CASE STUDY 018
[Machine Learning: R]
Election results

Difficulty Level: 3 of 3

Disclaimer: SuperDataScience has no affiliation with data sources. The scenario is made up for educational purposes.

You are a Data Analyst working for the Democratic Party who would like to understand the demographics behind the primary results

1. Merge the two datasets using an inner join
2. Count the number of counties won by each candidate
3. Split the data into training and testing data sets (80% in training set)
4. What are the levels of the factor winner?
5. Build a logistic regression model using all of the available variables
6. Rerun the model dropping the least significant variable
7. Predict the probability for the training data set
8. Try different thresholds for the probability at which we predict 'Hilary Clinton' (try, 0.4, 0.5, 0.6, 0.7)
9. Calculate the confusion matrix using the best threshold from above
10. Predict the results for the testing data set using the same threshold
11. Calculate the confusion matrix for the testing set

Good luck!

Difficulty note: this is a difficult assignment. Do not be surprised that there will be lots of nuances we have not covered off in the courses. But just like in the Real Life – there will be things training has not prepared you for and you will need to do research to find how to solve the problems at hand. If you get stuck, check the clues file. Data source: https://www.kaggle.com/benhamner/2016-us-election