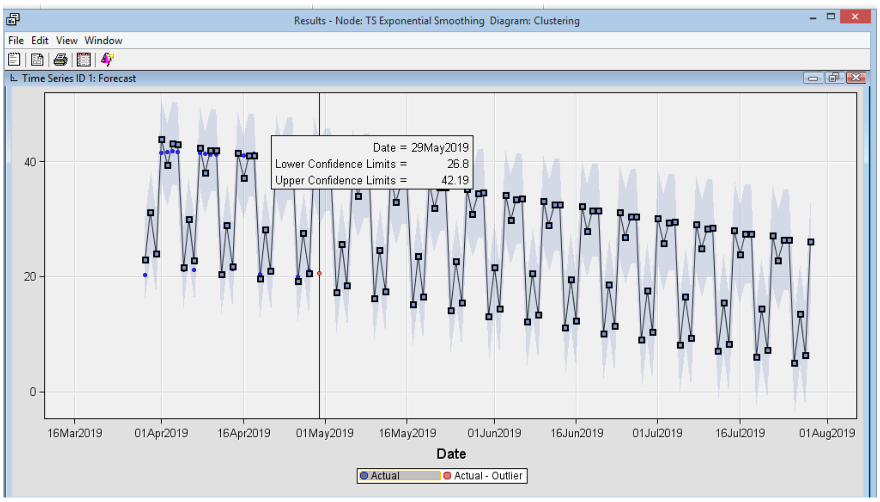


Figure 46 Forecast Plot

Actual stock price are shown as blue dots. Smoothed (fitted) values appear as a line that overlays blue squares. The light blue band represents the confidence interval for the fitted values.

Exponential smoothing has picked up on the seasonal variations in our data. Within each day the smoothed values show an early spike in the stock price followed a gradual decline.

By eyeballing the plot, it can gauge how well the smoothed values fit the observed values; by the Fit Statistics table , which contains error-based fit measures like Mean Square Error and Mean Absolute Percent Error.

Fit Statistics

Time Series ID	Time Series [illegible]e ID	Variable Name	Region	Degrees of Freedom Error	Number of Observations Used	Number of Observations	Number of Missing Actuals	Number of Missing Predicted Values	Number of Model Parameters	Total Sum of Squares	Corrected Total Sum of Squares	Sum of Square Error	Mean Square Error	Root Mean Square Error	Unbiased Mean Square Error	Unbiased Root Mean Square Error	Mean Absolute Percent Error
1	stockprice ...	_TSVALUE_	FIT	24	33	33	6	0	3	33299.97	2673.398	321.0685	11.89143	3.448395	13.37786	3.657575	6.36792

Figure 47 Fit Statistics table

Forecasts and their confidence intervals appear to the right of the vertical reference line. Consistent with the seasonal pattern, the forecast calls for a rise then a drop off in subsequent months.