Course Project: Part 2 –Model Development

Attn: Dr. Stephen Hill

BAN 502: Introduction to Predictive Analytics

Stephen AJ Goode

Before creating our models, we want to start by splitting the data into training and testing sets using a seed of 1234 and a 70/30 split. Now we can begin creating our models starting with “model1”. Let’s start with a Logistic Regression. I included only the most visually appealing variables from the bar graphs and boxplots in Part 1 such as Cloud9am, Cloud 3pm, Humidity9am, Humidity3pm, Pressure9am, Pressure3pm and Temp3pm. The model yielded an AIC of 14839. Let’s see if we can do better.

To create “model2”, I removed the insignificant variables (Temp3pm & Humidity9am) and added Temp9am, WindGustSpeed, Month and WindGustDir to optimize our Logistic Regression model and improve the AIC from to 14175. There’s a lot of variables in our dataset so let’s see if we can speed things up and create a better model without the manual plug & play.

Using stepwise regression, I was able to achieve a lower AIC of 13701 in both my forward and backward models. Both models were identical with the following significant variables: Humidity9am, Humidity3pm, WindGustSpeed, Cloud9am, Cloud3pm, Pressure9am, Pressure3pm, Temp9am, RainTodayYes, Rainfall, WindDir9am (SSE, SE, S, NNE), WindDir3pm (E, ESE, ENE, NE, SSE, N, NNE), WindSpeed9am, WindSpeed3pm, and Month (March – November). That’s the majority of our variables. The following variables were excluded: Temp3pm, Year, WindGustDir, Day, MinTemp and MaxTemp.

Now that we have a good model, we can develop an ROC curve and determine the probability threshold that best balances the sensitivity & specificity of our training set. Based on our code, the probability threshold that best balances sensitivity & specificity is 0.2255691. We can now use this number to create a balanced Confusion Matrix:

* Accuracy = 79.65%
* Sensitivity = 79.63%
* Specificity = 79.66%

Next, I created a classification tree to predict RainTomorrow in the training set and plotted the tree. I used the printcp & plotcp functions to evaluate tree performance as a function of the complexity parameter (CP). The optimal CP value is 0.208721 which results in a “root”, or a tree with no splits. Now we can prune our tree to the optimal CP value and use a Confusion Matrix to evaluate predictions using both the pruned and unpruned tree.

Starting with our unpruned tree, “tree1”, we can develop predictions from our training set and use caret’s confusionMatrix function to evaluate the accuracy. Our model has an Accuracy of 83%, Sensitivity of 34% and Specificity of 97.19%. The Accuracy is better than our naïve and the P-value is significant which are both good indicators.

Running our test data through our unpruned classification tree, it looks like our model yielded similar results. The Accuracy is 82.71% which is slightly below the Accuracy of our training set. The Sensitivity is 34.29% and the Specificity is 96.75%. The results are very close to our training set.

Predictions were also developed using the pruned “tree2” model on the training and test datasets. The results were very similar and there’s nothing noteworthy to add.

I’d also like to create a Random Forest model, however, there’s way too many variables in our data. I’m going to create two datasets – MORNING & AFTERNOON – and isolate our AM/PM variables against our response variable. I wasn’t sure which variables to exclude but I knew where to start – with our visually obvious & significant variables such as Humidity, Cloud Coverage & Atmospheric Pressure. From there I can plug & play using VarImp.

I removed Temp9am because it was the least important variable with an overall score of 0. I replaced Temp9am with WindSpeed9am but there was no change – WindSpeed9am was 0 as well so I replaced it with WindGustSpeed. So our MORNING model has Humidity9am, Cloud9am, Pressure9am WindGustSpeed and our response variable. Humidity is the most important variable followed by WindGustSpeed, Cloud9am and Pressure9am.

Naturally, I wanted to test these variables using their afternoon counterparts. I created the “Afternoon” dataset that includes Cloud3pm, Pressure3pm, Humidity3pm, WindGustSpeed and of course our response variable. I expected to see the same pecking order and ranking; it was close but that was not the case. Humidity was the strongest variable again followed by Cloud3pm, WindGustSpeed and Pressure3pm. In conclusion, the afternoon and morning variables do not have the exact same levels of importance.

I also tried to make a Confusion Matrix based on predictions I created via a Neural Network, however, I could not get the code to run past the Predictions on the training set.

I wanted to create a cluster analysis as well. To do this, I selected a handful numeric variables in our dataset and named it rain\_num. But first I went back and omitted the missing data from our original dataset “rain”, to cut down on processing time. Instead of 28,000 observations, we now have just under 14,000 which is a lot better. After scaling our data, we can now run an analysis and visualize our clusters. The also calculated the optimal number of clusters – three (3) – which is the amount I started out with.