Homework #3: Crime Prediction

Data 621 Business Analytics and Data Mining

Aadi Kalloo, Nathan Lim, Asher Meyers, Daniel Smilowitz, Logan Thomson

Due July 3, 2016

Contents

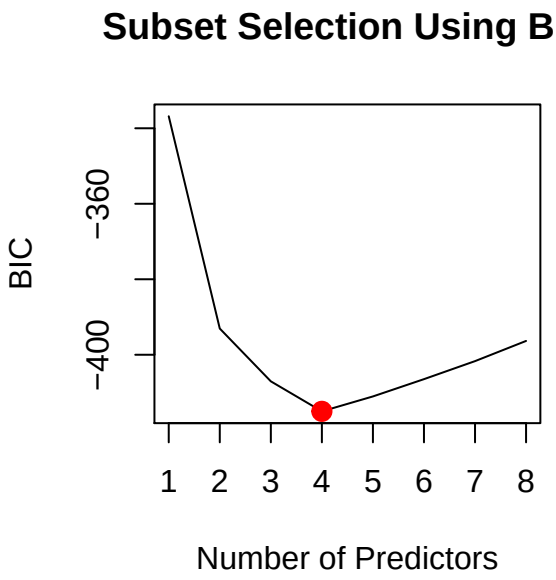
Data Exploration

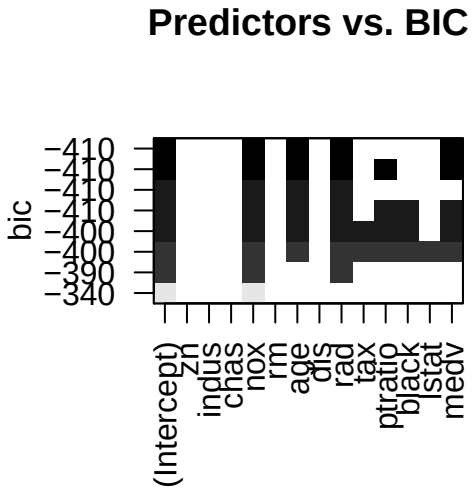
Data Preparation

Model Creation

Model 1: Bayesian Information Criterion

The first model created utilizes the Bayesian Information Criterion (BIC) to determine the number of predictors to use and which predictors should be used.





The left plot above shows that the BIC is minimized using 4 predictors. The plot on the right shows that the 4 predictors with the lowest BIC are `nox`, `age`, `rad`, and `medv`. As such, a model is created using these predictors; this model is presented below.

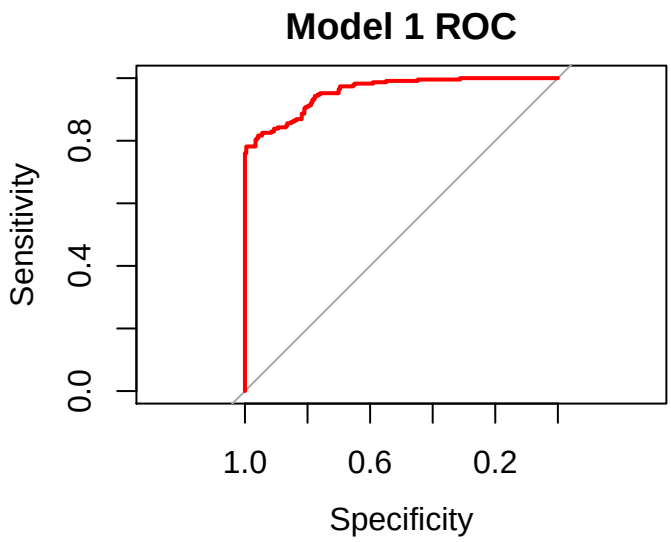
	Estimate	Std. Error	z value	Pr(>|z|)
nox	23.62	3.936	6.003	1.942e-09
age	0.01824	0.009172	1.989	0.04673
rad	0.4528	0.1093	4.144	3.413e-05
medv	0.04481	0.02319	1.932	0.05338
(Intercept)	-17.63	2.168	-8.131	4.246e-16

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	645.9 on 465 degrees of freedom
Residual deviance:	232.8 on 461 degrees of freedom

The coefficients in this model indicate that nitrogen oxides concentration has the strongest, as well as the most statistically significant, effect on on the target variable. Age, highway access, and home value all have far weaker effects on the target. All of the estimated coefficients are statistically significant at the $\alpha = 0.5$ level except `medv` – the p-value for this coefficient is 0.0534.

The receiver operating characteristic (ROC) curve and confusion matrix returned by this model are shown below:

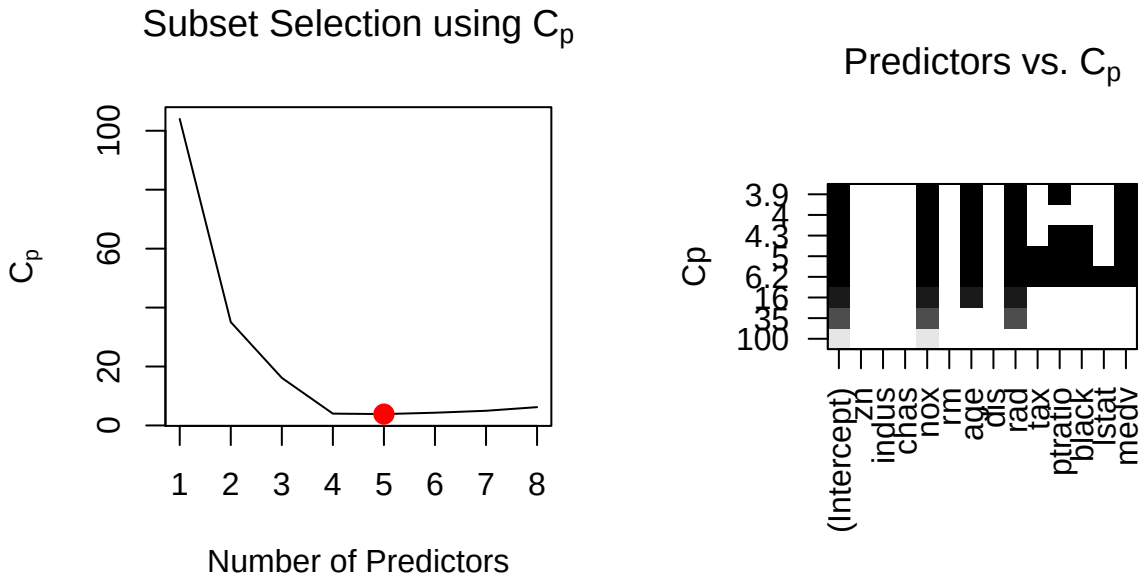


	0	1
0	214	23
1	37	192

This model provides an area under the curve of 0.9570 and an accuracy of 0.8691.

Model 2: Mallow's C_p

The second model created utilizes Mallow's C_p to determine the number of predictors to use and which predictors should be used.



The left plot above shows that C_p is minimized using 5 predictors. The plot on the right shows that the 5 predictors with the lowest C_p are `nox`, `age`, `rad`, `ptratio`, and `medv`. As such, a model is created using these predictors; this model is presented below.

	Estimate	Std. Error	z value	Pr(>|z|)
nox	25.33	4.084	6.203	5.53e-10
age	0.0194	0.009308	2.085	0.03711
rad	0.5126	0.1148	4.464	8.027e-06
ptratio	0.2742	0.09874	2.777	0.005486
medv	0.08544	0.02798	3.054	0.002259
(Intercept)	-24.94	3.683	-6.77	1.289e-11

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	645.9 on 465 degrees of freedom
Residual deviance:	224.7 on 460 degrees of freedom

As in model 1, the coefficients in this model indicate that nitrogen oxides concentration has the strongest, as well as the most statistically significant, effect on on the target variable. This may be due to the fact that a single part per 10 million in concentration merits a unit increase in this variable. Each of the estimated coefficients are statistically signficant.

The ROC curve and confusion matrix returned by this model are shown below:

This model provides an area under the curve of 0.9605 and an accuracy of 0.8691.

Model 3: Transformed BIC

Model 3 is created using the same BIC selection from Model 1, with modification from transformation. Based on the distributions of the `nox` and `rad` variables, log transformations of these variables are used. The model using these transformed variables, the model is presented below.

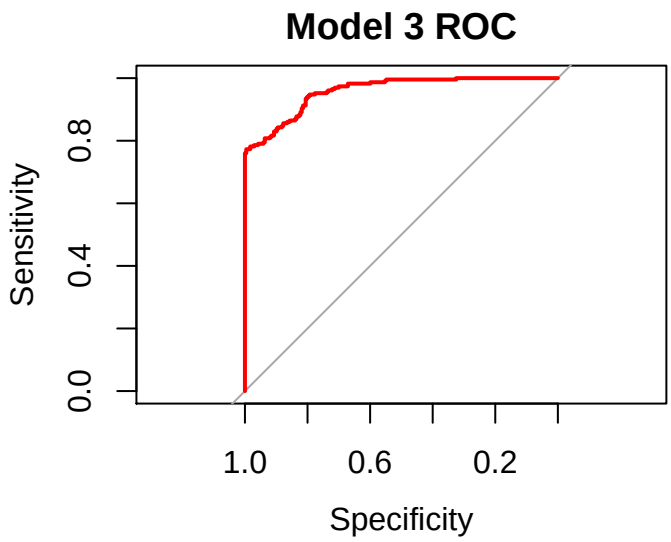
	Estimate	Std. Error	z value	Pr(>|z|)
log(nox)	13	2.097	6.199	5.681e-10
age	0.01696	0.009147	1.855	0.06366
log(rad)	2.269	0.4494	5.05	4.428e-07
medv	0.0489	0.02354	2.078	0.03774
(Intercept)	1.96	1.925	1.018	0.3086

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	645.9 on 465 degrees of freedom
Residual deviance:	231.7 on 461 degrees of freedom

In model 3, the coefficient associated with nitrogen oxide concentration is once again the largest in magnitude, although it has decreased in magnitude. The estimated coefficient for home value is now statistically signficant; however, the coefficient for age no longer is.

The ROC curve and confusion matrix returned by this model are shown below:



	0	1
0	213	24
1	37	192

This model has an area under the curve of 0.9584 and an accuracy of 0.8691.

Model 4: Significant BIC

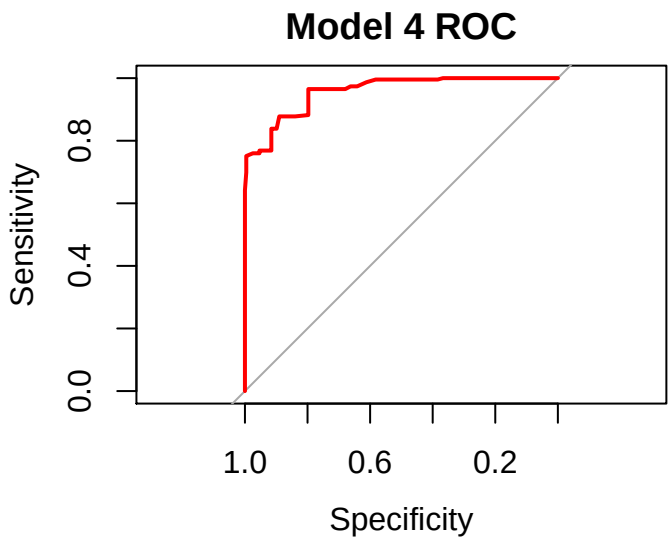
Like Model 1, this model is formed using best subsets regression with BIC as a criterion for choosing the number of predictors; however, with this model, predictors are removed sequentially until all predictors are statistically significant ($p < 0.05$). This leads to the removal of age and median home values as predictors.

	Estimate	Std. Error	z value	Pr(>|z|)
nox	27.2	3.232	8.415	3.915e-17
rad	0.5139	0.1082	4.75	2.036e-06
(Intercept)	-17.45	1.949	-8.956	3.376e-19

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	645.9 on 465 degrees of freedom
Residual deviance:	239.5 on 463 degrees of freedom

The coefficients for this model are similar to the coefficients associated with the two variables in the larger BIC model (Model 1), but show far greater statistical significance with the additional variables removed.



	0	1
0	213	24
1	37	192

This model has an area under the curve of 0.9575 and an accuracy of 0.8691.

Model 5: Best GLM

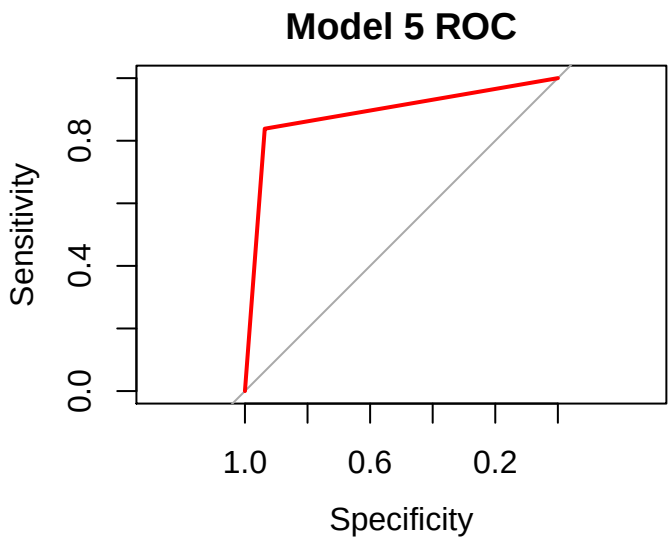
The `bestglm` package is implemented to "[select] the best subset of inputs for the GLM family." The BIC is still used as the information criteria. The model generated by the package is presented below.

	Estimate	Std. Error	z value	Pr(>|z|)
nox	35.63	4.524	7.877	3.35e-15
rad	0.6376	0.1194	5.338	9.377e-08
tax	-0.008146	0.002332	-3.493	0.0004776
(Intercept)	-19.87	2.368	-8.389	4.911e-17

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	645.9 on 465 degrees of freedom
Residual deviance:	224.5 on 462 degrees of freedom

The coefficients for this model again indicate the strong influence of a unit increase in nitrogen oxide concentration. The inclusion of property tax rate in the model lead to the first negative coefficient seen. Each of the coefficient estimates is of very high statistical significance.



Quitting from lines 201-207 (HW3_Master.Rmd) Error in confusionMatrix.default(train_df$target, train_df$target_model5, : the data cannot have more levels than the reference Calls: ... pander -> confusionMatrix -> confusionMatrix.default In addition: Warning message: In get_engine(options$engine) : Unknown language engine 'r model' (must be registered via knit_engines$set()).

This model has an area under the curve of 0.8876 and an accuracy of 0.8691.

Model 6: Transformed Best GLM

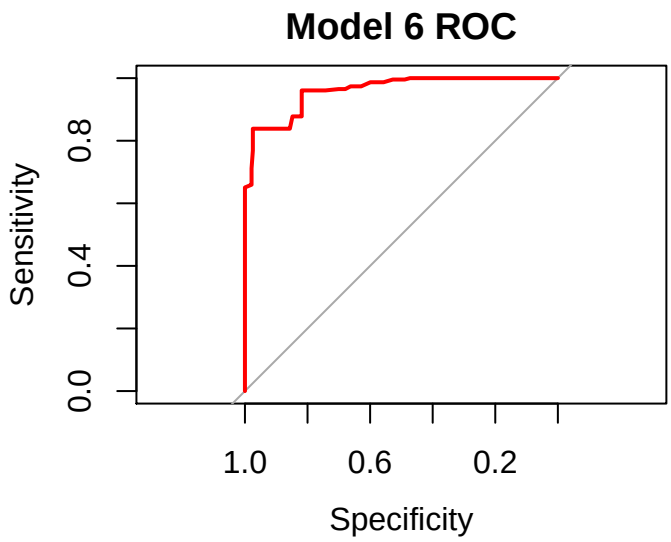
Using the three variables identified by the `bestglm` function for Model 5, the `nox` and `rad` predictors are transformed using logarithms, as in Model 3. The model using these transformations is presented below.

	Estimate	Std. Error	z value	Pr(>|z|)
log(nox)	19.35	2.519	7.682	1.564e-14
log(rad)	3.356	0.5447	6.162	7.196e-10
tax	-0.008214	0.002335	-3.518	0.0004345
(Intercept)	9.385	1.829	5.132	2.866e-07

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	645.9 on 465 degrees of freedom
Residual deviance:	223.5 on 462 degrees of freedom

The coefficient for nitrogen oxide concentration decreased following the transformation, as did the standard error of the estimate of the coefficient. While the p-value related to the coefficients changed following the tranformation, all estimates remain statistically significant.



	0	1
0	222	15
1	37	192

This model has an area under the curve of 0.9594 and an accuracy of 0.8884.

Model 7: Full Model

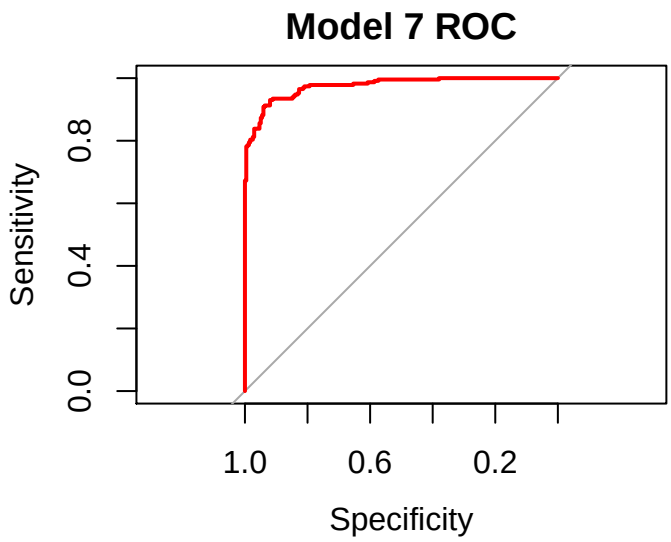
For completeness, a full generalized linear model is created using all explanatory variables.

	Estimate	Std. Error	z value	Pr(>|z|)
zn	-0.06172	0.03441	-1.794	0.07287
indus	-0.07258	0.04855	-1.495	0.1349
chas	1.032	0.7596	1.359	0.1741
nox	50.16	8.05	6.231	4.623e-10
rm	-0.6921	0.7414	-0.9335	0.3505
age	0.03452	0.01388	2.487	0.01289
dis	0.7658	0.2344	3.267	0.001087
rad	0.663	0.1651	4.015	5.945e-05
tax	-0.006593	0.003064	-2.152	0.03142
ptratio	0.4422	0.1322	3.344	0.0008252
black	-0.01309	0.00668	-1.96	0.04997
lstat	0.04757	0.05451	0.8727	0.3828
medv	0.1997	0.07102	2.812	0.004919
(Intercept)	-36.84	7.029	-5.241	1.595e-07

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	645.9 on 465 degrees of freedom
Residual deviance:	186.1 on 452 degrees of freedom

The 13 coefficients in the full model vary very widely in magnitude, and five of the coefficients are negative. There is also a wide range of signifcances – 5 of the coefficients are not significant under any reasonable α, and one is very nearly exactly 0.05.



	0	1
0	222	15
1	20	209

This model has an area under the curve of 0.9753 and an accuracy of 0.9249.

Model Comparison

The characteristics and performance of the six models are compared below:

Model #	# of Predictors	AUC	Accuracy
1	4	0.9570	0.8712
2	5	0.9605	0.8691
3	4	0.9584	0.8691
4	2	0.9575	0.8691
5	3	0.8876	0.8691
6	3	0.9594	0.8884
7	13	0.9753	0.9249

Model Selection and Prediction

The model selected used only the significant predictors (Model 3) was selected as the best model for prediction of `TARGET` in the crime data set. While the AUC value of this model the second-highest of the four models tested, its mean cross-validation error indicates that it is has the best predictive value for unseen data. Additionally, it is a parsimonious model, and the simplicity lends itself to easier understanding of the model by other users.

10-fold Cross Validation

Mean CV Error

Model1	36.6
Model2	46.69
Model3	15.48
Model4	22.14

The linear model is applied to an evaluation dataset containing response variables for 259 cases. A table of the predicted team wins is presented below.

0	1
250	216

0	1
237	229

Similar to the training dataset, the predictions for the test data set predictions are weighted more toward crime being below the median

A comparison of the full sets of predictions for the evaluation dataset is available in Appendix B.

Appendix A

Appendix B – Index-wise Results from Predictive Model

Appendix C – R Code