Predictive Modeling

For each country’s sample, ordinary logistic regression, stepwise logistic regression, and neural network methods are used to build predictive models to predict wine points. Ideally, random forest models would be implemented as well, but are too computationally intensive for this analysis. Historically, wines with points 90 and above are considered outstanding or even classic/timeless wines. For the purpose of this analysis, points scores from 80-89 are classified as “No” and scores from 91-100 are classified as “Yes.” The variable is renamed “Exceptional.” This classification converts the wine points variable to a binary variable and allows for the usage of ordinary logistic regression, stepwise logistic regression, and neural network models. Ordinary logistic regression will be used as the benchmark model. This model will use the variable, price to predict Exceptional. Price is the chosen variable for the benchmark model because it is the most definitive predictor for each country sample. The AIC value of the ordinary logistic regression model will be compared to other models. The Akaike information criterion (AIC) is an estimator of the relative quality of statistical models for a given set of data. Forward and backward stepwise regression models are effective for fine-tuning the model by adding or removing variables, respectively until predictive power, or R-squared, no longer increases. Predictor variables potentially included in the model include price, province, and variety. This, in combination with cross-validation, develops models that can be used for further analysis.

Each model will, initially, be run on the training set, where significant variables and model accuracy will be determined. Additional K-fold cross-validation (10 folds) will be applied to neural network models to fine tune the model. The model developed using the training data sets will be introduced to the test data set. Model accuracy will then be determined on the test data set and compared to the training set accuracy to determine overall fit. Accuracy, sensitivity, specificity, and overall fit will be the main determinants in deciding which statistical model will be the champion model. The target model will have the highest accuracy measure possible, while also having balanced sensitivity and specificity. Accuracy is the proportion of total correct observations to incorrect observations. Sensitivity is the proportion of observed positives that were predicted to be positive (true positives), whereas specificity is the proportion of observed negatives that were predicted to be negative (true negatives).

Logistic regression models will undergo threshold selection to balance sensitivity and specificity. Using the logistic regression model with the lowest AIC, predictions will be developed that predict probability of the positive response of the variable, Exceptional. The positive response is “Yes.” An ROCR curve is developed to visualize how sensitivity and (1 – specificity) changes for each threshold. The threshold with the most balanced sensitivity and specificity is selected. Next, a confusion matrix is developed that shows the amount true and false observations with the established threshold. Now the accuracy of the model is determined. Ideally, accuracy, sensitivity, and specificity are as close to 1 as possible. This model is then introduced to the test data set, where accuracy is again determined. The two accuracy measures are compared to determine the validity and fit of the model.

For the neural network models, the variables price and province are included to predict Exceptional. The addition of more variables is too computationally intensive with the tools and time limitations at hand at hand. 10 K-fold cross-validation is set up in addition to the training and test data split. A tuning grid is established with a specific size of hidden layers and a decay parameter. The hidden layers are used to develop weights so that the model output can successfully predict the class of newly introduced data. The size of hidden layers will be determined by the total number of variables plus their respective layers. Normally there should be an element of parameter tuning, whereas a tuning grid searches through hidden layer sizes of 1 to an amount determined by the total number of variables plus their respective factor levels. The tuning grid, in turn, selects the best hidden layer size. This tuning grid method is unfortunately too time-consuming and intensive to use for this analysis. In addition, some of the country’s data sets are too large and a certain size of hidden layers will be too computationally intensive. When faced with this issue, the largest size of hidden layer possible will be selected via trial and error. A delay value of 0.1 is a typical value to reduce overfitting. The nnet package within R is used to develop neural network models.

**US Wine**

**Benchmark Model**

Call:

glm(formula = Exceptional ~ price, family = "binomial", data = UStrain)

Deviance Residuals:

Min 1Q Median 3Q Max

-5.026 -0.821 -0.599 1.056 2.137

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -2.4496054 0.0193249 -126.8 <2e-16 \*\*\*

price 0.0548290 0.0005126 107.0 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 107294 on 81482 degrees of freedom

Residual deviance: 90021 on 81481 degrees of freedom

AIC: 90025

Number of Fisher Scoring iterations: 5

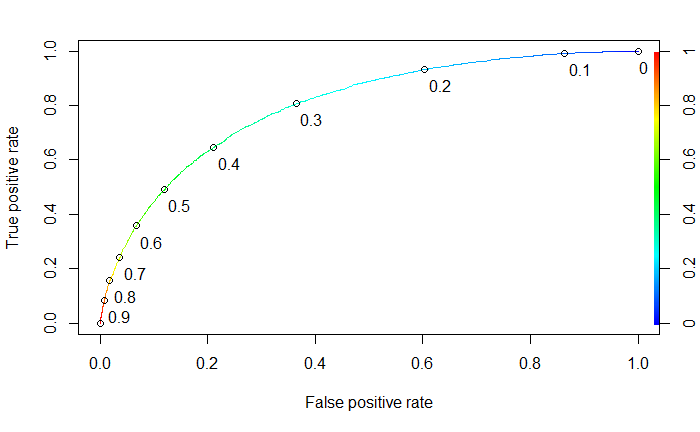
The ordinary logistic regression model yields an AIC of 90025. Price is a significant predictor of Exceptional and has a positive coefficient of 0.0548290.

**Stepwise Logistic Regression**

Forward and backward stepwise regression yielded identical AIC of 86535 and included every variable from the training set. This model defeats the benchmark model and will be used to develop predictions. Price and provinces of Colorado, Michigan, Texas, Virginia, and Washington are determined as significant. Price has a positive coefficient of 0.059363 which suggests that higher prices lead to a greater probability of wines being exceptional. Every significant province has a negative coefficient except for Washington.

The threshold that balances sensitivity and specificity for this model is 0.3549503. At this threshold, sensitivity is 0.7210672 and specificity is 0.7300961.

ROCR:



The confusion matrix for the training data set is as follows:

FALSE TRUE

No 37543 13879

Yes 8385 21676

Accuracy is measured at 0.7267651 for the training set. The confusion matrix for the test data set is as follows:

FALSE TRUE

No 16259 5779

Yes 3621 9262

Accuracy is measured at 0.730821 for the test set. There is no sign of overfitting as the accuracy measures for training and test sets are very similar.

**Neural Network**

The data summary of the US training set for neural networks is as follows:

Classes ‘tbl\_df’, ‘tbl’ and 'data.frame': 81483 obs. of 3 variables:

$ price : num 235 90 65 65 60 48 48 185 90 325 ...

$ province : Factor w/ 27 levels "America","Arizona",..: 3 3 20 20 3 20 20 20 20 3 ...

$ Exceptional: Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...

price province Exceptional

Min. : 4.00 California:56254 No :51422

1st Qu.: 20.00 Washington:12830 Yes:30061

Median : 29.00 Oregon : 6976

Mean : 34.98 New York : 3543

3rd Qu.: 44.00 Virginia : 895

Max. :2013.00 Idaho : 234

(Other) : 751

The neural network model is created with a hidden layer size of 30 and a decay of 0.1.

Training

Confusion Matrix and Statistics

Reference

Prediction No Yes

No 43635 13914

Yes 7787 16147

Accuracy : 0.7337

95% CI : (0.7306, 0.7367)

No Information Rate : 0.6311

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.4028

Mcnemar's Test P-Value : < 2.2e-16

Sensitivity : 0.5371

Specificity : 0.8486

Pos Pred Value : 0.6746

Neg Pred Value : 0.7582

Prevalence : 0.3689

Detection Rate : 0.1982

Detection Prevalence : 0.2937

Balanced Accuracy : 0.6929

'Positive' Class : Yes

According to the confusion matrix and statistics output from the neural network model developed on the training set, accuracy is 0.7337, no information rate is 0.6311, p-value is < 0.05, sensitivity is 0.5371, and specificity is 0.8486.

Test

Confusion Matrix and Statistics

Reference

Prediction No Yes

No 18666 5932

Yes 3372 6951

Accuracy : 0.7336

95% CI : (0.7289, 0.7382)

No Information Rate : 0.6311

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.4032

Mcnemar's Test P-Value : < 2.2e-16

Sensitivity : 0.5395

Specificity : 0.8470

Pos Pred Value : 0.6734

Neg Pred Value : 0.7588

Prevalence : 0.3689

Detection Rate : 0.1990

Detection Prevalence : 0.2956

Balanced Accuracy : 0.6933

'Positive' Class : Yes

The output developed from introducing the model to the test data set is very similar to the training data set. There is no sign of overfitting, but sensitivity and specificity are not very balanced.

Accuracy measures for logistic regression and neural network are similar. The logistic regression model involves one more variable and balances sensitivity and specificity. Because of this, the logistic regression model will act as the champion model for the US wine data set.

**French Wine**

**Benchmark Model**

Call:

glm(formula = Exceptional ~ price, family = "binomial", data = Francetrain)

Deviance Residuals:

Min 1Q Median 3Q Max

-8.4904 -0.7392 -0.6390 0.8935 1.9410

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -2.0927960 0.0303639 -68.92 <2e-16 \*\*\*

price 0.0467501 0.0008123 57.55 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 30600 on 22793 degrees of freedom

Residual deviance: 23861 on 22792 degrees of freedom

AIC: 23865

Number of Fisher Scoring iterations: 7

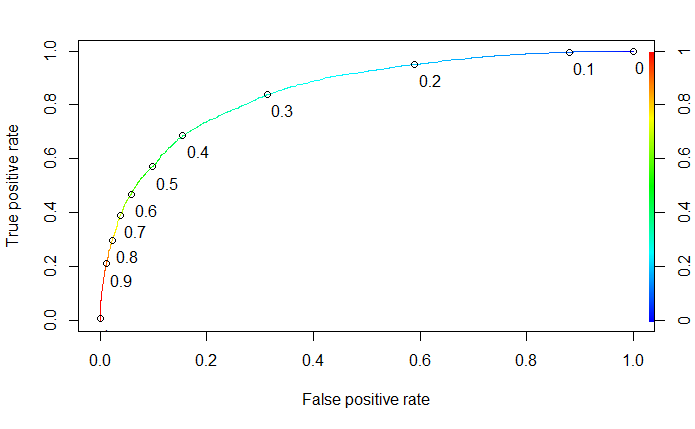
The ordinary logistic regression model yields an AIC of 23865. Price is a significant predictor of Exceptional and has a positive coefficient of 0.04657501.

**Stepwise Logistic Regression**

Forward and backward stepwise regression yielded identical AIC of 22581 and included every variable from the training set. This model defeats the benchmark model and will be used to develop predictions. Price and provinces of Beaujolais, Bordeaux, Burgundy, Champagne, France other, Languedoc-Roussilon, Loire Valley, Rha’ne Valley, and Southwest France are determined as significant. Also, varieties Pinot Blanc and Provence white blend are significant. Price has a positive coefficient of 0.054423 which suggests that higher prices lead to a greater probability of wines being exceptional. Every significant province has a negative.

The threshold that balances sensitivity and specificity for this model is 0.3429604. At this threshold, sensitivity is 0.7656059 and specificity is 0.7716878.

ROCR:



The confusion matrix for the training data set is as follows:

FALSE TRUE

No 10630 3145

Yes 2117 6902

Accuracy is measured at 0.7691498 for the training set. The confusion matrix for the test data set is as follows:

FALSE TRUE

No 4674 1229

Yes 923 2941

Accuracy is measured at 0.7796662 for the test set. There is no sign of overfitting as the accuracy measures for training and test sets are close.

**Neural Network**

The data summary of the US training set for neural networks is as follows:

Classes ‘tbl\_df’, ‘tbl’ and 'data.frame': 22794 obs. of 3 variables:

$ price : num 290 22 60 69 68 42 28 45 38 18 ...

$ province : Factor w/ 11 levels "Alsace","Beaujolais",..: 11 6 10 10 4 8 8 4 4 7 ...

$ Exceptional: Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...

price province Exceptional

Min. : 5.00 Bordeaux :4666 No :13775

1st Qu.: 16.00 Burgundy :4538 Yes: 9019

Median : 25.00 Alsace :2422

Mean : 43.25 Loire Valley :2091

3rd Qu.: 45.00 Southwest France:1809

Max. :3300.00 Champagne :1734

(Other) :5534

The neural network model is created with a hidden layer size of 15 and a decay of 0.1.

Training

Confusion Matrix and Statistics

Reference

Prediction No Yes

No 11538 2805

Yes 2237 6214

Accuracy : 0.7788

95% CI : (0.7734, 0.7842)

No Information Rate : 0.6043

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.5324

Mcnemar's Test P-Value : 1.404e-15

Sensitivity : 0.6890

Specificity : 0.8376

Pos Pred Value : 0.7353

Neg Pred Value : 0.8044

Prevalence : 0.3957

Detection Rate : 0.2726

Detection Prevalence : 0.3708

Balanced Accuracy : 0.7633

'Positive' Class : Yes

According to the confusion matrix and statistics output from the neural network model developed on the training set, accuracy is 0.7788, no information rate is 0.6311, p-value is < 0.05, sensitivity is 0.6890, and specificity is 0.8376.

Test

Confusion Matrix and Statistics

Reference

Prediction No Yes

No 4973 1177

Yes 930 2687

Accuracy : 0.7843

95% CI : (0.776, 0.7924)

No Information Rate : 0.6044

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.5439

Mcnemar's Test P-Value : 8.357e-08

Sensitivity : 0.6954

Specificity : 0.8425

Pos Pred Value : 0.7429

Neg Pred Value : 0.8086

Prevalence : 0.3956

Detection Rate : 0.2751

Detection Prevalence : 0.3703

Balanced Accuracy : 0.7689

'Positive' Class : Yes

The output developed from introducing the model to the test data set is very similar to the training data set. There is no sign of overfitting, but sensitivity and specificity are not very balanced.

Accuracy measures for logistic regression and neural network are similar. The logistic regression model involves one more variable and balances sensitivity and specificity. Because of this, the logistic regression model will act as the champion model for the France wine data set.

**Italy Wine**

**Benchmark Model**

Call:

glm(formula = Exceptional ~ price, family = "binomial", data = Italytrain)

Deviance Residuals:

Min 1Q Median 3Q Max

-5.6952 -0.6535 -0.5126 0.7060 2.1708

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -2.8429645 0.0346253 -82.11 <2e-16 \*\*\*

price 0.0586475 0.0008441 69.48 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 32214 on 24989 degrees of freedom

Residual deviance: 23903 on 24988 degrees of freedom

AIC: 23907

Number of Fisher Scoring iterations: 5

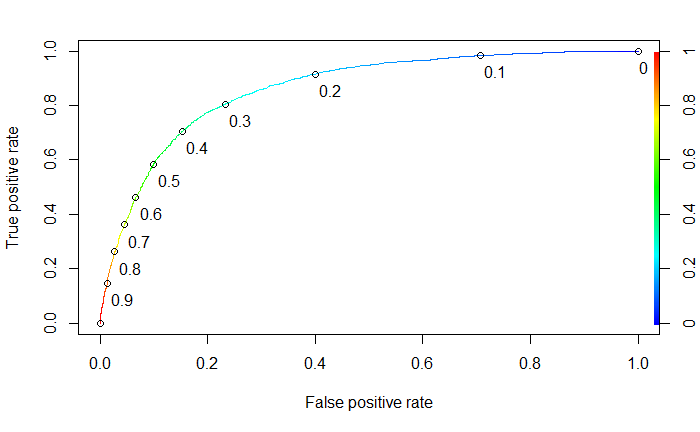
The ordinary logistic regression model yields an AIC of 23907. Price is a significant predictor of Exceptional and has a positive coefficient of 0.0586475.

**Stepwise Logistic Regression**

Forward and backward stepwise regression yielded identical AIC of 22213 and included every variable from the training set. This model defeats the benchmark model and will be used to develop predictions. Price, several provinces, and several varieties are determined as significant predictors. Significant Provinces include Northeastern Italy, Sicily & Sardinia, Southern Italy, Tuscany, and Veneto. Significant varieties include Aleatico, Arneis, Barbera, Bordeaux-style Red Blend, Cabernet Sauvignon, Cabernet-Sauvignon-Merlot, Cannonau, Carignano, Chardonnay, Corcina, Rondinella, Molinara, Dolcetto, Falanghina, Frappato, Friulano, Gargenega, Gewarztraminer, Glera, Grillo, Insolia, Kerner, Lagrein, Lambrusco, Lambrusco Grasparossa, Maller-Thurgau, Malvasia, Marzemino, Merlot, Montepulciano, Moscato, Nebbiolo, Negroamaro, Nerello Mascalese, Nero D-Avola, Nosiola, Pallagrello, Passerina, Pecorino, Piedirosso, Pinot Bianco, Pinot Grigio, Pinot Nero, Primitivo, Prosecco, Prugnolo Gentile, Red Blend, Ribolla Gialla, Riesling, Rosa, Rosato, Sagrantino, Sangiovese, Sangiovese Grosso, Sauvignon, Sauvignon Blanc, Schiava, Sparkling Blend, Sylvaner, Syrah, Tocai Friulano, Trebbiano, Turbiana, Uva di Troia, Verdicchio, Vermentino, Vernaccia, and Viognier. Price has a positive coefficient of 0.0519319 which suggests that higher prices lead to a greater probability of wines being exceptional. Provinces with positive coefficients include Northeastern Italy, Sicily & Sardinia, and Tuscany.

The threshold that balances sensitivity and specificity for this model is 0.337666. At this threshold, sensitivity is 0.7712895 and specificity is 0.8056727.

ROCR:



The confusion matrix for the training data set is as follows:

FALSE TRUE

No 13180 3179

Yes 1975 6656

Accuracy is measured at 0.7937575 for the training set. The confusion matrix for the test data set is as follows:

FALSE TRUE

No 5619 1391

Yes 875 2823

Accuracy is measured at 0.7883825 for the test set. There is no sign of overfitting as the accuracy measures for training and test sets are close.

**Neural Network**

The data summary of the US training set for neural networks is as follows:

Classes ‘tbl\_df’, ‘tbl’ and 'data.frame': 24990 obs. of 3 variables:

$ price : num 80 23 29 39 30 50 100 60 59 85 ...

$ province : Factor w/ 10 levels "Central Italy",..: 4 9 9 9 9 9 9 9 6 6 ...

$ Exceptional: Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...

price province Exceptional

Min. : 5.00 Tuscany :7723 No :16359

1st Qu.: 18.00 Piedmont :4215 Yes: 8631

Median : 27.00 Veneto :3803

Mean : 38.43 Northeastern Italy:2674

3rd Qu.: 50.00 Sicily & Sardinia :2513

Max. :900.00 Central Italy :1668

(Other) :2394

The neural network model is created with a hidden layer size of 15 and a decay of 0.1.

Training

Confusion Matrix and Statistics

Reference

Prediction No Yes

No 13728 2578

Yes 2631 6053

Accuracy : 0.7916

95% CI : (0.7865, 0.7966)

No Information Rate : 0.6546

P-Value [Acc > NIR] : <2e-16

Kappa : 0.5397

Mcnemar's Test P-Value : 0.4712

Sensitivity : 0.7013

Specificity : 0.8392

Pos Pred Value : 0.6970

Neg Pred Value : 0.8419

Prevalence : 0.3454

Detection Rate : 0.2422

Detection Prevalence : 0.3475

Balanced Accuracy : 0.7702

'Positive' Class : Yes

According to the confusion matrix and statistics output from the neural network model developed on the training set, accuracy is 0.7916, no information rate is 0.6546, p-value is < 0.05, sensitivity is 0.7013, and specificity is 0.8392.

Test

Confusion Matrix and Statistics

Reference

Prediction No Yes

No 5845 1121

Yes 1165 2577

Accuracy : 0.7865

95% CI : (0.7786, 0.7942)

No Information Rate : 0.6547

P-Value [Acc > NIR] : <2e-16

Kappa : 0.5292

Mcnemar's Test P-Value : 0.3685

Sensitivity : 0.6969

Specificity : 0.8338

Pos Pred Value : 0.6887

Neg Pred Value : 0.8391

Prevalence : 0.3453

Detection Rate : 0.2407

Detection Prevalence : 0.3495

Balanced Accuracy : 0.7653

'Positive' Class : Yes

The output developed from introducing the model to the test data set is very similar to the training data set. There is no sign of overfitting, but sensitivity and specificity are not very balanced.

Accuracy measures for logistic regression and neural network are similar. The logistic regression model involves one more variable, variety, which has many significant predictors, and balances sensitivity and specificity. Because of this, the logistic regression model will act as the champion model for the Italy wine data set.

**Spain Wine**

**Benchmark Model**

Call:

glm(formula = Exceptional ~ price, family = "binomial", data = Spaintrain)

Deviance Residuals:

Min 1Q Median 3Q Max

-6.6503 -0.5470 -0.4643 -0.4222 2.2189

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -2.82154 0.04987 -56.58 <2e-16 \*\*\*

price 0.04987 0.00143 34.87 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 10546.2 on 10313 degrees of freedom

Residual deviance: 8294.9 on 10312 degrees of freedom

AIC: 8298.9

Number of Fisher Scoring iterations: 6

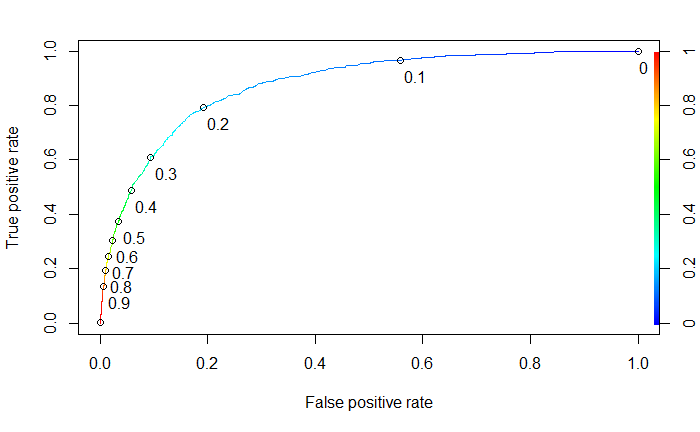
The ordinary logistic regression model yields an AIC of 8298.9. Price is a significant predictor of Exceptional and has a positive coefficient of 0.04987.

**Stepwise Logistic Regression**

Forward and backward stepwise regression yielded identical AIC of 7786.9 and included every variable from the training set. This model defeats the benchmark model and will be used to develop predictions. Price and the province, Central Spain are determined as significant predictors. Price has a positive coefficient of 0.046139 which suggests that higher prices lead to a greater probability of wines being exceptional. Central Spain has a coefficient of -1.633915.

The threshold that balances sensitivity and specificity for this model is 0.1861174. At this threshold, sensitivity is 0.8125874 and specificity is 0.7894479.

ROCR:



The confusion matrix for the training data set is as follows:

FALSE TRUE

No 6449 1720

Yes 408 1737

Accuracy is measured at 0.7936785 for the training set. The confusion matrix for the test data set is as follows:

FALSE TRUE

No 2762 738

Yes 173 746

Accuracy is measured at 0.7938448 for the test set. There is no sign of overfitting as the accuracy measures for training and test sets are close.

**Neural Network**

The data summary of the US training set for neural networks is as follows:

Classes ‘tbl\_df’, ‘tbl’ and 'data.frame': 10314 obs. of 3 variables:

$ price : num 65 110 79 17 26 17 44 30 22 18 ...

$ province : Factor w/ 8 levels "Andalucia","Catalonia",..: 6 6 6 4 1 4 6 6 2 2 ...

$ Exceptional: Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 1 1 1 1 ...

price province Exceptional

Min. : 4.00 Northern Spain:6040 No :8169

1st Qu.: 12.00 Catalonia :1747 Yes:2145

Median : 18.00 Central Spain : 844

Mean : 27.93 Galicia : 681

3rd Qu.: 29.00 Levante : 625

Max. :770.00 Andalucia : 273

(Other) : 104

The neural network model is created with a hidden layer size of 12 and a decay of 0.1.

Training

Confusion Matrix and Statistics

Reference

Prediction No Yes

No 7552 997

Yes 617 1148

Accuracy : 0.8435

95% CI : (0.8364, 0.8505)

No Information Rate : 0.792

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.4918

Mcnemar's Test P-Value : < 2.2e-16

Sensitivity : 0.5352

Specificity : 0.9245

Pos Pred Value : 0.6504

Neg Pred Value : 0.8834

Prevalence : 0.2080

Detection Rate : 0.1113

Detection Prevalence : 0.1711

Balanced Accuracy : 0.7298

'Positive' Class : Yes

According to the confusion matrix and statistics output from the neural network model developed on the training set, accuracy is 0.8435, no information rate is 0.792, p-value is < 0.05, sensitivity is 0.5352, and specificity is 0.9245.

Test

Confusion Matrix and Statistics

Reference

Prediction No Yes

No 3255 436

Yes 245 483

Accuracy : 0.8459

95% CI : (0.8349, 0.8564)

No Information Rate : 0.792

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.4934

Mcnemar's Test P-Value : 3.318e-13

Sensitivity : 0.5256

Specificity : 0.9300

Pos Pred Value : 0.6635

Neg Pred Value : 0.8819

Prevalence : 0.2080

Detection Rate : 0.1093

Detection Prevalence : 0.1647

Balanced Accuracy : 0.7278

'Positive' Class : Yes

The output developed from introducing the model to the test data set is very similar to the training data set. There is no sign of overfitting, but sensitivity and specificity are not very balanced.

Accuracy measures are substantially better for the neural network model than for the logistic regression model. Because of the better accuracy and the lack of significant predictors with the logistic regression model, the neural network model will act as the champion model for the Spain wine data set.

**Chile Wine**

**Benchmark Model**

Call:

glm(formula = Exceptional ~ price, family = "binomial", data = Chiletrain)

Deviance Residuals:

Min 1Q Median 3Q Max

-6.4683 -0.4111 -0.3317 -0.3042 2.4885

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -3.93774 0.08389 -46.94 <2e-16 \*\*\*

price 0.08878 0.00309 28.73 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 5600.2 on 7127 degrees of freedom

Residual deviance: 4009.1 on 7126 degrees of freedom

AIC: 4013.1

Number of Fisher Scoring iterations: 6

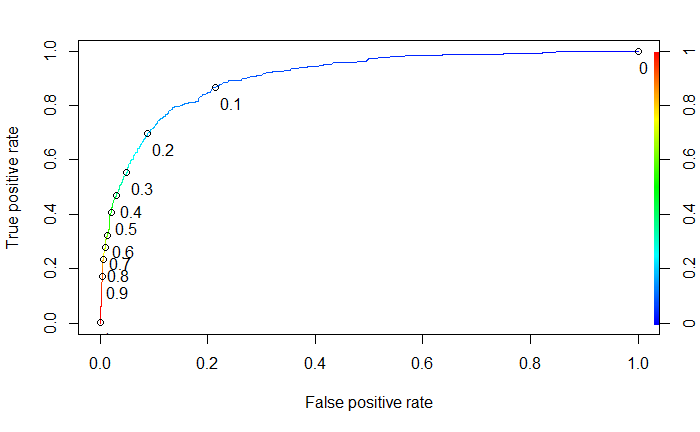
The ordinary logistic regression model yields an AIC of 4013.1. Price is a significant predictor of Exceptional and has a positive coefficient of 0.08878.

**Stepwise Logistic Regression**

Forward and backward stepwise regression yielded identical AIC of 3829.6 and included every variable from the training set. This model defeats the benchmark model and will be used to develop predictions. Price, several provinces, and several varieties are determined as significant predictors. Significant Provinces include Aconcagua Valley, Central Valley, Elqui Valley, and Repel Valley-Casablanca Valley. Significant varieties include Cabernet Sauvignon, Cabernet Sauvignon-Carmena re, Cabernet Sauvignon-Syrah, Carmena re, Chardonnay, Malbec, Merlot, Pinot Noir, Sauvignon Blanc, and Shiraz. Price has a positive coefficient of 0.072781 which suggests that higher prices lead to a greater probability of wines being exceptional. All other coefficients are negative.

The threshold that balances sensitivity and specificity for this model is 0.1307330. At this threshold, sensitivity is 0.8054679 and specificity is 0.8473369.

ROCR:



The confusion matrix for the training data set is as follows:

FALSE TRUE

No 5234 943

Yes 185 766

Accuracy is measured at 0.8417508 for the training set. The confusion matrix for the test data set is as follows:

FALSE TRUE

No 2304 342

Yes 71 336

Accuracy is measured at 0.8647232 for the test set. There is no sign of overfitting as the accuracy measures for training and test sets are close.

**Neural Network**

The data summary of the US training set for neural networks is as follows:

Classes ‘tbl\_df’, ‘tbl’ and 'data.frame': 7128 obs. of 3 variables:

$ price : num 30 19 30 20 20 20 24 30 100 17 ...

$ province : Factor w/ 44 levels "Aconcagua Costa",..: 19 16 16 28 9 6 34 16 16 22 ...

$ Exceptional: Factor w/ 2 levels "No","Yes": 1 1 2 1 1 1 2 2 2 1 ...

price province Exceptional

Min. : 5.00 Colchagua Valley :1427 No :6177

1st Qu.: 11.00 Maipo Valley :1065 Yes: 951

Median : 14.00 Casablanca Valley: 971

Mean : 19.95 Maule Valley : 678

3rd Qu.: 20.00 Central Valley : 626

Max. :400.00 CuricÃ³ Valley : 395

(Other) :1966

The neural network model is created with a hidden layer size of 21 and a decay of 0.1.

Training

Confusion Matrix and Statistics

Reference

Prediction No Yes

No 6018 535

Yes 159 416

Accuracy : 0.9026

95% CI : (0.8955, 0.9094)

No Information Rate : 0.8666

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.4944

Mcnemar's Test P-Value : < 2.2e-16

Sensitivity : 0.43743

Specificity : 0.97426

Pos Pred Value : 0.72348

Neg Pred Value : 0.91836

Prevalence : 0.13342

Detection Rate : 0.05836

Detection Prevalence : 0.08067

Balanced Accuracy : 0.70585

'Positive' Class : Yes

According to the confusion matrix and statistics output from the neural network model developed on the training set, accuracy is 0.8435, no information rate is 0.792, p-value is < 0.05, sensitivity is 0.5352, and specificity is 0.9245.

Test

Confusion Matrix and Statistics

Reference

Prediction No Yes

No 2584 221

Yes 63 186

Accuracy : 0.907

95% CI : (0.8961, 0.9171)

No Information Rate : 0.8667

P-Value [Acc > NIR] : 4.053e-12

Kappa : 0.5183

Mcnemar's Test P-Value : < 2.2e-16

Sensitivity : 0.45700

Specificity : 0.97620

Pos Pred Value : 0.74699

Neg Pred Value : 0.92121

Prevalence : 0.13327

Detection Rate : 0.06090

Detection Prevalence : 0.08153

Balanced Accuracy : 0.71660

'Positive' Class : Yes

The output developed from introducing the model to the test data set is very similar to the training data set. There is no sign of overfitting, but sensitivity and specificity are not very balanced.

Accuracy measures are better for neural network, but there is too much risk with the low sensitivity measure. The logistic regression model involves one more variable, variety, which has many significant predictors, and balances sensitivity and specificity. Because of this, the logistic regression model will act as the champion model for the Chile wine data set.

**Portugal Wine**

**Benchmark Model**

Call:

glm(formula = Exceptional ~ price, family = "binomial", data = Portugaltrain)

Deviance Residuals:

Min 1Q Median 3Q Max

-4.8698 -0.7015 -0.5861 0.7543 2.0253

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -2.470518 0.062708 -39.4 <2e-16 \*\*\*

price 0.079601 0.002535 31.4 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 8249.5 on 6336 degrees of freedom

Residual deviance: 6316.5 on 6335 degrees of freedom

AIC: 6320.5

Number of Fisher Scoring iterations: 6

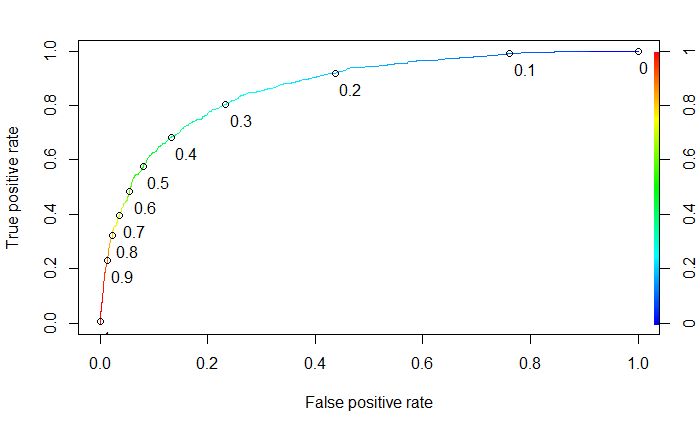
The ordinary logistic regression model yields an AIC of 6320.5. Price is a significant predictor of Exceptional and has a positive coefficient of 0.079601.

**Stepwise Logistic Regression**

Forward and backward stepwise regression yielded identical AIC of 3829.6 and included every variable from the training set. This model defeats the benchmark model and will be used to develop predictions. Price and the province, Beira Atlantico are determined as significant predictors. Price has a positive coefficient of 0.087297 which suggests that higher prices lead to a greater probability of wines being exceptional. Central Spain has a coefficient of 3.214917.

The threshold that balances sensitivity and specificity for this model is 0.3137183. At this threshold, sensitivity is 0.7839397 and specificity is 0.7913299.

ROCR:



The confusion matrix for the training data set is as follows:

FALSE TRUE

No 3239 844

Yes 494 1760

Accuracy is measured at 0.7936785 for the training set. The confusion matrix for the test data set is as follows:

FALSE TRUE

No 1385 364

Yes 187 778

Accuracy is measured at 0.7969786 for the test set. There is no sign of overfitting as the accuracy measures for training and test sets are close.

**Neural Network**

The data summary of the US training set for neural networks is as follows:

Classes ‘tbl\_df’, ‘tbl’ and 'data.frame': 6337 obs. of 3 variables:

$ price : num 10 12 7 20 55 41 40 35 95 18 ...

$ province : Factor w/ 45 levels "Alenquer","Alentejano",..: 13 33 2 13 13 13 2 45 13 2 ...

$ Exceptional: Factor w/ 2 levels "No","Yes": 1 1 1 1 2 2 2 2 2 2 ...

price province Exceptional

Min. : 4.00 Douro :1405 No :4083

1st Qu.: 11.00 Alentejano :1029 Yes:2254

Median : 17.00 Port : 744

Mean : 26.05 Tejo : 641

3rd Qu.: 28.00 Lisboa : 532

Max. :980.00 Vinho Verde: 465

(Other) :1521

The neural network model is created with a hidden layer size of 20 and a decay of 0.1.

Training

Confusion Matrix and Statistics

Reference

Prediction No Yes

No 3593 795

Yes 490 1459

Accuracy : 0.7972

95% CI : (0.7871, 0.8071)

No Information Rate : 0.6443

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.5438

Mcnemar's Test P-Value : < 2.2e-16

Sensitivity : 0.6473

Specificity : 0.8800

Pos Pred Value : 0.7486

Neg Pred Value : 0.8188

Prevalence : 0.3557

Detection Rate : 0.2302

Detection Prevalence : 0.3076

Balanced Accuracy : 0.7636

'Positive' Class : Yes

According to the confusion matrix and statistics output from the neural network model developed on the training set, accuracy is 0.7972, no information rate is 0.6443, p-value is < 0.05, sensitivity is 0.6473, and specificity is 0.88.

Test

Confusion Matrix and Statistics

Reference

Prediction No Yes

No 1547 334

Yes 202 631

Accuracy : 0.8025

95% CI : (0.787, 0.8173)

No Information Rate : 0.6444

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.5554

Mcnemar's Test P-Value : 1.528e-08

Sensitivity : 0.6539

Specificity : 0.8845

Pos Pred Value : 0.7575

Neg Pred Value : 0.8224

Prevalence : 0.3556

Detection Rate : 0.2325

Detection Prevalence : 0.3069

Balanced Accuracy : 0.7692

'Positive' Class : Yes

The output developed from introducing the model to the test data set is very similar to the training data set. There is no sign of overfitting, but sensitivity and specificity are not very balanced.

Accuracy measures for logistic regression and neural network are similar. The logistic regression model involves one more variable, variety, which has many significant predictors, and balances sensitivity and specificity. Because of this, the logistic regression model will act as the champion model for the Portugal wine data set.

**Argentina Wine**

**Benchmark Model**

Call:

glm(formula = Exceptional ~ price, family = "binomial", data = Argentinatrain)

Deviance Residuals:

Min 1Q Median 3Q Max

-5.2935 -0.4473 -0.3650 -0.3226 2.5394

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -3.776549 0.082095 -46.00 <2e-16 \*\*\*

price 0.084700 0.002792 30.33 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 5900.1 on 6540 degrees of freedom

Residual deviance: 4033.8 on 6539 degrees of freedom

AIC: 4037.8

Number of Fisher Scoring iterations: 5

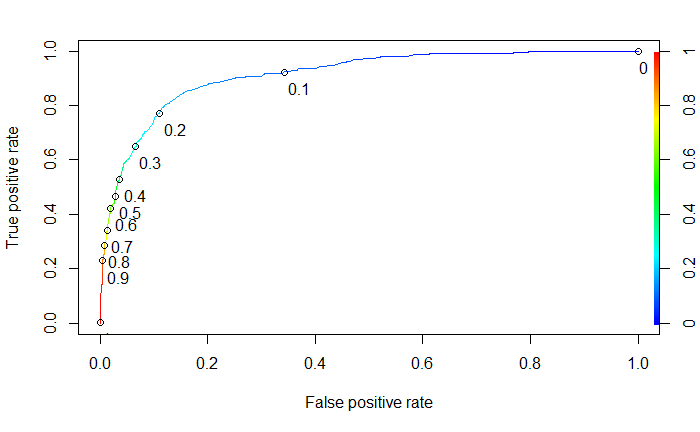
The ordinary logistic regression model yields an AIC of 4037.8. Price is a significant predictor of Exceptional and has a positive coefficient of 0.0847.

**Stepwise Logistic Regression**

Forward and backward stepwise regression yielded identical AIC of 3680.9 and includes 72/74 varieties, 1/2 provinces, and price in the models. These models defeat the benchmark model and will be used to develop predictions. Price and the province, other are determined as significant predictors. Price has a positive coefficient of 0.073525109 which suggests that higher prices lead to a greater probability of wines being exceptional. Other has a coefficient of -1.88624552.

The threshold that balances sensitivity and specificity for this model is 0.1636628. At this threshold, sensitivity is 0.8470696 and specificity is 0.8451092.

ROCR:



The confusion matrix for the training data set is as follows:

FALSE TRUE

No 4606 843

Yes 169 923

Accuracy is measured at 0.8452836 for the training set. The confusion matrix for the test data set is as follows:

FALSE TRUE

No 1976 359

Yes 75 392

Accuracy is measured at 0.8451106 for the test set. There is no sign of overfitting as the accuracy measures for training and test sets are close.

**Neural Network**

The data summary of the US training set for neural networks is as follows:

Classes ‘tbl\_df’, ‘tbl’ and 'data.frame': 6541 obs. of 3 variables:

$ price : num 30 25 26 15 15 23 15 14 25 26 ...

$ province : Factor w/ 2 levels "Mendoza Province",..: 1 1 1 1 1 1 1 1 1 1 ...

$ Exceptional: Factor w/ 2 levels "No","Yes": 2 2 1 1 1 1 1 1 1 1 ...

price province Exceptional

Min. : 4.00 Mendoza Province:5554 No :5449

1st Qu.: 12.00 Other : 987 Yes:1092

Median : 15.00

Mean : 22.33

3rd Qu.: 22.00

Max. :250.00

The neural network model is created with a hidden layer size of 5 and a decay of 0.1.

Training

Confusion Matrix and Statistics

Reference

Prediction No Yes

No 5204 502

Yes 245 590

Accuracy : 0.8858

95% CI : (0.8778, 0.8934)

No Information Rate : 0.8331

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.5468

Mcnemar's Test P-Value : < 2.2e-16

Sensitivity : 0.5403

Specificity : 0.9550

Pos Pred Value : 0.7066

Neg Pred Value : 0.9120

Prevalence : 0.1669

Detection Rate : 0.0902

Detection Prevalence : 0.1277

Balanced Accuracy : 0.7477

'Positive' Class : Yes

According to the confusion matrix and statistics output from the neural network model developed on the training set, accuracy is 0.8858, no information rate is 0.8331, p-value is < 0.05, sensitivity is 0.5403, and specificity is 0.955.

Test

Confusion Matrix and Statistics

Reference

Prediction No Yes

No 2223 214

Yes 112 253

Accuracy : 0.8837

95% CI : (0.8712, 0.8953)

No Information Rate : 0.8333

P-Value [Acc > NIR] : 4.697e-14

Kappa : 0.5411

Mcnemar's Test P-Value : 2.221e-08

Sensitivity : 0.54176

Specificity : 0.95203

Pos Pred Value : 0.69315

Neg Pred Value : 0.91219

Prevalence : 0.16667

Detection Rate : 0.09029

Detection Prevalence : 0.13026

Balanced Accuracy : 0.74690

'Positive' Class : Yes

The output developed from introducing the model to the test data set is very similar to the training data set. There is no sign of overfitting, but sensitivity and specificity are not very balanced.

Accuracy measures are better for the neural network model than for the logistic regression model. Because of the better accuracy and the lack of significant predictors with the logistic regression model, the neural network model will act as the champion model for the Argentina wine data set.

**Australia Wine**

**Benchmark Model**

Call:

glm(formula = Exceptional ~ price, family = "binomial", data = Australiatrain)

Deviance Residuals:

Min 1Q Median 3Q Max

-4.2077 -0.7045 -0.5823 0.6511 2.0581

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -2.47167 0.06953 -35.55 <2e-16 \*\*\*

price 0.06023 0.00218 27.63 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 6441.7 on 5032 degrees of freedom

Residual deviance: 5004.0 on 5031 degrees of freedom

AIC: 5008

Number of Fisher Scoring iterations: 6

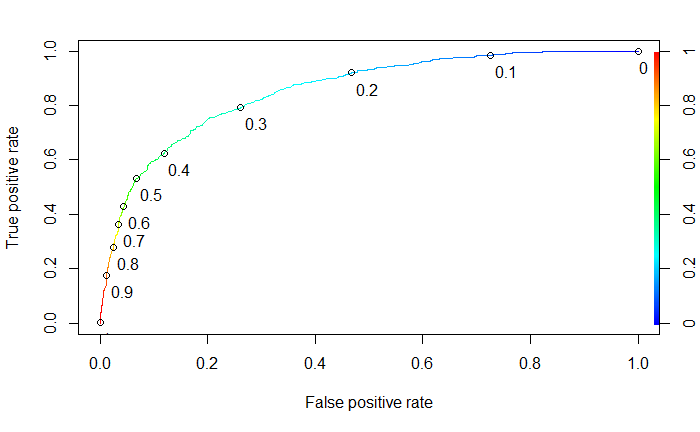
The ordinary logistic regression model yields an AIC of 5008. Price is a significant predictor of Exceptional and has a positive coefficient of 0.06023.

**Stepwise Logistic Regression**

Forward and backward stepwise regression yielded identical AIC of 4761 and includes all variables. These models defeat the benchmark model and will be used to develop predictions. Price, several provinces, and several varieties are determined as significant predictors. Significant Provinces include South Wales, Queensland, South Australia, Tasmania, Victoria, and Western Australia. Significant varieties include Cabernet Sauvignon, G-S-M, Grenache, Muscat, Rhane-style Red Blend, Riesling, Sacmillon, Shiraz, Grenache, and Shiraz-Viognier. Price has a positive coefficient of 0.054992 which suggests that higher prices lead to a greater probability of wines being exceptional. All province and variety coefficients are positive.

The threshold that balances sensitivity and specificity for this model is 0.3302429. At this threshold, sensitivity is 0.7539636 and specificity is 0.7981982.

ROCR:



The confusion matrix for the training data set is as follows:

FALSE TRUE

No 2658 672

Yes 419 1284

Accuracy is measured at 0.7832307 for the training set. The confusion matrix for the test data set is as follows:

FALSE TRUE

No 1132 294

Yes 170 559

Accuracy is measured at 0.7846868 for the test set. There is no sign of overfitting as the accuracy measures for training and test sets are close.

**Neural Network**

The data summary of the US training set for neural networks is as follows:

Classes ‘tbl\_df’, ‘tbl’ and 'data.frame': 5033 obs. of 3 variables:

$ price : num 20 65 65 25 25 33 125 17 34 150 ...

$ province : Factor w/ 7 levels "Australia Other",..: 4 4 5 4 4 6 4 4 6 4 ...

$ Exceptional: Factor w/ 2 levels "No","Yes": 2 2 2 1 2 2 2 2 1 2 ...

price province Exceptional

Min. : 5.0 Australia Other : 551 No :3330

1st Qu.: 15.0 New South Wales : 230 Yes:1703

Median : 20.0 Queensland : 1

Mean : 32.1 South Australia :2993

3rd Qu.: 35.0 Tasmania : 58

Max. :850.0 Victoria : 682

Western Australia: 518

The neural network model is created with a hidden layer size of 12 and a decay of 0.1.

Training

Confusion Matrix and Statistics

Reference

Prediction No Yes

No 3021 762

Yes 309 941

Accuracy : 0.7872

95% CI : (0.7756, 0.7984)

No Information Rate : 0.6616

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.4917

Mcnemar's Test P-Value : < 2.2e-16

Sensitivity : 0.5526

Specificity : 0.9072

Pos Pred Value : 0.7528

Neg Pred Value : 0.7986

Prevalence : 0.3384

Detection Rate : 0.1870

Detection Prevalence : 0.2484

Balanced Accuracy : 0.7299

'Positive' Class : Yes

According to the confusion matrix and statistics output from the neural network model developed on the training set, accuracy is 0.7872, no information rate is 0.6616, p-value is < 0.05, sensitivity is 0.5526, and specificity is 0.9072.

Test

Confusion Matrix and Statistics

Reference

Prediction No Yes

No 1276 313

Yes 150 416

Accuracy : 0.7852

95% CI : (0.7672, 0.8023)

No Information Rate : 0.6617

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.4924

Mcnemar's Test P-Value : 5.122e-14

Sensitivity : 0.5706

Specificity : 0.8948

Pos Pred Value : 0.7350

Neg Pred Value : 0.8030

Prevalence : 0.3383

Detection Rate : 0.1930

Detection Prevalence : 0.2626

Balanced Accuracy : 0.7327

'Positive' Class : Yes

The output developed from introducing the model to the test data set is very similar to the training data set. There is no sign of overfitting, but sensitivity and specificity are not very balanced.

Accuracy measures for logistic regression and neural network are similar. The logistic regression model involves one more variable, variety, which has many significant predictors, and balances sensitivity and specificity. Because of this, the logistic regression model will act as the champion model for the Australia wine data set.

**Austria Wine**

**Benchmark Model**

Call:

glm(formula = Exceptional ~ price, family = "binomial", data = Austriatrain)

Deviance Residuals:

Min 1Q Median 3Q Max

-3.4254 -0.9772 0.2961 1.0622 1.6745

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -1.817821 0.090718 -20.04 <2e-16 \*\*\*

price 0.069834 0.003189 21.90 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 5103.1 on 3697 degrees of freedom

Residual deviance: 4345.3 on 3696 degrees of freedom

AIC: 4349.3

Number of Fisher Scoring iterations: 6

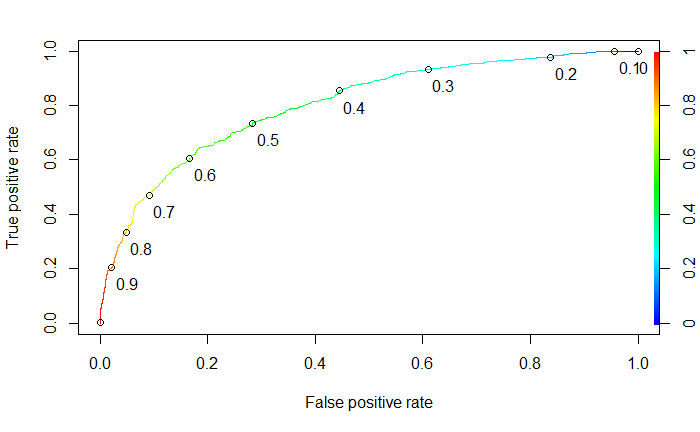
The ordinary logistic regression model yields an AIC of 4349.3. Price is a significant predictor of Exceptional and has a positive coefficient of 0.069834.

**Stepwise Logistic Regression**

Forward and backward stepwise regression yielded identical AIC of 4761 and includes all variables. These models defeat the benchmark model and will be used to develop predictions. Price and several varieties are determined as significant predictors. Significant varieties include Cabernet Sauvignon, Gelber Muskateller, Merlot, Pinot Noir, Red Blend, Sauvignon Blanc, St. Laurent, Traminer, Zweigelt. Price has a positive coefficient of 0.069984 which suggests that higher prices lead to a greater probability of wines being exceptional. All variety coefficients are negative.

The threshold that balances sensitivity and specificity for this model is 0.4954528. At this threshold, sensitivity is 0.7369739 and specificity is 0.7162162.

ROCR:



The confusion matrix for the training data set is as follows:

FALSE TRUE

No 2658 672

Yes 419 1284

Accuracy is measured at 0.7268794 for the training set. The confusion matrix for the test data set is as follows:

FALSE TRUE

No 536 193

Yes 207 648

Accuracy is measured at 0.6912879 for the test set. There is a small sign of overfitting as the accuracy measure for training is a little bit higher than the test set accuracy.

**Neural Network**

The data summary of the US training set for neural networks is as follows:

Classes ‘tbl\_df’, ‘tbl’ and 'data.frame': 3698 obs. of 3 variables:

$ price : num 19 27 19 17 17 18 22 24 17 22 ...

$ province : Factor w/ 30 levels "Ã–sterreichischer Perlwein",..: 4 15 22 26 26 8 15 7 8 4 ...

$ Exceptional: Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...

price province Exceptional

Min. : 7.00 Burgenland :840 No :1702

1st Qu.: 18.00 NiederÃ¶sterreich:615 Yes:1996

Median : 25.00 Wachau :449

Mean : 31.31 Kamptal :409

3rd Qu.: 38.00 Kremstal :373

Max. :1100.00 Thermenregion :153

(Other) :859

The neural network model is created with a hidden layer size of 31 and a decay of 0.1.

Training

Confusion Matrix and Statistics

Reference

Prediction No Yes

No 1166 468

Yes 536 1528

Accuracy : 0.7285

95% CI : (0.7139, 0.7428)

No Information Rate : 0.5398

P-Value [Acc > NIR] : < 2e-16

Kappa : 0.4519

Mcnemar's Test P-Value : 0.03447

Sensitivity : 0.7655

Specificity : 0.6851

Pos Pred Value : 0.7403

Neg Pred Value : 0.7136

Prevalence : 0.5398

Detection Rate : 0.4132

Detection Prevalence : 0.5581

Balanced Accuracy : 0.7253

'Positive' Class : Yes

According to the confusion matrix and statistics output from the neural network model developed on the training set, accuracy is 0.7285, no information rate is 0.5398, p-value is < 0.05, sensitivity is 0.7655, and specificity is 0.6851.

Test

Confusion Matrix and Statistics

Reference

Prediction No Yes

No 517 211

Yes 212 644

Accuracy : 0.733

95% CI : (0.7104, 0.7546)

No Information Rate : 0.5398

P-Value [Acc > NIR] : <2e-16

Kappa : 0.4625

Mcnemar's Test P-Value : 1

Sensitivity : 0.7532

Specificity : 0.7092

Pos Pred Value : 0.7523

Neg Pred Value : 0.7102

Prevalence : 0.5398

Detection Rate : 0.4066

Detection Prevalence : 0.5404

Balanced Accuracy : 0.7312

'Positive' Class : Yes

The output developed from introducing the model to the test data set is very similar to the training data set. There is no sign of overfitting, but sensitivity and specificity are not very balanced.

Accuracy measures for logistic regression and neural network are similar. The logistic regression model involves one more variable, variety, which has many significant predictors, and balances sensitivity and specificity. Because of this, the logistic regression model will act as the champion model for the Austria wine data set.

**New Zealand Wine**

**Benchmark Model**

Call:

glm(formula = Exceptional ~ price, family = "binomial", data = NZtrain)

Deviance Residuals:

Min 1Q Median 3Q Max

-3.3322 -0.6763 -0.5773 0.5764 2.0568

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -2.823827 0.106159 -26.60 <2e-16 \*\*\*

price 0.069764 0.003706 18.83 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 3561.8 on 3060 degrees of freedom

Residual deviance: 3050.0 on 3059 degrees of freedom

AIC: 3054

Number of Fisher Scoring iterations: 4

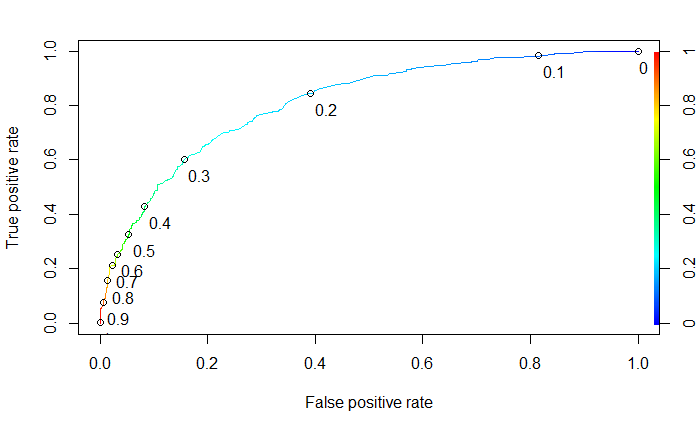
The ordinary logistic regression model yields an AIC of 3054. Price is a significant predictor of Exceptional and has a positive coefficient of 0.069764.

**Stepwise Logistic Regression**

Forward and backward stepwise regression yielded identical AIC of 2972 and includes all variables. These models defeat the benchmark model and will be used to develop predictions. Price, several provinces, and several varieties are determined as significant predictors. Significant provinces include Canterbury, Central Otago, East Coast, Hawke’s Bay, Marlborough, Martinborough, Nelson, New Zealand, Waiheke Island, and Wairapara. Significant varieties include Gewarztraminer, Merlot, and White Blend. Price has a positive coefficient of 0.08718 which suggests that higher prices lead to a greater probability of wines being exceptional. All province and variety coefficients are negative, except for White Blend.

The threshold that balances sensitivity and specificity for this model is 0.2194863. At this threshold, sensitivity is 0.7627737 and specificity is 0.7101385.

ROCR:



The confusion matrix for the training data set is as follows:

FALSE TRUE

No 1590 649

Yes 195 627

Accuracy is measured at 0.7242731 for the training set. The confusion matrix for the test data set is as follows:

FALSE TRUE

No 722 237

Yes 76 276

Accuracy is measured at 0.761251 for the test set. There is a small sign of overfitting as the accuracy measure for training is a little bit higher than the test set accuracy.

**Neural Network**

The data summary of the US training set for neural networks is as follows:

Classes ‘tbl\_df’, ‘tbl’ and 'data.frame': 3061 obs. of 3 variables:

$ price : num 57 33 40 25 34 65 61 29 28 35 ...

$ province : Factor w/ 28 levels "Awatere Valley",..: 10 11 3 16 10 3 13 11 13 11 ...

$ Exceptional: Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...

price province Exceptional

Min. : 7.00 Marlborough :1649 No :2239

1st Qu.: 16.00 Central Otago: 334 Yes: 822

Median : 20.00 Hawke's Bay : 320

Mean : 24.92 Martinborough: 226

3rd Qu.: 28.00 Nelson : 107

Max. :125.00 Waipara : 86

(Other) : 339

The neural network model is created with a hidden layer size of 32 and a decay of 0.1.

Training

Confusion Matrix and Statistics

Reference

Prediction No Yes

No 2081 523

Yes 158 299

Accuracy : 0.7775

95% CI : (0.7624, 0.7922)

No Information Rate : 0.7315

P-Value [Acc > NIR] : 2.602e-09

Kappa : 0.3411

Mcnemar's Test P-Value : < 2.2e-16

Sensitivity : 0.36375

Specificity : 0.92943

Pos Pred Value : 0.65427

Neg Pred Value : 0.79916

Prevalence : 0.26854

Detection Rate : 0.09768

Detection Prevalence : 0.14930

Balanced Accuracy : 0.64659

'Positive' Class : Yes

According to the confusion matrix and statistics output from the neural network model developed on the training set, accuracy is 0.7775, no information rate is 0.7315, p-value is < 0.05, sensitivity is 0.36375, and specificity is 0.92943.

Test

Confusion Matrix and Statistics

Reference

Prediction No Yes

No 895 202

Yes 64 150

Accuracy : 0.7971

95% CI : (0.7743, 0.8186)

No Information Rate : 0.7315

P-Value [Acc > NIR] : 2.12e-08

Kappa : 0.4103

Mcnemar's Test P-Value : < 2.2e-16

Sensitivity : 0.4261

Specificity : 0.9333

Pos Pred Value : 0.7009

Neg Pred Value : 0.8159

Prevalence : 0.2685

Detection Rate : 0.1144

Detection Prevalence : 0.1632

Balanced Accuracy : 0.6797

'Positive' Class : Yes

The output developed from introducing the model to the test data set is very similar to the training data set. There is no sign of overfitting, but sensitivity and specificity are not very balanced.

Accuracy measures are better for neural network, but there is too much risk with the low sensitivity measure. The logistic regression model involves one more variable, variety, which has many significant predictors, and balances sensitivity and specificity. Because of this, the logistic regression model will act as the champion model for the New Zealand wine data set.