Tyler Bradley
Methodology and Results of Exercise 01 - RTI
02/10/2015

The data was read in, collapsed into one table, and saved within R. Then the variables were investigated with minor cleaning ("Never-worked" for class and "Holand-Netherlands" for country were removed because of complete separation). After this, the data was partitioned into training, validation, and testing sets for modeling. A basic logistic model with all variables was fit first. Then automatic selection techniques and LASSO with standardized continuous inputs were used to reduce the number of terms. A random forest was fit in case the relationships in the data followed this model better. ROC curves on the validation set were used to compare the different models. A logistic model with standardized inputs and variables selected by LASSO was found to be the best by comparing model fit and complexity. The optimal cutoff for deciding between 0 and 1 was found using misclassification error on the validation set. The final model's misclassification error on the testing set is 14.86%. The parameters for the final model built on all the data are in the following table.

Estimates for predictor variables (continuous are standardize)									Intercept
									-3.19
Marital Status		Occupation							Sex
Married-Civ-Spouse	Never Married	Exec-managerial	Farming-fishing	Handlers-cleaners	Other-service	Prof-specialty	Sales	Tech-support	Male
1.79	-0.44	0.84	-0.94	-0.63	-0.86	0.61	0.35	0.62	0.80
Relationship		Class			Age	Education	Capital		Hours
Own-child	Wife	Federal-gov	Self-emp-inc	Self-emp-not-inc	Years	Number	Loss	Gain	Per week
-0.99	1.17	0.57	0.20	-0.46	0.32	0.72	0.26	2.35	0.38

Chart showing an important relationship: