WQD180024

Teng Lung Yun

WQD7005 DATA MINING

ASSIGNMENT
MILESTONE 5
PREDICTIVE MODELING

The stock I am analyzing are metal stocks mainly in long steel industry. The 4 main stocks are Lionind, Masteel, Ssteel, and Annjoo. Hence, I crawled the all the news for these 4 stocks for the past one year for all these stocks. The news is crawled from https://klse.i3investor.com.

Preprocessing Data.

	A	B	C	D	E	F	G	H	I
		Lionind	Masteel	Ssteel	mean	date	convertvalue	Sentiment	
1	1	0.111111	0	0	0.037037	4/30/2018	positive	1	
2	2	0	NA	0.058632	0.029316	5/4/2018	positive	1	
3	3	0.333333	NA	NA	0.333333	5/17/2018	positive	1	
4	4	0.4	NA	0.030651	0.215326	5/28/2018	positive	1	
5	5	0	0	0	0	5/28/2018	neutral	0	
6	6	0	0	0	0	5/28/2018	neutral	0	
7	7	0	NA	0	0	6/7/2018	neutral	0	
8	8	0.2	0.333333	0.052632	0.195322	6/11/2018	positive	1	
9	9	0.2	NA	-0.05128	0.074359	6/21/2018	positive	1	
10	10	0	NA	0.022222	0.011111	6/21/2018	positive	1	
11	11	0	0.5	0.055344	0.185115	7/4/2018	positive	1	
12	12	0.375	NA	NA	0.375	7/4/2018	positive	1	
13	13	0	0.142857	0	0.047619	7/4/2018	positive	1	
14	14	0	0	0.047686	0.015895	7/4/2018	positive	1	
15	15	0	0	0.065574	0.021858	7/6/2018	positive	1	
16	16	0.166667	NA	0.076923	0.121795	7/19/2018	positive	1	
17	17	0	NA	NA	0	7/19/2018	neutral	0	
18	18	0	0	0	0	7/23/2018	neutral	0	
19	19	0.166667	NA	0.083333	0.125	7/25/2018	positive	1	
20	20	0	0	0.065778	0.021926	8/2/2018	positive	1	
21	21	0	0	0	0	8/2/2018	neutral	0	
22	22	0	NA	-0.02071	-0.01036	8/4/2018	negative	-1	
23	23	0	NA	0	0	8/9/2018	neutral	0	
24	24	0	NA	-0.02564	-0.01282	8/10/2018	negative	-1	
25	25	0.2	NA	0.025641	0.112821	8/13/2018	positive	1	
26	26	0.2	NA	-0.04878	0.07561	8/13/2018	positive	1	
27	27	0	NA	0	0	8/17/2018	neutral	0	

Figure 1: News sentiment.

Stock price data for the past 1 year is collected from https://www.investing.com/equities/southern-steel-bhd-historical-data.

The same is repeated for Lionind, Masteel and Annjoo.

```
masteel$ï..Date <- mdy(masteel$ï..Date)
names(masteel)[1] <- 'date'
str(masteel)
masteel$date <- as.Date(masteel$date, format = "%d-%m-%Y")

masteel_senti <- masteel %>% left_join(senti, by = "date")
masteel_senti_1 <- masteel_senti[-c(8:13)]
masteel_senti_1$Sentiment[is.na(masteel_senti_1$Sentiment)]<-0
masteel_senti_1$name <- "masteel"

Annjoo <- read.csv("ANNJ.csv")
Annjoo$ï..Date <- mdy(Annjoo$ï..Date)
names(Annjoo)[1] <- 'date'
str(Annjoo)
Annjoo$date <- as.Date(Annjoo$date, format = "%d-%m-%Y")

Annjoo_senti <- Annjoo %>% left_join(senti, by = "date")
Annjoo_senti_1 <- Annjoo_senti[-c(8:13)]
Annjoo_senti_1$Sentiment[is.na(Annjoo_senti_1$Sentiment)]<-0
Annjoo_senti_1$name <- "Annjoo"

all <- rbind(sstee_senti_1, Annjoo_senti_1, masteel_senti_1, lionind_senti_1)
write.csv(all, "all.csv")
```

Figure 2: Preprocess data and combine dataset

The information collected from investing.com are price, open, high, and low with time stamp. The sentiment is then combined with the collected information with matching date.

Using R package, the time-stamped news sentiment polarity is calculated. The 4 stocks' news polarities are then combined and averaged. The news polarities are then correlated with the stock price for the 1st year using SAS.

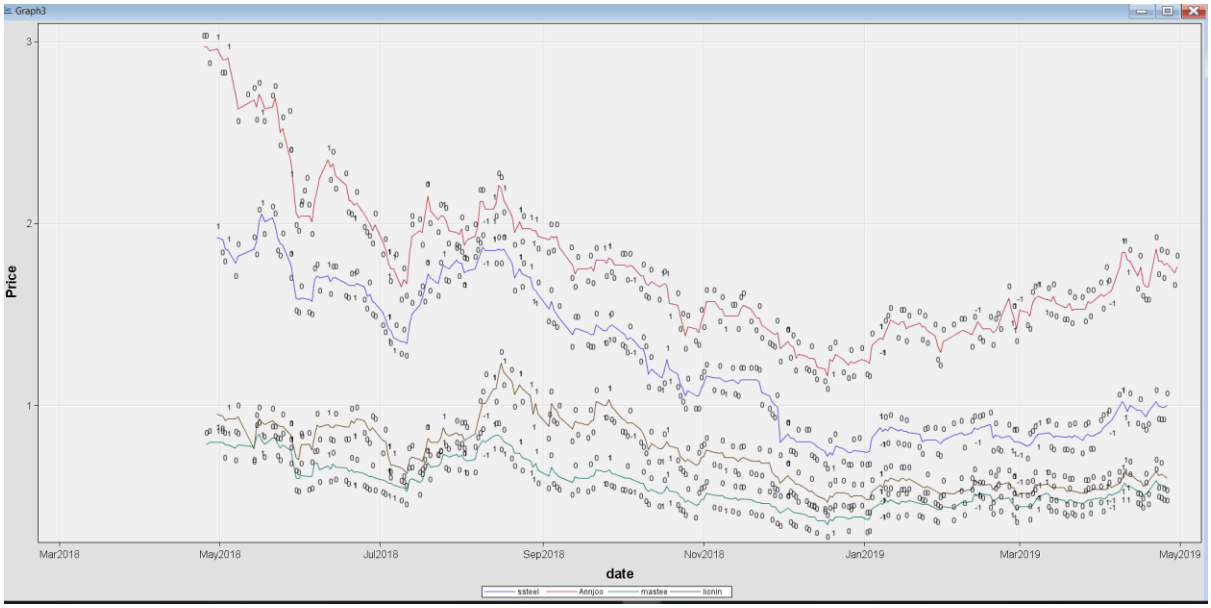


Figure 3: Closing Price of 4 Metal Stocks with respect to Time.

1 and 0 pertained to news sentiment. Positive sentiment is labelled as 1 while zero and negative is labelled as 0. From Figure 1, it is observed that there is a surge of closing price when there is positive news and vice versa.

Exploration of dataset.

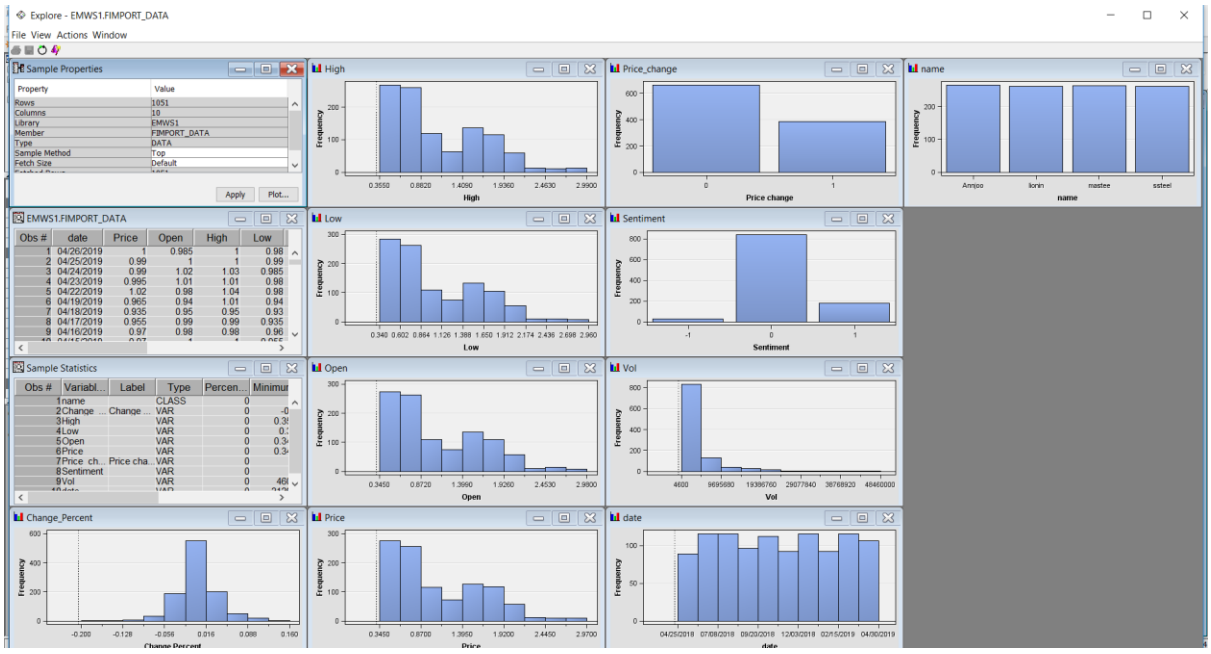


Figure 4: Exploration of data

Procedures:

1. Data loaded into SAS as SAS table.
2. Data partitioned into 50:30:20 training to validation to test set.
3. All interval attributes are transformed to log.
4. Decision Tree and Logistic Regression Models are constructed using the transformed data.
5. Use Model Comparison to compare two model.

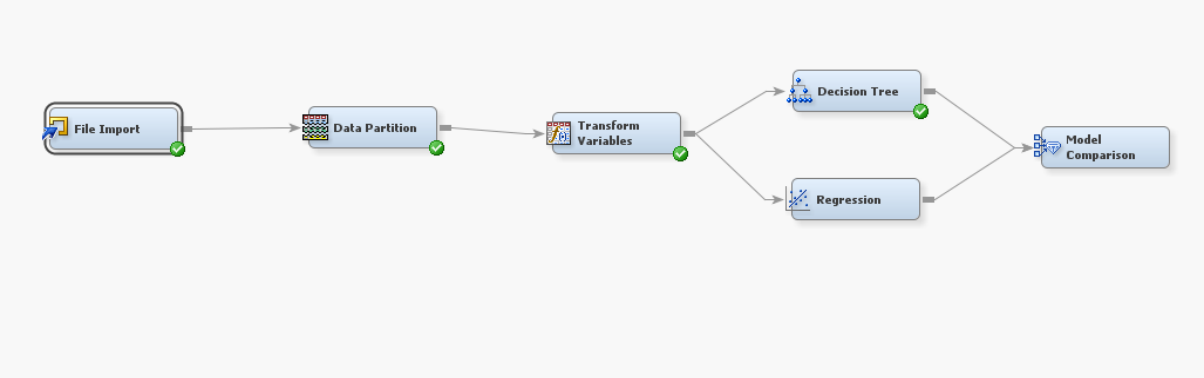
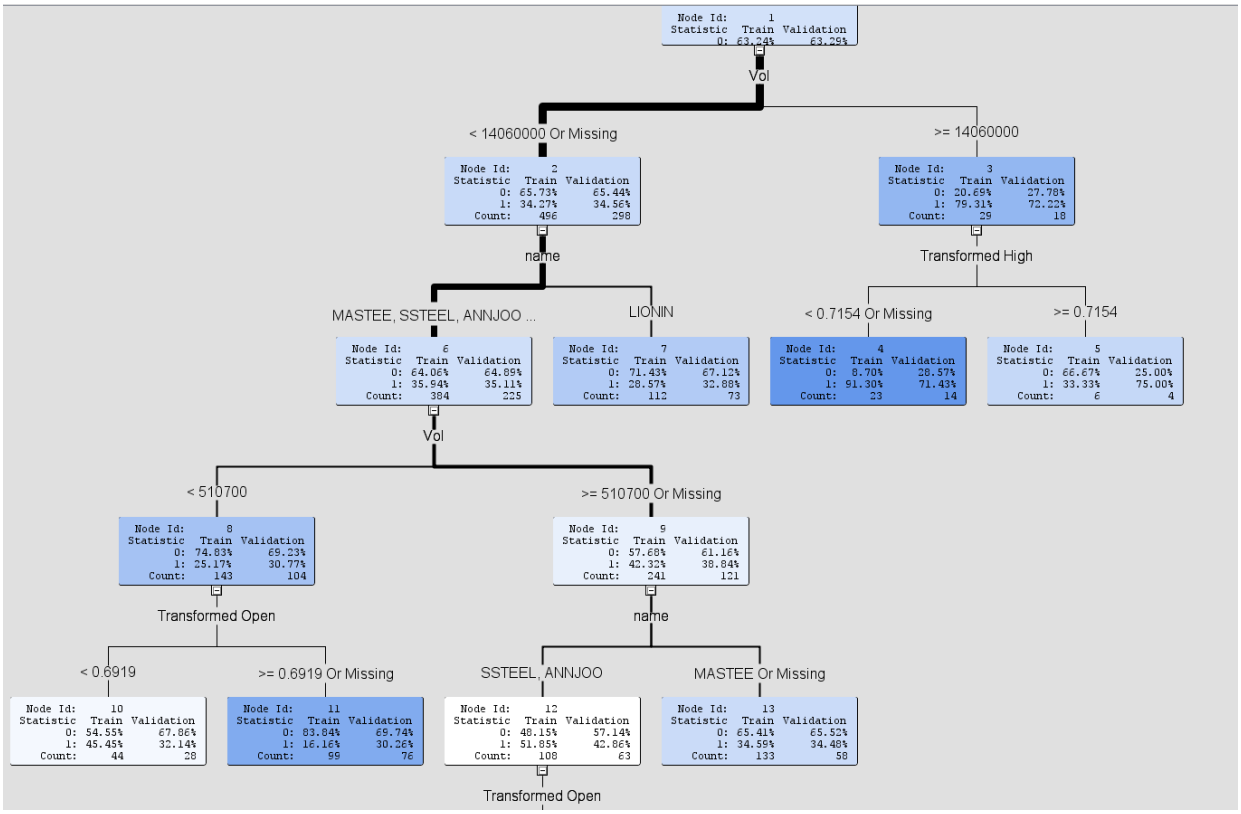


Figure 5: Steps/Procedures for Analysis

Modeling using Decision Tree



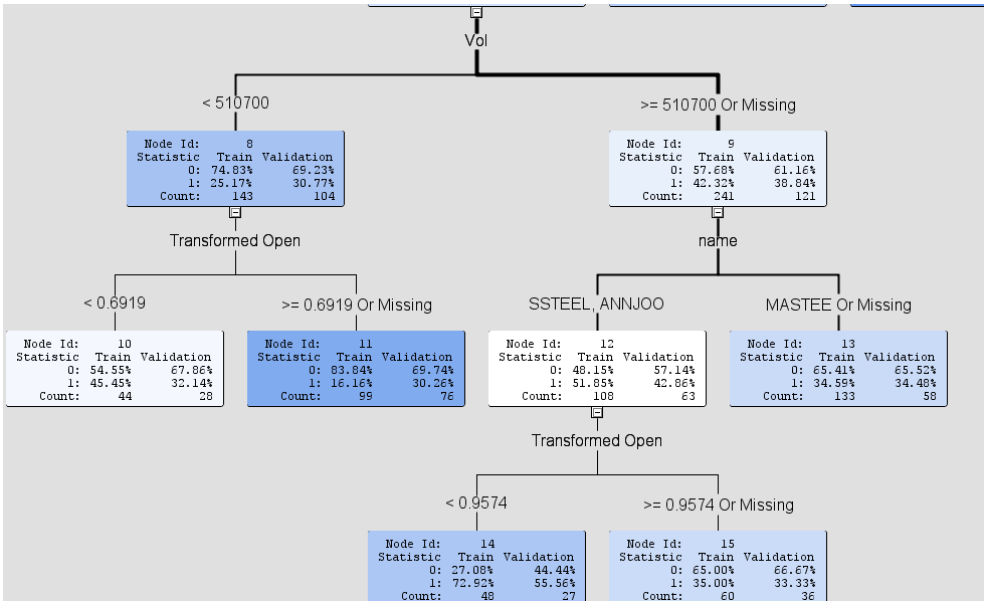


Figure 6: Decision Tree

The above figure shows the tree grown based on the value of information gain in the interactive mode.

Classification Table

Data Role=TRAIN Target Variable=Price_change Target Label=Price change

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	65.7258	98.1928	326	62.0952
1	0	34.2742	88.0829	170	32.3810
0	1	20.6897	1.8072	6	1.1429
1	1	79.3103	11.9171	23	4.3810

Data Role=VALIDATE Target Variable=Price_change Target Label=Price change

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	65.4362	97.5000	195	61.7089
1	0	34.5638	88.7931	103	32.5949
0	1	27.7778	2.5000	5	1.5823
1	1	72.2222	11.2069	13	4.1139

Figure 7: Decision Tree Accuracy Results

From Figure 7, classification results show 66.48% accuracy for train dataset and 65.81% for test dataset.

Data Role=TRAIN Target=Price_change Target Label=Price change

False Negative	True Negative	False Positive	True Positive
170	326	6	23

Data Role=VALIDATE Target=Price_change Target Label=Price change

False Negative	True Negative	False Positive	True Positive
103	195	5	13

Figure 8: Decision Tree Confusion Matrix

Modeling using Regression Tree

Classification Table

Data Role=TRAIN Target Variable=Price_change Target Label=Price change

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	87.1345	89.7590	298	56.7619
1	0	12.8655	22.7979	44	8.3810
0	1	18.5792	10.2410	34	6.4762
1	1	81.4208	77.2021	149	28.3810

Data Role=VALIDATE Target Variable=Price_change Target Label=Price change

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	83.8863	88.5000	177	56.0127
1	0	16.1137	29.3103	34	10.7595
0	1	21.9048	11.5000	23	7.2785
1	1	78.0952	70.6897	82	25.9494

Figure 9: Regression Accuracy Results

From Figure 9, classification results show 85.14% accuracy for train dataset and 81.95% for test dataset.

Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	Reg	Regression	0.18038	0.10403	0.14857	0.12395
	Tree	Decision Tree	0.33861	0.19726	0.28952	0.23369

Figure 10: Accuracy of both models using "Model Comparison"

Comparing both model, regression shows better accuracy for train and test dataset.

Event Classification Table
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Model Node	Model Description	Data Role	Target	Target Label	False Negative	True Negative	False Positive	True Positive
Tree	Decision Tree	TRAIN	Price_change	Price change	137	317	15	56
Tree	Decision Tree	VALIDATE	Price_change	Price change	91	184	16	25
Reg	Regression	TRAIN	Price_change	Price change	44	298	34	149
Reg	Regression	VALIDATE	Price_change	Price change	34	177	23	82

Figure 11: Confusion Matrix of Both Models

Youtube: https://www.youtube.com/watch?v=wYkgqAywb_c&feature=youtu.be