**ISE 365/465 – Data Mining Homework 3**

Chumeng Lyu

Chl614@lehigh.edu

**2. A one page write-up of your modeling strategy and method. Be sure to explain unique steps that you took to get better model results including new variables you created and settings you used in your modeling. This write-up will be judged on how well you explain what you did and why you did it. Make sure to include the explanation of your best model in this step.**

Firstly, I used IBM SPSS to preprocess the data. I used Derive node to create a new variable WineType, the WineType of red wine is red, and WineType of white wine is white. Then I use the Append node to combine the two given data sets. Then used another Derive node to create the new variable Quality\_rating, set quality>=7 to be T and else to be F. Then I imported the new data file into Enterprise Miner.

While building the model, I used StatExplore node to inspect input variables, and used SAS Code to see their correlation. To simplify the model and interpret the model more easily, in order to assure there are no variables have correlation more than 0.5, I decided to drop density because it is highly correlated with alcohol, and drop free sulfur dioxide and residual sugar, because their high correlation with total sulfur dioxide. Then I used partition node partition the data into 70/30 training/validation data set and used decision tree to establish model. I used 3 different types of nominal criterion to establish the model.

I also built decision tree models in another way: Firstly, I used Variable Selection node to eliminated variables that not highly correlated to the target, then used decision trees to establish the model.

By observing the results from Graphboard node in SPSS, we can see that some subsections of attribute have higher correlation with target. Thus, I also built a binned data set in IBM SPSS, by setting the number of bin as 5. And do the correlation elimination in Enterprise Miner, then used partition node and used decision tree to establish model.

What’s more, I built the decision trees of original data to get the control data.

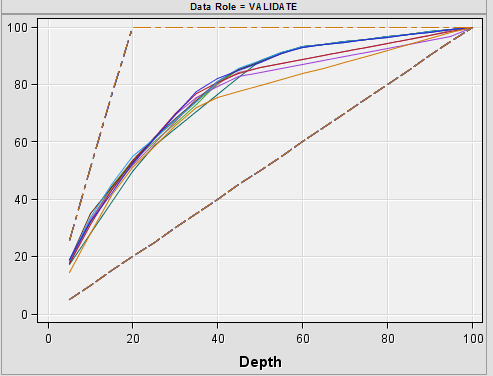
Finally, I use the Comparison node to compare the decision trees described above.

**3. A one page summary of the modeling results for steps 1 to 4 above. This summary will be graded on how well you explain the variables in your models and their contribution to the solution as well as an analysis of the quality of your models using techniques we have learned in class.**

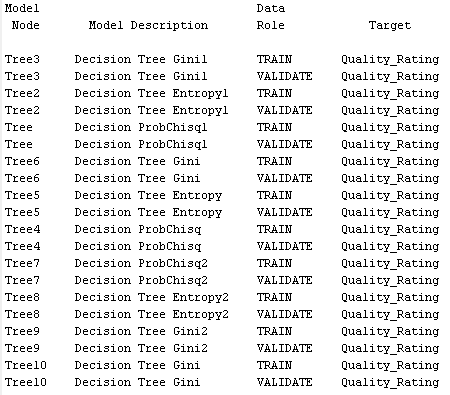
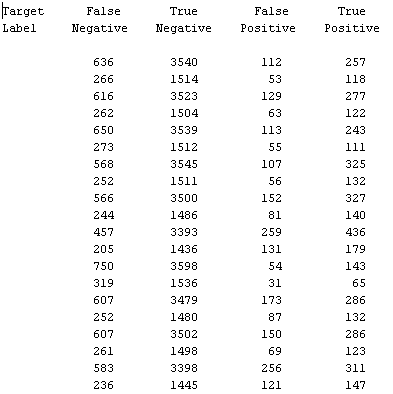
As shown from StatExplore Node, alcohol has the highest worth. And it has high correlation with density. This indicates that the alcohol content may partly decide the density of wine.

From the results of Graphboard node in SPSS and MultiPlot Node in Enterprise Miner, I found the proportion of high quality wine in high alcohol is more than low alcohol. Chlorides has the third highest worth. Wine with lower chlorides are more likely to be high quality. Citric acid is a type of weak acid. Too high or too low citric acid wine both have low proportion high quality. By looking up some related materials, we can know that total sulfur dioxide is the quantity of sulfur dioxide in wine. And free sulfur dioxide is part of total sulfur dioxide. The correlation between total sulfur dioxide and free sulfur dioxide are 0.72. Volatile acidity refers to the steam distillable acids present in wine, primarily acetic acid but also lactic, formic, butyric, and propionic acids. From MultiPlot Node, it looks like that low volatile acidity wine is more likely to be high quality. Residual sugar is usually measured in grams of sugar per liter of wine. It seems low residual sugar are more likely to be high quality, too. Traditionally total acidity is divided into two groups, namely the volatile acids and fixed acids. And fixed acids includes citric acid. Winemakers use pH as a way to measure ripeness in relation to acidity. According to reference, low pH wines will taste tart and crisp, while higher pH wines are more susceptible to bacterial growth. So too high or too low pH are less likely to be high quality, and this data file prove this. Surphates are some types of sulfur dioxide.

As a wine reviewer, I want to predict which wine will be of high quality, which means we want the percentage of true high quality wine among the wine that are predicted to be high quality is as high as possible. In this case we don’t care about those wine with low quality. Thus I think we should use precision to evaluate the model, and meanwhile we need keep accuracy in an acceptable range. From the chart below we can see that the results from training sets and validation sets are similar, and we use the results of validation sets. The precision, sensitivity, and accuracy are shown below.



**Exhibit 1 - Cumulative % Captured Response**

**Exhibit 2 - Results from Model Comparison node**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| DT Type | FN | TN | | FP | TP | sensitivity | precision | accuracy |
| **DT ProbChisq** | 205 | 1436 | 131 | | 179 | **0.466145833** | 0.57741935 | 0.827781 |
| **DT Entropy** | 244 | 1486 | 81 | | 140 | 0.364583333 | 0.63348416 | 0.833419 |
| **DT Gini** | 252 | 1486 | 81 | | 140 | 0.357142857 | 0.63348416 | **0.842132** |
| **DT ProbChisq with Drop** | 273 | 1512 | 55 | | 111 | 0.2890625 | 0.6686747 | 0.831881 |
| **DT Entropy with Drop** | 262 | 1504 | 63 | | 112 | 0.299465241 | 0.64 | 0.83319 |
| **DT Gini with Drop** | 266 | 1514 | 53 | | 118 | 0.307291667 | **0.69005848** | 0.836494 |
| **DT ProbChisq with Variable Selection** | 319 | 1536 | 31 | | 65 | 0.169270833 | 0.67708333 | 0.820605 |
| **DT Entropy with Variable Selection** | 252 | 1480 | 87 | | 132 | 0.34375 | 0.60273973 | 0.826243 |
| **DT Gini with Variable Selection** | 261 | 1498 | 69 | | 123 | 0.3203125 | 0.640625 | 0.830856 |
| **DT Gini use bin** | 236 | 1445 | 121 | | 147 | 0.38381201 | 0.54850746 | 0.816829 |

**Exhibit 3 – Sensitivity, precision, accuracy table**

The table above shows that the accuracy of all decision trees are generally close to each other, while the Gini decision tree built based on original data has the highest accuracy. From my opinion, the Gini decision tree built after dropping the highly correlated variables is the best model I have. It has the highest precision value and its accuracy is also acceptable. In this problem, I want to make sure accuracy is more than 80%. The result shows that by using this model (Gini decision tree with drop), about 69% of the wine that I predicted to be high quality are true high quality wine. Even the ProbChisq decision tree model(using original data) has higher sensitivity value, its precision is much lower, which means the correctness of the high quality wine prediction is worse than that using Gini decision tree with drop.