**Penalized Logistic Regression and Wine Quality**

**Susan Wiggins**

**Cases in Business Analytics BAN 525**

**Professor: Dr. Cetin Ciner**



**Table of Content:**

Introduction………………………………………………………………………………………………………………………………………3-5

Analysis and Model Comparison……………………………………………………………………………………………………….5-7

Interpretation……………………………………………………………………………………………………………………………………8-9

References………………………………………………………………………………………………………………………………………..10

Appendix A **(Ordinary Logistic Regression)**………………………………………………………………………………………11-12

Appendix B **(General Regression Lasso)**……………………………………………………………………………………………12-13

Appendix C **(General Regression Adaptive Lasso)**……………………………………………………………………………14-15

Appendix D **(General Regression Elastic Net)**…………………………………………………………………………………..16-17

Appendix E **(General Regression Adaptive Elastic Net)**……………………………………………………………………18-19

Appendix F **(Comparison Model)**………………………………………………………………………………………………………20

Appendix G **(Adaptive Lasso Final Profiler Results)**………………………………………………………………………….21-22

Appendix H **(Profiler Predictions)**…………………………………………………………………………………………………….23-24

Appendix I **(Summary Report Variable**)……………………………………………………………………………………………25-27

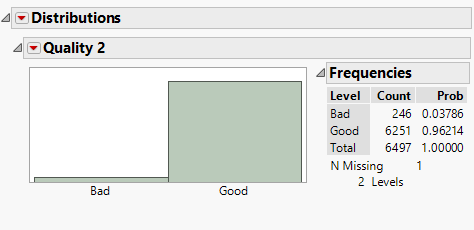
Appendix J **(Prediction of Sample Probability)**…………………………………………………………………………………28

**Introduction**

Red wine a pleasure of many wine lovers is known for helping with in health efforts such as lowering the risk of heart disease, Stroke and Death. Only a small dosage is recommended for such preventive measures. This case study involves an investigation into asking what physiochemical aspects determine whether the probability that a wine sample is “Bad” or “Good”. Physiochemical is a category of the science field that deals with chemical and biological aspects. The dataset consists of wine from Portuguese called “Vinho Verde”. Vinho Verde wine is considered difficult to produce in the Minho, Portuguese region. The majority of Vinho Verde are white wines. Red grapes are difficult to produce in this region due to the rainy weather and cool atmosphere. By far this region produces a rare wine just based on the grape itself who rosy nature presents a sense of distinct acidity and noted fruity taste. The map below notes the locality of Minho region where Vino Verde is produced. The Minho region takes on the Atlantic’s Challenges of weather where it protrudes out within the country’s geography.



The dataset has been prepared based on the following criterion ranking from 1 (lowest quality) to 10 (best quality). Ranking were further deduced based on the adjustments as Bad (unacceptable) noted 3 and 4 as well as Good (acceptable) being any value above five. The dependent variable being measured in this study is Quality (Bad) and the predicator candidates are Fixed acidity, Volatile acidity, Citric Acid, Residual Sugar, Chlorides, Free Sulfur Dioxide, Total Sulfur Dioxide, Density, Sulphates, Alcohol, and Color. In order to understand some of the terminology associated within this dataset a brief explanation of variables is noted; for instance, P. Cortez defines that, “Chlorides are the amount of salt associated in the wine. Alcohol is the percent alcohol content in wine. Density refers to the density of water being closer to water depending on percent alcohol and sugar content. pH describes the basicity or acidity character of a wine. The scale for this pH level is from 0 (very acidic) to 14 (very basic) and most wines fall between a 3-4 category. In addition, sulphates are a wine additive which can contribute to sulfur dioxide levels, which act as an antimicrobial and antioxidant. Fixed acidity involves wines are fixed or nonvolatile (poor evaporation). Volatile acidity is the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste. Citric acid is found in small quantities, citric acid can add 'freshness' and flavor to wines. Residual sugar is the amount of sugar remaining after the fermentation process. Sweet wines are noted to measure at greater than 45 grams/liter and rare wines are found at the measurements for 1 gram/ liter. Next, Free sulfur dioxide is the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine. The final terminology used as a predicator is total sulfur dioxide which is the amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine” (1-2). Five methods will be addressed in this case ordinary logistics regression, penalized logistic regression lasso, penalized logistic regression adaptive lasso, penalized logistic regression elastic net, and penalized logistic regression adaptive elastic net. Logistic Regression is used in this case based on the categorical variable (Quality) being used to determine the outcome of the analysis. The distribution of Quality already informs the reader that there are a great many more of good wines than bad wines as seen below.



The probability of good is reported at 96 % whereas the probability of bad only reports probability of

3.8 %. Logistic Regression a power house for categorical data in essence will report the probability of an event through predictive factors used in the equation.

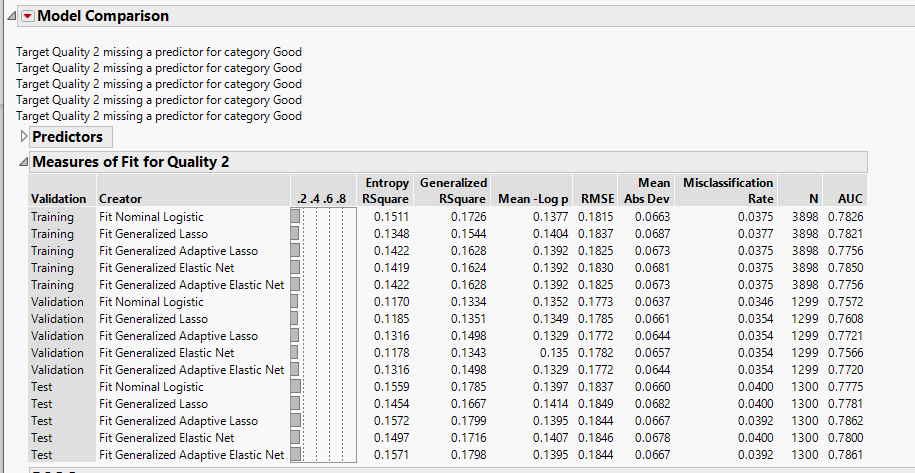
**Analysis and Model Comparison**

The dataset employed in this case study is large and sometimes modeling can be affected when Big data is considered. Ordinary Logistic Regression a modeling technique most highly used in prediction methods in categorical data can cause problems when Big data is being used. For instance, Ordinary Logistic Regression modeling can cause large variances in the data, models random noise, and poor forecasting can exists. Therefore, other machine learning techniques need to be implemented such as Penalized Logistic Regression (Lasso, Adaptive Lasso, Elastic Net, Adaptive Elastic Net). In order to reach the best results in the analysis one must depend on cross validation which will decrease the ability to interpret random noise. In cross validation, data is held out and the predication criterion is based on this interpretation. Estimates are then built on the data after the completion of each modeling affect occurs. In the case predication of Quality (Bad), the cross validation is established on 60/20/20 split of the data with a random seed of 123. The final interpretation of the model will be based on the results of the test data due to the basis of new observations.

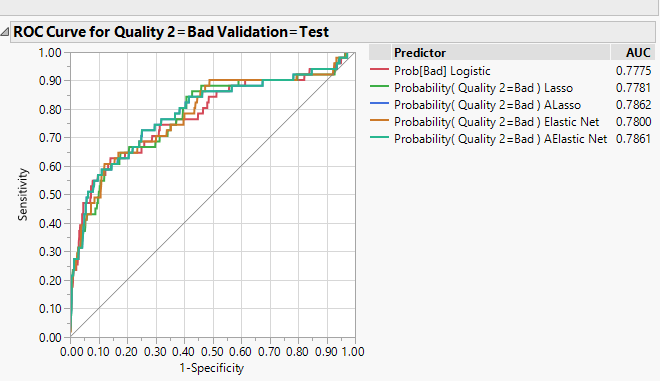
Next, the following penalized modeling methods were performed Ordinary Logistic Regression, Lasso, Adaptive Lasso, Elastic Net, and Adaptive Elastic Net. I found that all these modeling methods where not identical to the Ordinary Logistic Regression due to the Lambda training value not being zero. All Lambda values where positive in these modeling methods indicating something was affected using these models. The following table indicates the lambda values reported in this case study for predication of Quality.

|  |  |
| --- | --- |
| **Name Model Method** | **Training Lambda** |
| General Logistic Regression Lasso | 0.1561832 |
| General Logistic Regression Adaptive Lasso | 2.5401818 |
| General Logistic Regression Elastic Net | 0.0804973 |
| General Logistic Regression Adaptive Elastic Net | 2.5658402 |

In addition, the results of each one of these models are listed in the appendix. A model comparison was run for analysis of best results of modeling predications and chart is listed below.

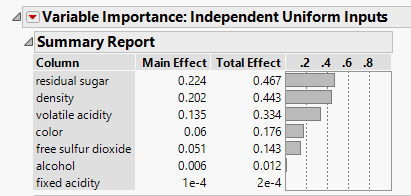


Due to the factor that each one of the new modeling techniques brought new information earlier based on the lambda values being a positive value, I was able to compare the best modeling for selection of Adaptive Lasso. The reason behind this selection is based on the comparison values of Misclassification being the lowest at .0392. In addition, the Area Under the Curve (AUC) reveals how well the data is being sorted in the modeling technique and an indication of a higher value towards the upper level of the curve at 1.0 informs the reader that the model sorting is better. The AUC for the Adaptive Lasso is the highest towards value in the reported model comparison at .7862. Therefore, the Adaptive Lasso method accounts for 78.62 % of the testing data for Quality (Bad). The ROC curve below is further evidence of comparison for AUC values. The ROC curve is a graphical representation of true positive (sensitivity) verses false positive rate (specifity). In this case the sensitivity is the classifying correctness of Quality “Bad”. Specifity deals with when the model classifies as Quality but when the Quality is “Good” in fact. The perfect test for a ROC curve hugs the upper quadrant of the graph and the gray vertical line indicates that the modeling technique would be useless below this reference line. Truly, the Adaptive Lasso is the model comparison of choice for the case study of wine quality.



**Interpretation**

The parameter estimates for the Adaptive Lasso model are reported as follows fixed acidity, volatile acidity, residual sugar, free sulfur dioxide, density, alcohol, and color. The positive relationships listed in the parameter estimates are fixed acidity, volatile acidity, and density and have a greater chance towards effecting the quality of wine interpretation of “Bad”. On the other hand, the negative relationships are residual sugar, free sulfur dioxide, alcohol, and color. The variables with a negative relationship will be of less importance towards the quality of wine. Next, the variable importance is included in the predication of quality and the following table gives us the results below.



The total effect considers the variations in the data and residual sugar accounts for 46.7 % of the variations in wine processed as “Bad”. Density and Volatile Acidity have significant importance in the model. Density 44% and Volatile Acidity 33% are noted. Color is the fourth variable of importance in Vinho Verde wine being labeled as Quality (Bad). Free Sulfur Dioxide has only a 14.3 % significance followed by alcohol and fixed acidity being very low in variable importance. The most important variable in this case study of quality of wine is residual sugar. Residual sugar stated earlier is the sugar remaining in the fermentation process after stopping. The model prediction profiler reveals that the color of wine has a higher possibility to be bad if red and a lower possibility if white. In addition, the quality of wine is more likely to be “Bad” when alcohol is at lower levels whereas higher level (15). The higher the density the higher the probability the quality is bad also. Another parameter profiler free sulfur dioxide ranging at 100 gives more chances for being “Bad” whereas from 100 to 300 indicates a less likely chances for significant influence. As residual sugar increases the likely of quality being “Bad” decreases in this model variable. Next, volatile acidity increases the likelihood that quality is bad within the Vinho Verde wine. The final predication profiler fixed acidity also increases the outcome of quality label “Bad”; therefore, as fixed acidity rises so does the chance for the wine quality to decrease.

A final step for the analysis of this case study was to predict a probability based on supplied parameters for field values in our study listed in Appendix I. The wine sample for “Bad” quality was to be reported based on the Adaptive Lasso method chosen and under the parameters in Appendix I reported a 4.8% likelihood of being “Bad”.

In conclusion, the selected model to predict Quality (Bad) was Penalized Logistic Regression Adaptive Lasso. The highest variable for fluctuations in quality is related to residual sugar at 46.7%. Density and Volatile Acidity are the next higher variables of importance, followed by color and free sulfur dioxide. As residual sugar increases the likely of quality being “Bad” decreases in this model variable. Density variable importance level is 44% and the higher the density the increased rate for the Vinho Verde to be “Bad”. The third significant variable listed at 33% was Volatile acidity and likelihood of quality being “Bad” increases with this variable. Color is ranked at fourth significance in variable importance. Red wine in the Minho region can have an increased likelihood of being “Bad” verses White wine of this region. Free sulfur dioxide is the fifth highest variable at 14%. Free sulfur dioxide ranging at 100 gives more chances for being “Bad” whereas from 100 to 300 indicates a less likely chances for significant influence. Alcohol and fixed acidity have very low levels of variable importance in this model. Finally, the wine sample probability in the Adaptive Lasso based on the parameters in Appendix I reported a 4.8% likelihood of being classified as “Bad”. Therefore, the parameter given for this estimation indicate the wine would be “Good”.

**Reference**

Ciner, Cetin. September 2019. Module 3. “Penalized Logistic Regression and Titanic Passengers.” Slides 1-43. Retrieved from: <https://uncw.instructure.com/courses/24252/pages/module-3-lecture-materials-penalized-logistic-regression-and-titanic-passengers?module_item_id=524507>

Ciner, Cetin. September 2019. Wine Quality Background. “Quality.” Slide 1-2. reference [Cortez et al., 2009] Retrieved from: <https://uncw.instructure.com/courses/24252/assignments/161360?module_item_id=524509>

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009

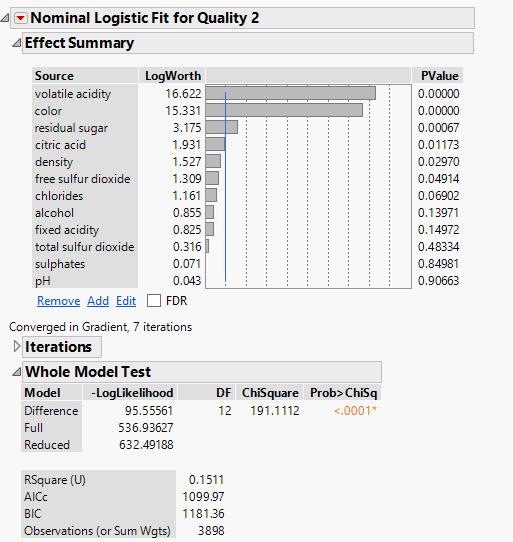
Puckette, Madeline. “101 Guides to Vinho Verde Wine from Portugal.” Wine Folly. Wine Folly Inc. 30 July 2019. Web. 13 Sept. 2019. <https://winefolly.com/review/vinho-verde-the-perfect-poolside-wine-from-portugal/>

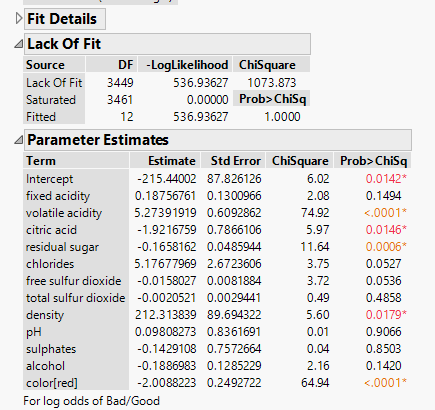
Web. Picture. <https://www.foodbusinessafrica.com/2019/05/27/kwal-to-invest-us30m-in-new-manufacturing-facility-to-expand-operations/>

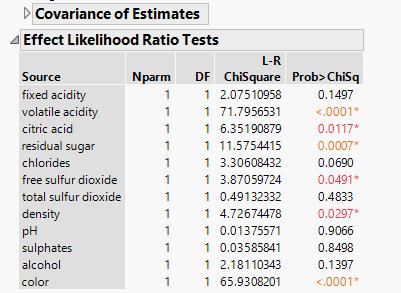
Website. Health. <https://www.google.com/search?ei=OZF7XfXnC-_B_Qb8wY7oDQ&q=what+make+red+wine+bad&oq=what+make+red+wine+bad&gs_l=psy-ab.3..33i22i29i30.249720.255590..256269...0.3..0.103.1224.17j1......0....1..gws-wiz.......0i71j0j0i22i30.1h3myIc_pok&ved=0ahUKEwi16Lvl683kAhXvYN8KHfygA90Q4dUDCAo&uact=5>

**Appendix A**

**Logistic Regression**

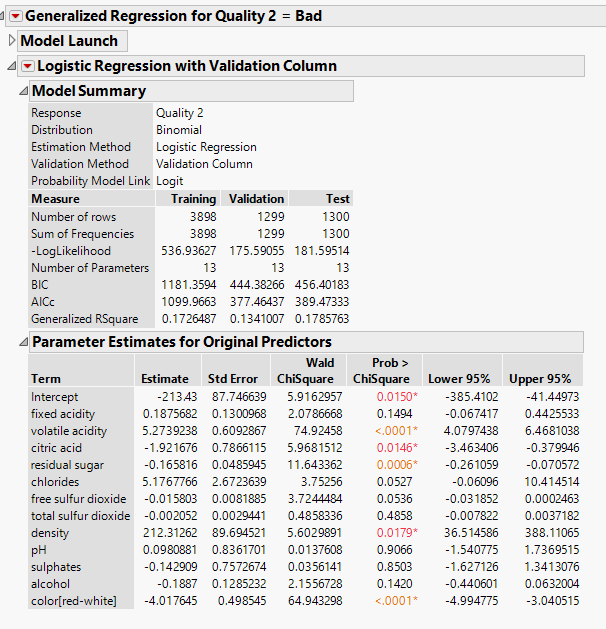


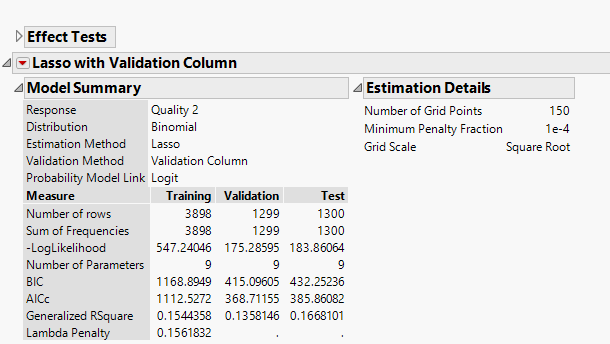


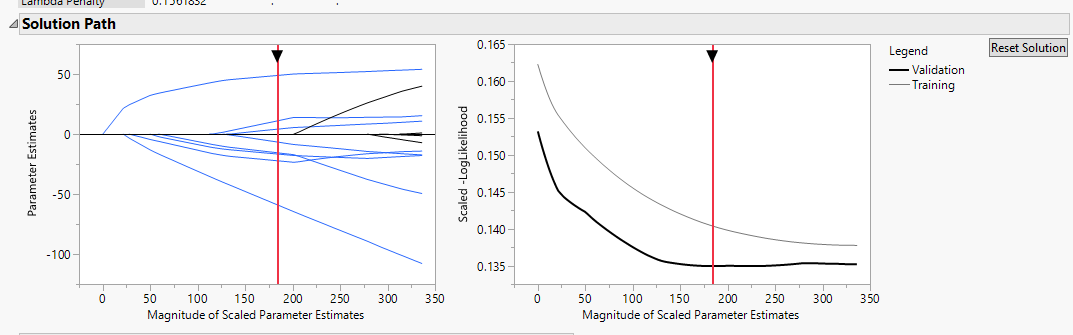


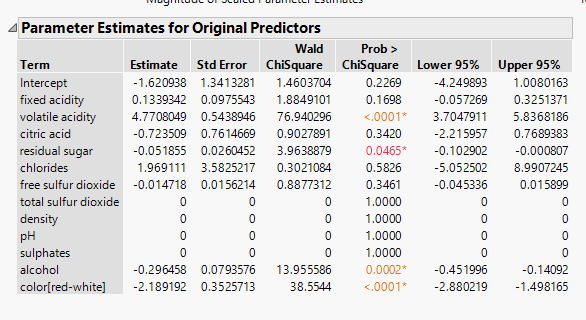
**Appendix B**

**General Regression Lasso**



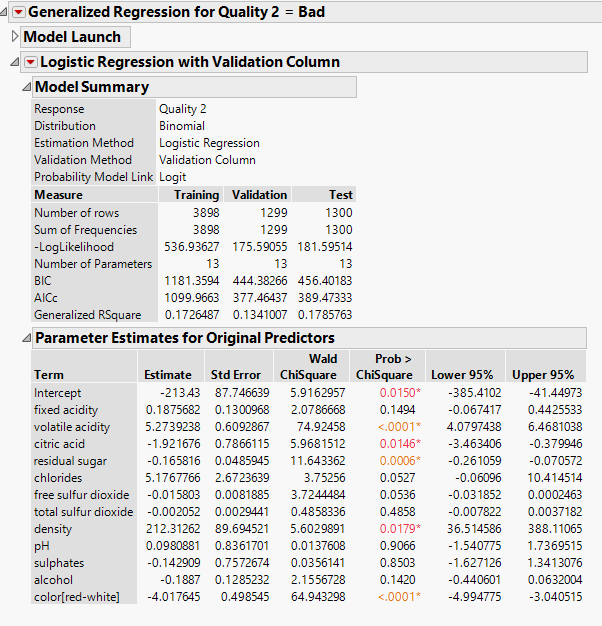


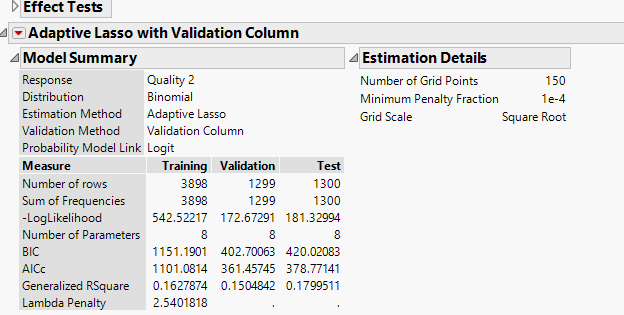


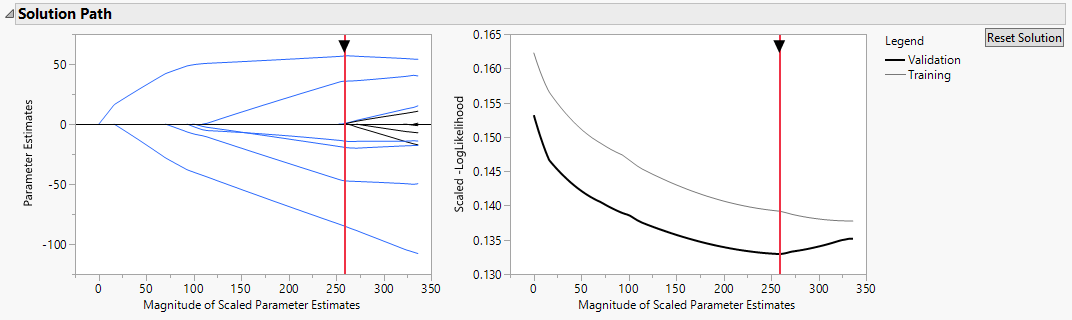


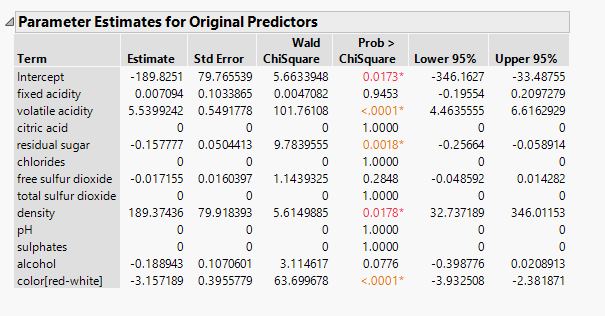
**Appendix C**

**General Regression Adaptive Lasso**



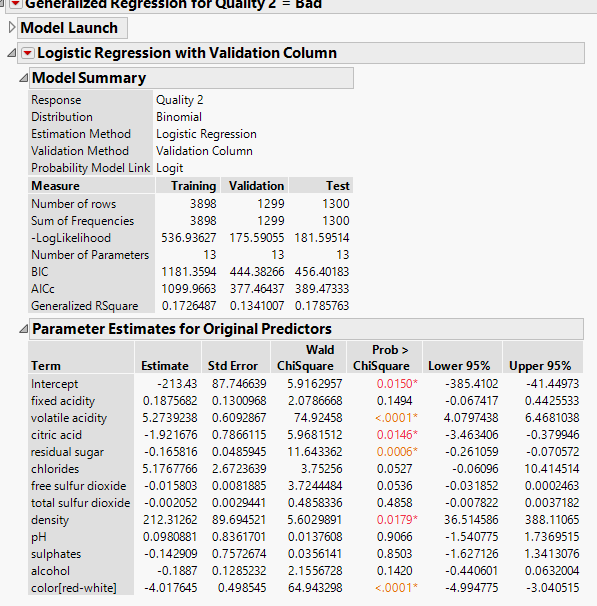


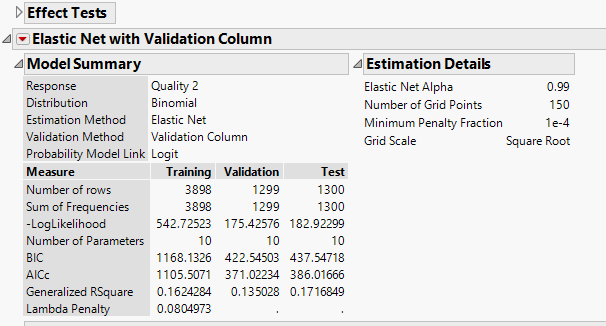


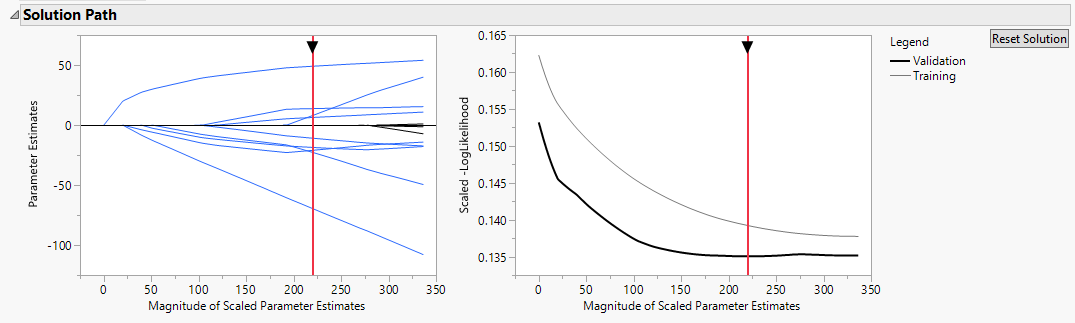


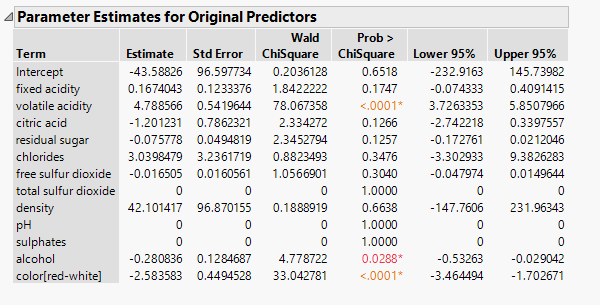
**Appendix D**

**General Regression Elastic Net**



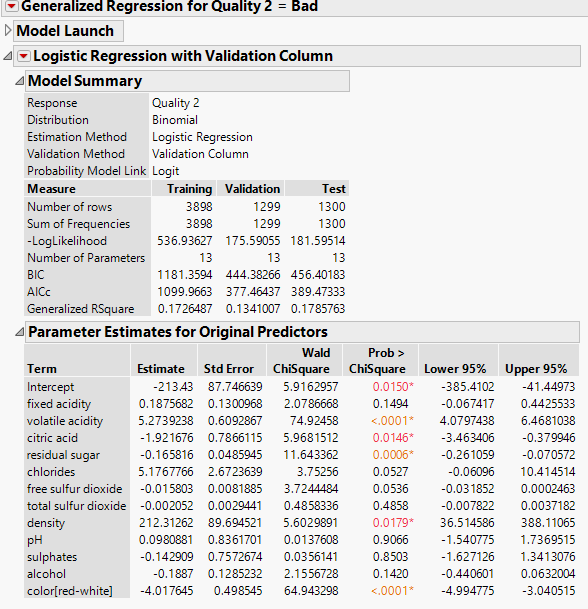


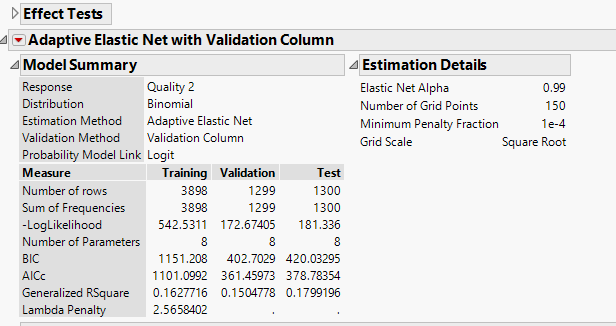


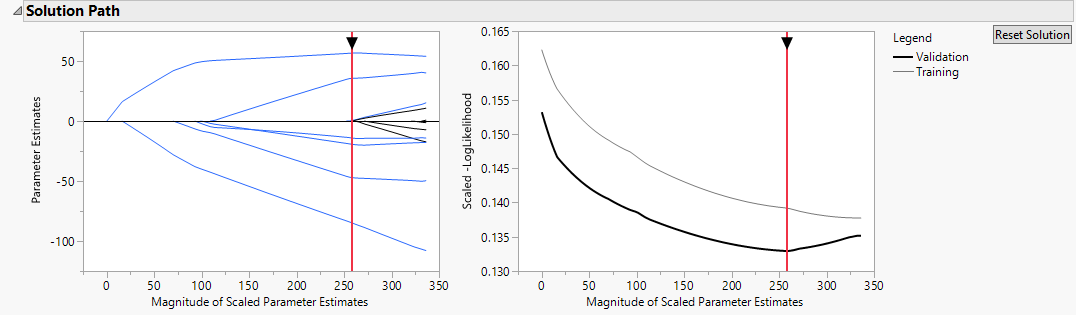


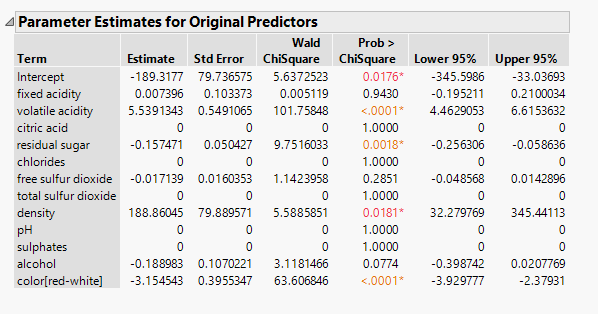
**Appendix E**

**General Regression Adaptive Elastic Net**



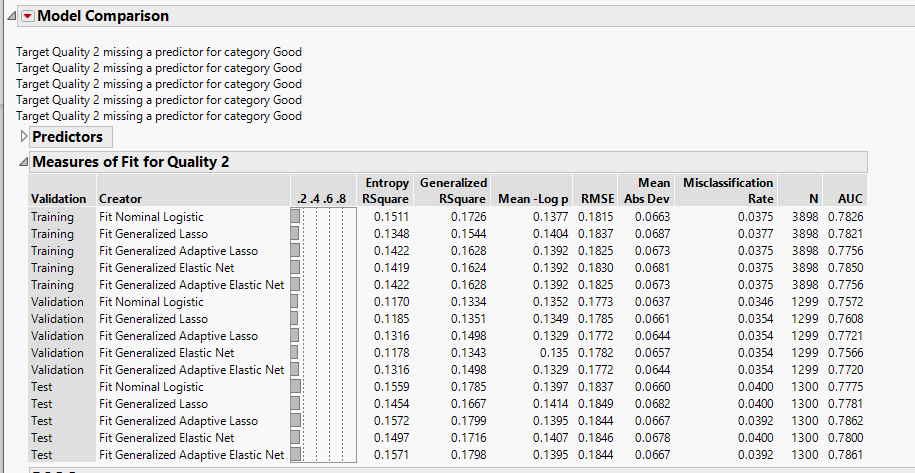






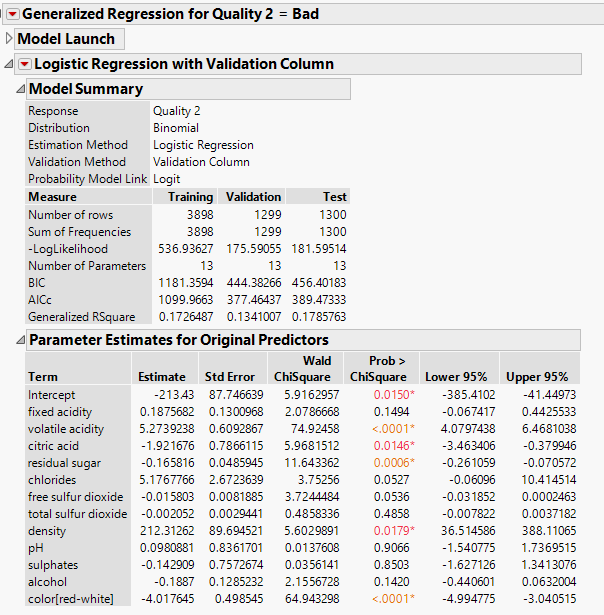
**Appendix F**

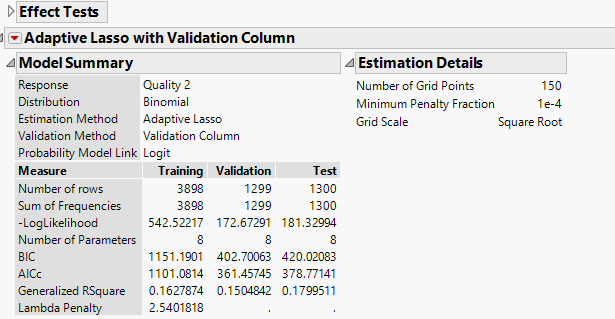
**Model Comparison**

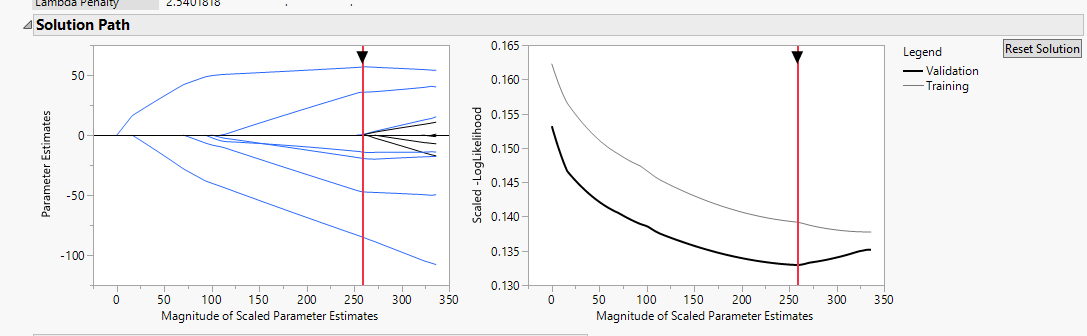


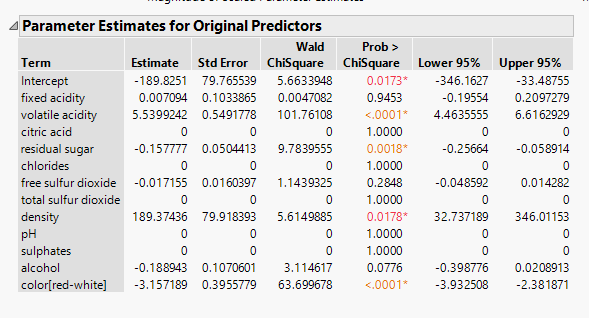
**Appendix G**

**Final Model Analysis Adative Lasso**



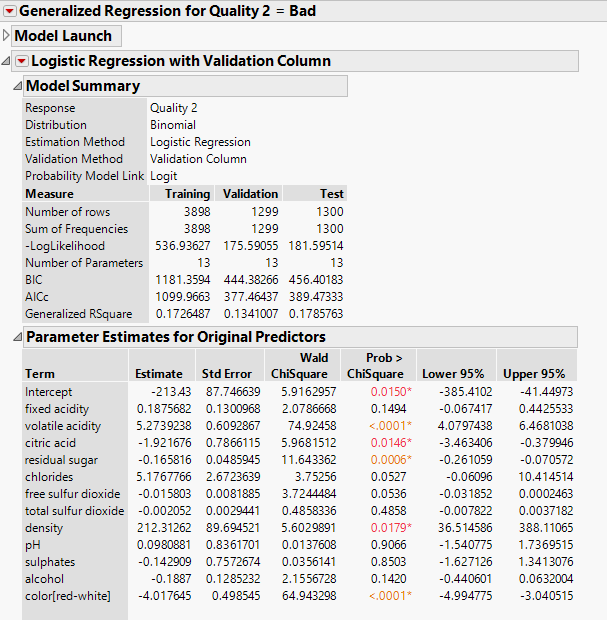


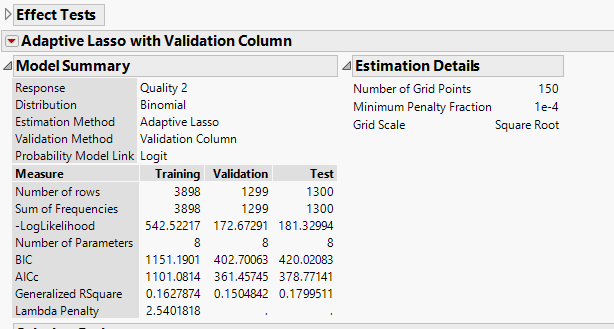


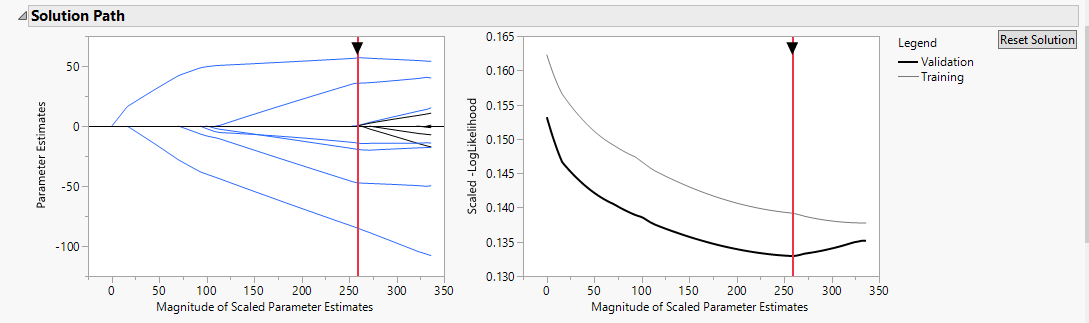


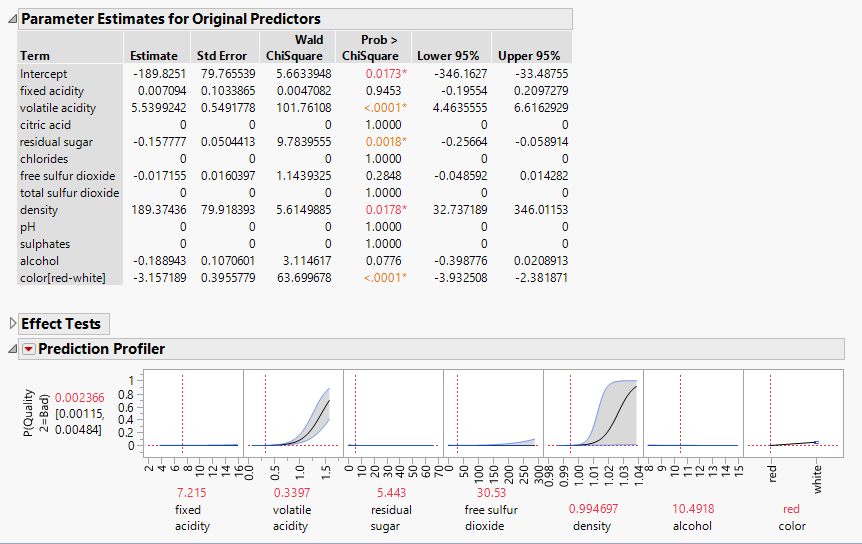
**Appendix H**

**Profiler Predictions**



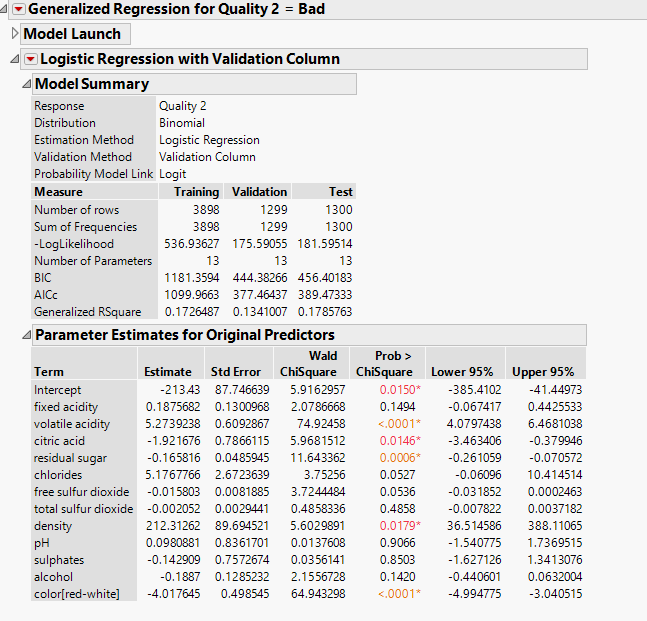


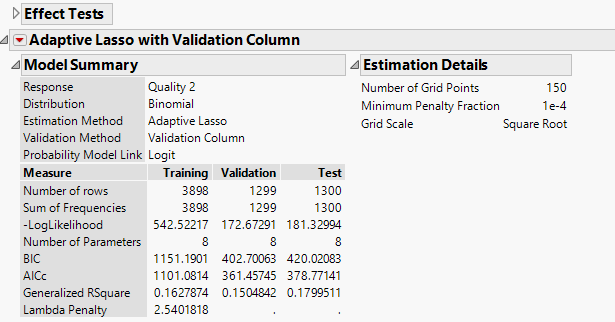


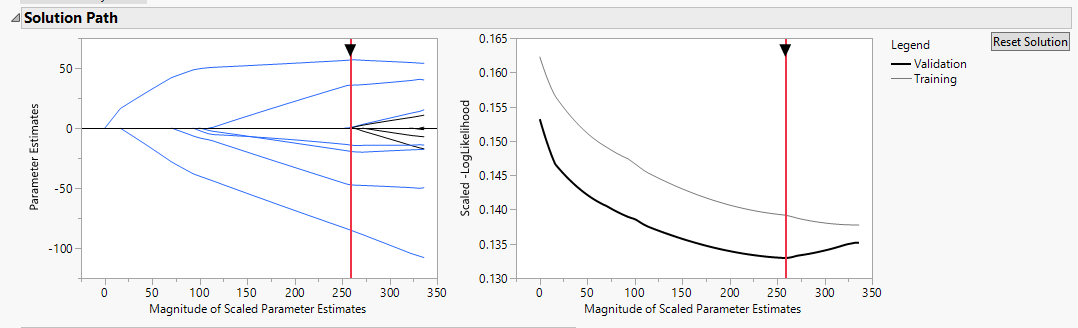


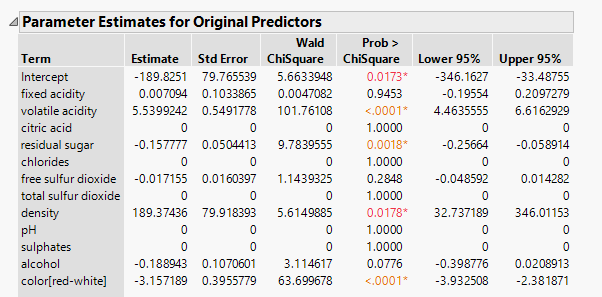
**Appendix I**

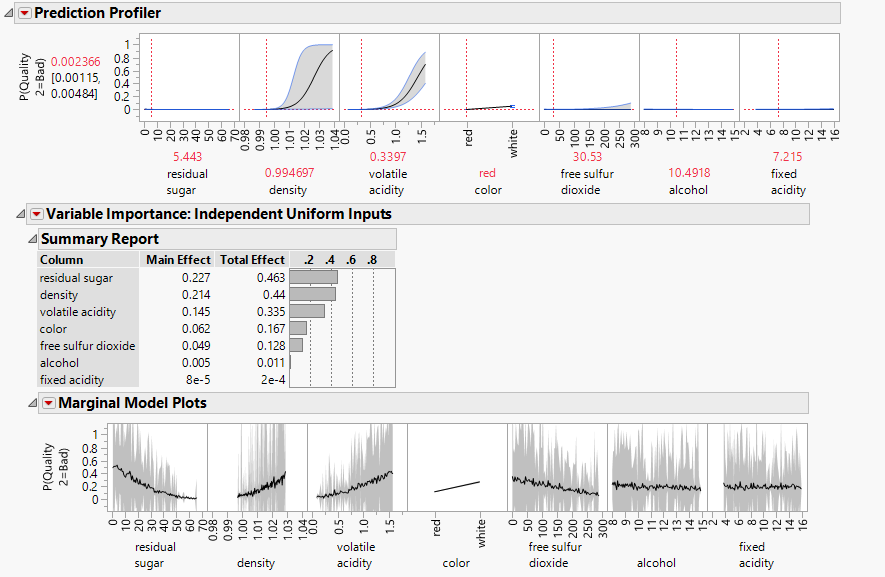
**Summary Report for Variable Importance**











**Appendix J**

Fixed acidity=8.9, Volatile acidity= .90, Citric Acid=.35, pH=3.20, Residual sugar=4.8, Chlorides=.09, Free sulfur dioxide=4, total sulfur dioxide=40, density=.99, sulphates=.65, alcohol=9.0, color=red

What is the prediction of the probability of this sample to be evaluated as BAD, according to your chosen model?

