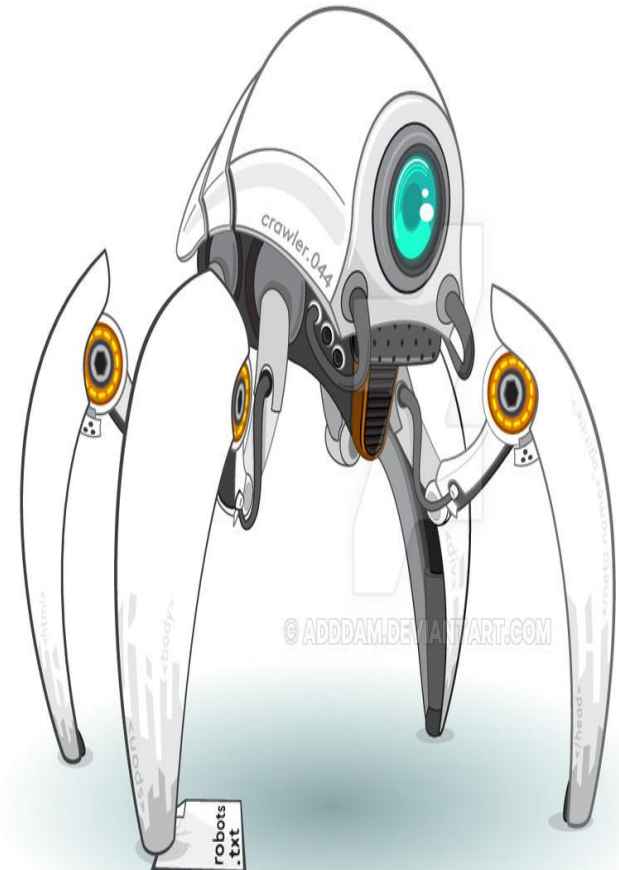# DATA MINING

ASSIGNMENT: MILESTONE 5

NORAISHA YUSUF (WQD180008)

UPDATE

Decision Tree and

Logistic Regression

Predictive model

- Sample data

Explore - LABTEST2.MERGEDSTOCK

File   View   Actions   Window

Sample Properties

Property	Value
Rows	7569
Columns	14
Library	LABTEST2
Member	MERGEDSTOCK
Type	DATA
Sample Method	Random
Fetch Size	Max
Fetched Rows	7569
Random Seed	12345

Apply   Plot...

Sample Statistics

Obs #	Variabl...	Label	Type	Percen...	Minimum	Maximum	Mean	Numbe...	M
1	board		CLASS	0	.	.	.	3	8
2	filters		CLASS	0	.	.	.	128+	3
3	name		CLASS	0	.	.	.	128+	2
4	senticlass		CLASS	0	.	.	.	3	6
5	target		CLASS	0	.	.	.	3	3
6	avg_senti		VAR	0	-0.9559	0.9922	0.204229	.	
7	close		VAR	0	0.005	150.4	3.430016	.	
8	day		VAR	0	21552	21637	21596.55	.	
9	dif		VAR	0	-1.6	1.18	0.002876	.	
10	difpct		VAR	0	-0.5	1.666667	0.003765	.	
11	high		VAR	0	0.01	150.6	3.460304	.	
12	low		VAR	0	0.005	149.3	3.396417	.	
13	open		VAR	0	0.005	149.6	3.428273	.	
14	volume		VAR	0	100	6.1665E8	5476020.	.	

LABTEST2.MERGEDSTOCK

Obs #	filters	day	name	board	close	open	high	low	volume	dif	difpct	avg_senti	senticl
290	AIRASIA OR AIRASIA GROUP	2019-01-17	AIRASIA	Main	3.05	3.08	3.1	3.04	5781000	0	-0.0032786885245901 23	0	Neutral
291	AIRASIA OR AIRASIA GROUP	2019-01-18	AIRASIA	Main	3.04	3.05	3.05	3	3216200	-0.009999999999999787	-0.0032786885245901 23	0	Neutral
292	AIRASIA OR AIRASIA GROUP	2019-01-22	AIRASIA	Main	3.01	3	3.04	2.99	4245200	-0.03000000000000025	-0.0099868421052631637	-0.020566666666666667	Negative
293	AIRASIA OR AIRASIA GROUP	2019-01-24	AIRASIA	Main	3.08	3.04	3.08	3.03	7260500	0.03000000000000025	0.0098360655573770592	0.81055	Positive
294	AIRASIA OR AIRASIA GROUP	2019-01-25	AIRASIA	Main	3.09	3.09	3.1	3.05	3371300	0.009999999999999787	0.0032467532467532756	0.7249	Positive
295	AIRASIA OR AIRASIA GROUP	2019-01-28	AIRASIA	Main	3.09	3.09	3.1	3.05	5113200	0	0	0.22155833333333333	Positive
296	AIRASIA OR AIRASIA GROUP	2019-01-29	AIRASIA	Main	3.13	3.1	3.16	3.08	19211100	0.04000000000000036	0.012944983818770295	-0.018175000000000024	Negative
297	AIRASIA OR AIRASIA GROUP	2019-01-31	AIRASIA	Main	3.04	3.04	3.09	3.03	5724400	-0.009999999999999787	-0.0032786885245901 23	0.46035000000000004	Positive
298	AIRASIA OR AIRASIA GROUP	2019-02-04	AIRASIA	Main	3.14	3.1	3.14	3.08	6236700	0.1000000000000007	0.03289473684210531	0.19296666666666665	Positive
299	AIRASIA OR AIRASIA GROUP	2019-02-07	AIRASIA	Main	3.13	3.14	3.14	3.09	2571100	-0.010000000000000231	-0.0031847133757962887	0	Neutral
300	AIRASIA OR AIRASIA GROUP	2019-02-12	AIRASIA	Main	3.12	3.13	3.13	3.08	4235400	0	0	0.50725	Positive
301	AIRASIA OR AIRASIA GROUP	2019-02-13	AIRASIA	Main	3.1	3.12	3.12	3.09	4158200	-0.02000000000000014	-0.0064102564102563875	0	Neutral
302	AIRASIA OR AIRASIA GROUP	2019-02-15	AIRASIA	Main	3.05	3.07	3.08	3.05	3913400	-0.04000000000000036	-0.012944983818770184	0.29505	Positive
303	AIRASIA OR AIRASIA GROUP	2019-02-18	AIRASIA	Main	2.99	3.08	3.08	2.98	7090800	-0.05999999999999961	-0.01967213114754085	0.7381846153846149	Positive
304	AIRASIA OR AIRASIA GROUP	2019-02-19	AIRASIA	Main	2.95	2.99	3	2.93	7511900	-0.04000000000000036	-0.013377926421404673	0.0027833333333333323	Positive
305	AIRASIA OR AIRASIA GROUP	2019-02-20	AIRASIA	Main	3.14	2.98	3.14	2.97	11483600	0.18999999999999995	0.06440677966101682	0.3030333333333333	Positive
306	AIRASIA OR AIRASIA GROUP	2019-02-21	AIRASIA	Main	3.29	3.12	3.29	3.12	14385600	0.14999999999999999	0.047770700636942554	0.18036666666666668	Positive

Decision Tree

- 3 classes – up, down, unchanged

Data partition 50%:50%

```
Summary Statistics for Class Targets

Data=DATA

               Numeric     Formatted     Frequency
Variable        Value        Value          Count        Percent      Label

  target          .         Down             2815        37.1912
  target          .         Unchanged        1754        23.1735
  target          .         Up               3000        39.6354


Data=TRAIN

               Numeric     Formatted     Frequency
Variable        Value        Value          Count        Percent      Label

  target          .         Down             1407        37.1927
  target          .         Unchanged         876        23.1562
  target          .         Up               1500        39.6511


Data=VALIDATE

               Numeric     Formatted     Frequency
Variable        Value        Value          Count        Percent      Label

  target          .         Down             1408        37.1896
  target          .         Unchanged         878        23.1907
  target          .         Up               1500        39.6197
```
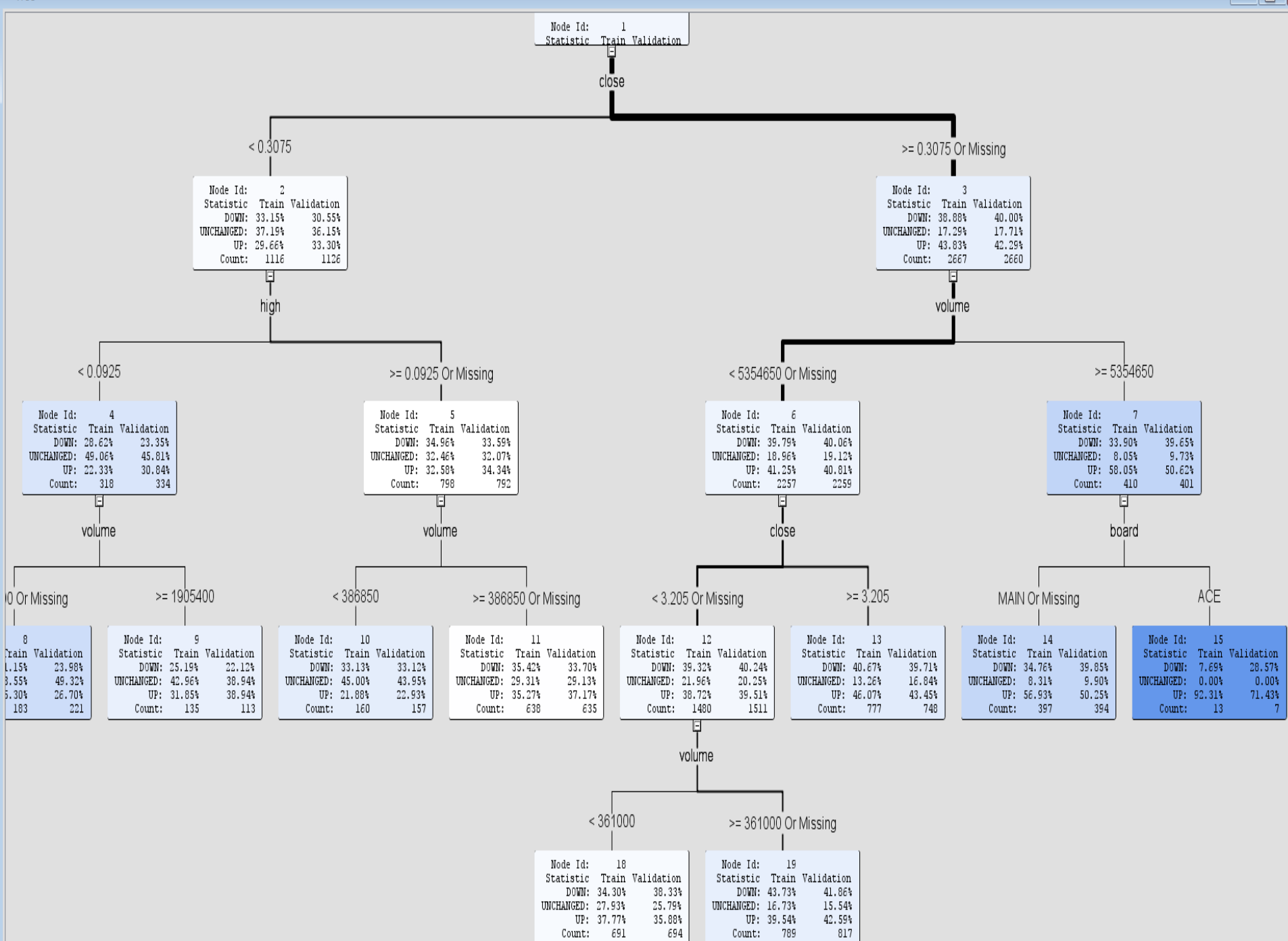
Node Id: 1
Statistic Train Validation

Split: **close**

- **< 0.3075**
- **>= 0.3075 Or Missing**

Node Id: 2
Statistic	Train	Validation
DOWN:	33.15%	30.55%
UNCHANGED:	37.19%	36.15%
UP:	29.66%	33.30%
Count:	1116	1126

Node Id: 3
Statistic	Train	Validation
DOWN:	38.88%	40.00%
UNCHANGED:	17.29%	17.71%
UP:	43.83%	42.29%
Count:	2667	2660

Split (Node 2): **high**
- **< 0.0925**
- **>= 0.0925 Or Missing**

Split (Node 3): **volume**
- **< 5354650 Or Missing**
- **>= 5354650**

Node Id: 4
Statistic	Train	Validation
DOWN:	28.62%	23.35%
UNCHANGED:	49.06%	45.81%
UP:	22.33%	30.84%
Count:	318	334

Node Id: 5
Statistic	Train	Validation
DOWN:	34.96%	33.59%
UNCHANGED:	32.46%	32.07%
UP:	32.58%	34.34%
Count:	798	792

Node Id: 6
Statistic	Train	Validation
DOWN:	39.79%	40.06%
UNCHANGED:	18.96%	19.12%
UP:	41.25%	40.81%
Count:	2257	2259

Node Id: 7
Statistic	Train	Validation
DOWN:	33.90%	39.65%
UNCHANGED:	8.05%	9.73%
UP:	58.05%	50.62%
Count:	410	401

Split (Node 4): **volume**
- **...0 Or Missing**
- **>= 1905400**

Split (Node 5): **volume**
- **< 386850**
- **>= 386850 Or Missing**

Split (Node 6): **close**
- **< 3.205 Or Missing**
- **>= 3.205**

Split (Node 7): **board**
- **MAIN Or Missing**
- **ACE**

Node 8
Statistic	Train	Validation
DOWN:	..15%	23.98%
UNCHANGED:	..55%	49.32%
UP:	..30%	26.70%
Count:	183	221

Node Id: 9
Statistic	Train	Validation
DOWN:	25.19%	22.12%
UNCHANGED:	42.96%	38.94%
UP:	31.85%	38.94%
Count:	135	113

Node Id: 10
Statistic	Train	Validation
DOWN:	33.13%	33.12%
UNCHANGED:	45.00%	43.95%
UP:	21.88%	22.93%
Count:	160	157

Node Id: 11
Statistic	Train	Validation
DOWN:	35.42%	33.70%
UNCHANGED:	29.31%	29.13%
UP:	35.27%	37.17%
Count:	638	635

Node Id: 12
Statistic	Train	Validation
DOWN:	39.32%	40.24%
UNCHANGED:	21.96%	20.25%
UP:	38.72%	39.51%
Count:	1480	1511

Node Id: 13
Statistic	Train	Validation
DOWN:	40.67%	39.71%
UNCHANGED:	13.26%	16.84%
UP:	46.07%	43.45%
Count:	777	748

Node Id: 14
Statistic	Train	Validation
DOWN:	34.76%	39.85%
UNCHANGED:	8.31%	9.90%
UP:	56.93%	50.25%
Count:	397	394

Node Id: 15
Statistic	Train	Validation
DOWN:	7.69%	28.57%
UNCHANGED:	0.00%	0.00%
UP:	92.31%	71.43%
Count:	13	7

Split (Node 12): **volume**
- **< 361000**
- **>= 361000 Or Missing**

Node Id: 18
Statistic	Train	Validation
DOWN:	34.30%	38.33%
UNCHANGED:	27.93%	25.79%
UP:	37.77%	35.88%
Count:	691	694

Node Id: 19
Statistic	Train	Validation
DOWN:	43.73%	41.86%
UNCHANGED:	16.73%	15.54%
UP:	39.54%	42.59%
Count:	789	817

Logistic Regression

- Only 2 classes considered: up and down

- Impute missing values

Output

```
54            Model Information
55
56    Training Data Set           WORK.EM_DMREG.VIE
57    DMDB Catalog                WORK.REG_DMDB
58    Target Variable             target
59    Target Measurement Level    Nominal
60    Number of Target Categories 2
61    Error                       MBernoulli
62    Link Function               Logit
63    Number of Model Parameters  385
64    Number of Observations      2907
65
66
67            Target Profile
68
69    Ordered                     Total
70    Value       target       Frequency
71
72       1        Up             1499
73       2        Down           1408
74
```

Results - Node: Impute Diagram: labtest2

File Edit View Window

Imputation Summary

Variable Name	Impute Method	Imputed Variable	Indicator Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
volume	MEAN	IMP_volume	M_volume	4193234	INPUT	INTERVAL		50

Output

```
43
44    Imputation Summary
45    Number Of Observations
46
47                                                          Number of
48    Variable  Impute   Imputed    Indicator  Impute              Measurement        Missing
49    Name      Method   Variable   Variable   Value     Role      Level       Label  for TRAIN
50
51    volume    MEAN     IMP_volume M_volume   4193234.05 INPUT    INTERVAL           50
52
53
54
55
56    Variable Distribution Training Data
57
58           Number of
59           Missing    Number of   Percent of
60    Obs    for TRAIN  Variables   Variables
61
62    1         50          1          100
63
```

Output   Plot...

```
138
139         Type 3 Analysis of Effects
140
141                              Wald
142    Effect          DF    Chi-Square    Pr > ChiSq
143
144    IMP_volume      1       0.0401       0.8413
145    M_volume        1       0.1811       0.6704
146    avg_senti       1       0.1546       0.6941
147    board           1       0.0035       0.9527
148    close           1       0.0026       0.9597
149    difpct          1      11.3234       0.0008
150    high            1       0.2442       0.6212
151    low             1       0.3459       0.5564
152    name          290    3309.4867      <.0001
153    open            1       0.0762       0.7824
154    senticlass      2       0.0909       0.9556
155
```

THE END