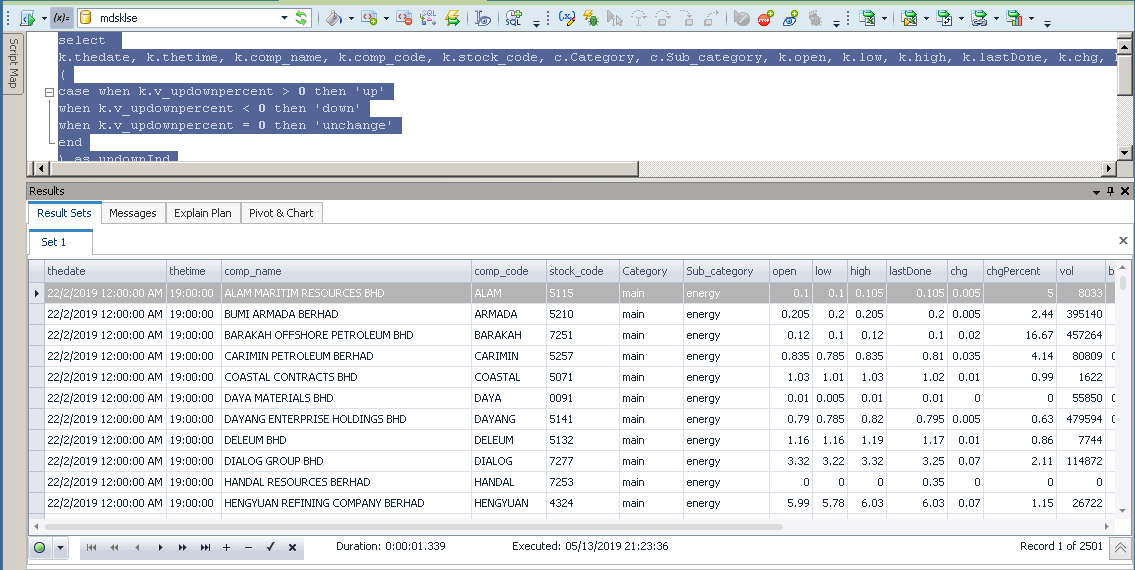
Milestone 5

To create a model of the stock data.

1. Query the data from mysql which is only select the data for Category: Main and Subcategory: energy

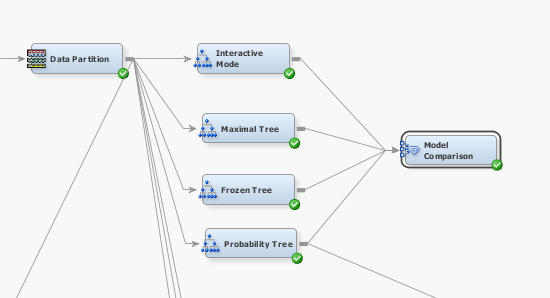
|  |
| --- |
| select  k.thedate, k.thetime, k.comp\_name, k.comp\_code, k.stock\_code, c.Category, c.Sub\_category, k.open, k.low, k.high, k.lastDone, k.chg, k.chgPercent, k.vol, k.buy, k.buyval, k.sell, k.sellval, k.crawl\_Timestamp, k.v\_updownpercent,  (  case when k.v\_updownpercent > **0** then 'up'  when k.v\_updownpercent < **0** then 'down'  when k.v\_updownpercent = **0** then 'unchange'  end  ) as updownInd  from upd\_klse  k  inner join compcat c  on k.comp\_code = c.compCode  where c.category = 'main'  and c.Sub\_category = 'energy'; |

The data is queried from MySQL server:



The extracted data will be exported to a csv file and import into SAS Enterprise miner.

1. The decision tree with different setting are created:



Decision Tree is selected to build a model, the attributes {open, low, high, last\_done, vol, buy and sell} of the stock data set whether are the factors to predict a stock price will rise or down or remain unchanged.

There are 4 kind of setting which described as below:

* Decision Tree – Interactive Mode

This tree is done by manually split the nodes.

* Decision Tree – Maximal Tree

This tree is done by auto splitting by Enterprise miner.

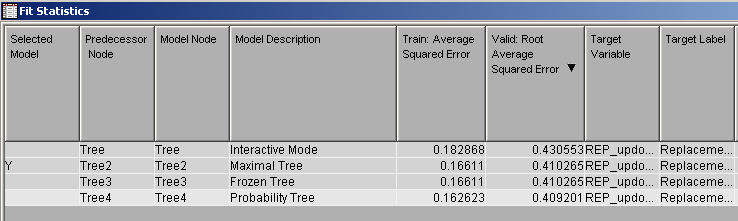
* Decision Tree – Frozen tree

This tree is done by applying the frozen tree setting

* Probability Tree

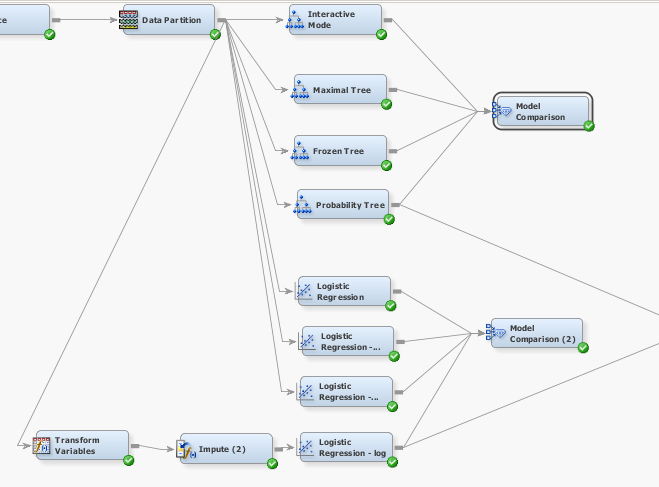
This tree is done by applying the Assessment Measures with “Average Square Error”

Among the 4 decision trees, by comparing the assessment measurement of each trees is Probability tree which having the minimum squared error.



1. Logistic Regression model

There are 4 Logistic Regression models have been created by applying different setting.



There are 4 kind of setting which described as below:

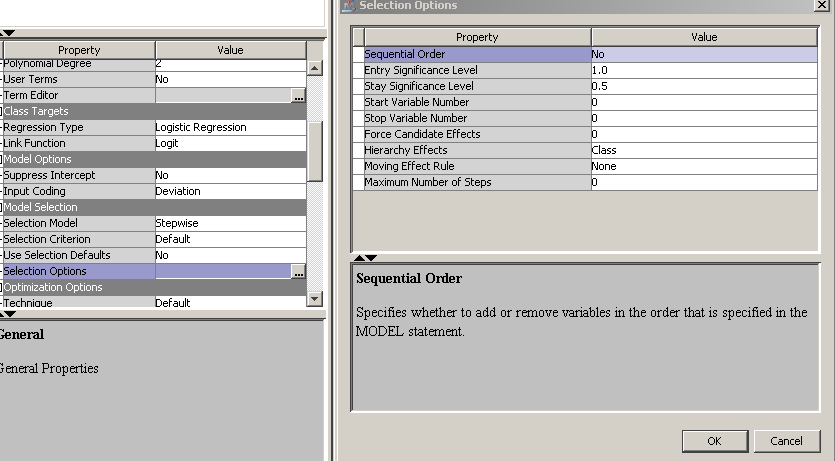
* Logistic Regression

This is default setting for the Logistic Regression.

* Logistic Regression - StepWise

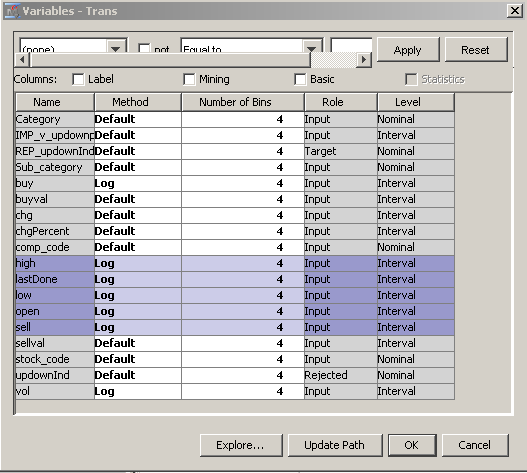
This regression is applied the setting at the Model selection : Model Select : Selection Model = Stepwise

* Logistic Regression - StepWise (set significant)

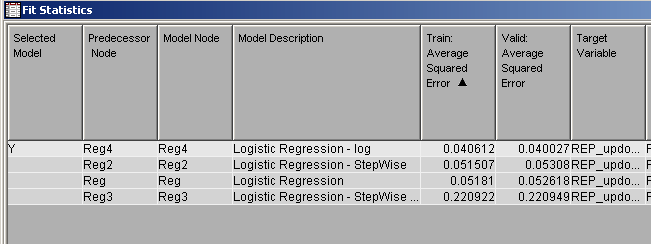
This regression is applied the selection option:  


* Logistic Regression - log

The input attributes are transformed to log value.

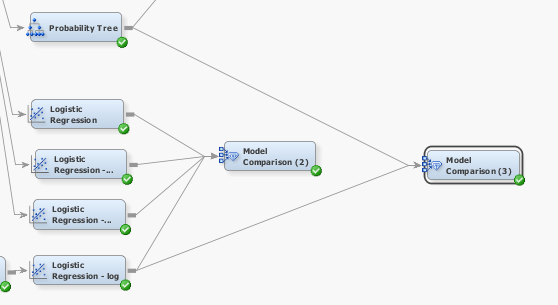


Among the 4 logistic regression, by comparing the assessment measurement of each regression is Logistic Regression - log which having the minimum squared error.

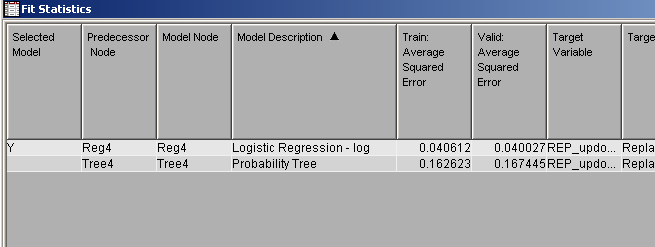


1. Comparison between Decision Tree and Logistic Regression:

Base on the data, to find which model (Decision Tree and Logistic Regression) is the best to describe the data, the comparison of these two model is done:



The result of the comparison between the best Decision Tree (Probability Tree) and Logistic Regression (Logistic Regression - log) by comparing the Average Squared Error:



1. Conclusion  
   Base on the comparison, Logistic Regression – log (by transforming the input attribute values by log) is the best model for the available data.