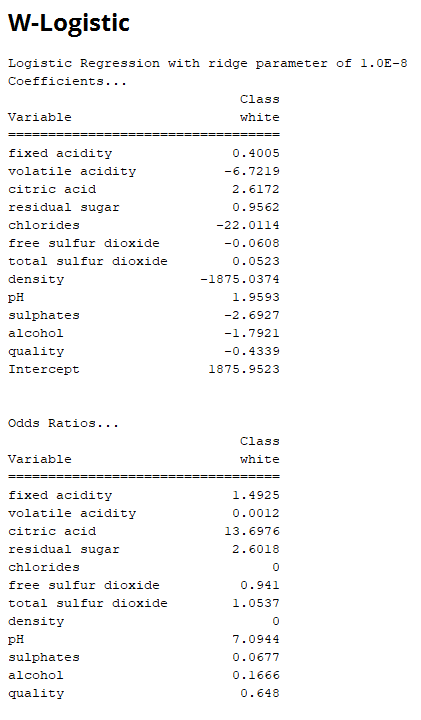
Nicholas Iudiciani

Prof. Deokar

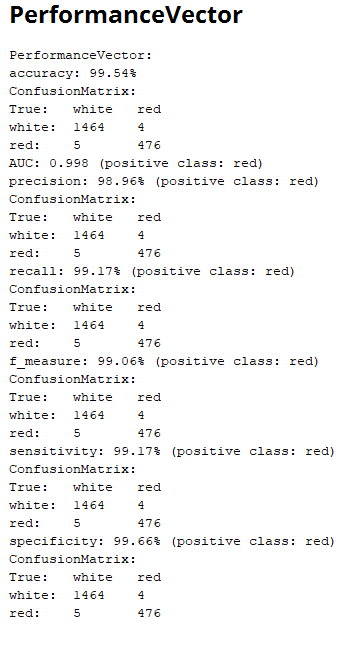
MIST4060 – Data Mining for Business Intel

Assignment 4

**ANALYSIS**

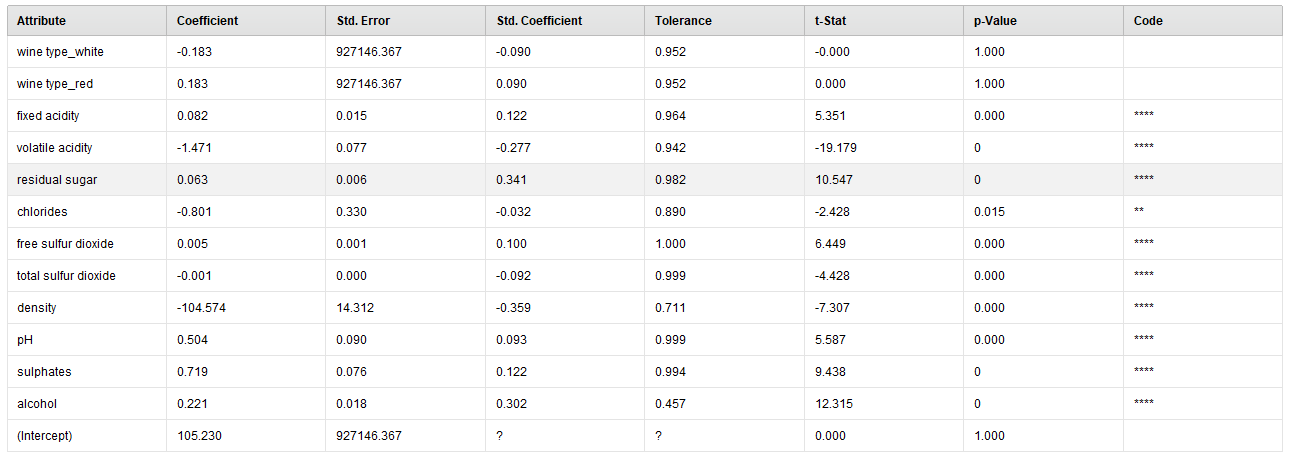
A)

B) For this W-Logistic model, the predicted class variable is wine of quality “White”. White becomes the logit (what is being predicted) and these coefficients show how good they would be at predicting the logit. The closer the numbers are to 0, the less influence that category has on predicting our wine type. Therefore the farther away they are from 0, the more influence they have, but I believe that the degree of each coefficient plays a role in its predictions. If it is greater than 0 (positive), then it will accurately predict the wine type positively, but if the coefficient is less than 0 (negative) then it will predict false negatives instead for the wine type. Density, chlorides, and volatile acidity would be the heaviest negative predictors, whereas citric acid and pH balance – though having far less influence than the negative influencers – would be the best positive predictors. As for the odds ratios, the ones that are greater than 1.0 (fixed acidity, citric acid, residual sugar, total sulfur dioxide, and pH balance) would increase the chances of predicting white wine quality. Those ratios that are equal to 1 will have no affect on the wine quality (which is none of them), and those that are less than will decrease the logit (volatile acidity, chlorides, free sulfur dioxide, density, alcohol, quality).

C)

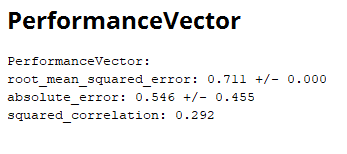
D) 99.54% accuracy means that the model correctly predicted almost 100% of all true and predicted examples of the batch, in this case the wine quality of 1469 white wines and 480 red wines. An AUC of 0.998 with red wine being the positive class indicates out of 1, the expected proportion of positive red wine predictions before a uniformly drawn random white (negative) prediction would be 0.998. Precision of 98.96% with red being the positive class indicates that the model predicted almost 99% of the actual positive examples of red wine as actually being of red wine quality. A recall and sensitivity of 99.17% indicates that the model correctly predicted almost all of the true positive examples relative to all positive examples in the batch. Since the denominator includes true positives, this means that the number of false negatives is very low – 4 to be exact, as opposed to 476 true positives. An f-measure of 99.06% merely shows that despite the presence of false positives and false negatives for this model, the number of true positives greatly outweighs everything else. Specificity of 99.66% with red wine as the positive class shows that out of 1469 negative examples of white wine, 1464 of them were true negative predictions of white wine.

**ANALYSIS 2**

A) 

B) For this model, we are predicting wine quality on a scale of 1-10 though the values range from 3-9 in this exampleset. Similar to W-logistic, the coefficients show the type and degree of influence on predicting the target variable which is wine quality. I would not use red and white wine typing as good predictors due to the standard error, we won’t look at those. Density has a strong negative degree of predicting power, whereas sulphates is the leader for positive prediction power, though it is less than +1.

C)



A root mean squared error of 0.711 is not an ideal number for this data set. A value of 0 would indicate that the example set is a perfect fit for the data. This RMSE is used to show the degree of differences in a model’s estimation versus all of the observed values. Absolute error is similar to RMSE, except RMSE squares the two values to avoid getting negative numbers, whereas absolute error uses absolute values instead of merely squaring. This is why there is a 0.546 +/- 0.455 window here. Basically the absolute error can be between 0.091 and 1. |.546-.455| = .091 and .546+.455 = ~1, which shows that this model can be very unreliable in its predictions. The squared correlation coefficient of 0.292 shows that the degree of wine quality and these other variables is more positively random than it is definitive. Basically, correlation coefficient or R squared should be closer to 1 for a strong positive correlation or closer to -1 for a strong negative correlation. But since it is less than 0.3, then this data can see a vague positive correlation between wine quality and the other variables, though it is more random than defined.