Thom Miano - RTI CDS Analytics Exercise 01 - Methodology

Business Understanding & Data Understanding

- Considered implications of predicting the target (i.e., people who earn >$50k/yr).
- Became familiar with features in the data set.

Data Preparation

- Retrieved the data and wrote R code that connected to the SQLite db and flattened the data into a CSV file.
- Ran simple exploratory data analysis, plotting histograms and bar plots.
- Removed all rows that contained missing values (i.e., "?").

Modeling & Evaluation

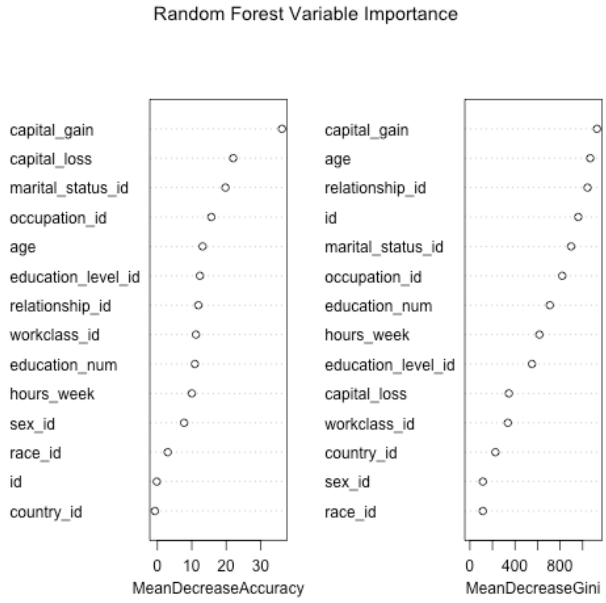
- Due to prevalence of categorical variables, used Random Forests classification.
- Randomly sampled 20% of data to create testing set, and randomly sampled 20% of the remaining data to create the validation set. Remaining data used for training.
- Ran four validation models to find a good value for number of trees parameter:

	model	sensitivity	specificity	accuracy	ntree	model.time	seed	train.sample	validate.sample
1	random forest	0.2676056	0.9968870	0.8179934	100	0.210	756	28941	7236
2	random forest	0.2687324	0.9974364	0.8186844	200	0.393	757	28941	7236
3	random forest	0.2687324	0.9968870	0.8182698	300	0.592	758	28941	7236
4	random forest	0.2681690	0.9968870	0.8181316	400	0.767	759	28941	7236

- Ran the model on the test data four times:

	model	sensitivity	specificity	accuracy	ntree	model.time	seed	train.sample	test.sample
1	random forest	0.2759670	0.9967378	0.8133776	100	0.379	600	28941	9045
2	random forest	0.2803129	0.9967378	0.8144831	100	0.231	601	28941	9045
3	random forest	0.2768362	0.9965896	0.8134881	100	0.297	602	28941	9045
4	random forest	0.2768362	0.9976275	0.8142620	100	0.245	603	28941	9045

- Used sensitivity, specificity, and accuracy as evaluation metrics.



Final Thoughts:

- Given more time I would reevaluate my method for calculating sensitivity, specificity, and accuracy. I would use the randomForest() confusion matrix.
- I would also use 10-fold cross-validation, and compare the random forest performance to other models.