ECON 386

Dr. Levkoff

Prediction Project

The purpose of this project is to demonstrate your ability to preprocess data for application in a predictive analytics environment. You will be assigned to a group of 4 other data scientists (your classmates). The specific nature of the (supervised) prediction problem for the project is a *classification problem*. There are several different types of modeling methodologies that can be applied to solve this type of problem and your group will have to choose a variety of methods and assess them.

You will be provided with a large set of training data (with the inputs/features and outputs) as well as a set of test data, which includes only 20 cases of input features only (no outputs). The data set should be somewhat familiar. Both files can be found in the course repository. Your objective is to build a model from the training data and use the trained model to predict, out-of-sample, on the test dataset. You should do this according to the following parameters:

- Each group will submit a single RMarkdown (.Rmd) file as a single document knitted in either html or pdf formatted output titled **prediction.Rmd**. You can find details related to this type of output file in RStudio here: https://rmarkdown.rstudio.com/authoring_quick_tour.html.
- The document should contain R code chunks (just like in a usual R Script) as well as written narrative between the code chunks to describe and document the steps taken in your analysis.
- Each group member will propose, estimate, and test/validate a unique model to perform the classification task using the training data with the intention of applying the model to the testing data. From the five models your group proposes, you will select one of them (ideally the best one) to apply to the test data to see how the algorithm performs "out-of-sample."
- You will be assessed on your effort with respect to your entire group as well as your group's final proposed model's performance on the test data.
- Your document should include 7 sections. The first section should include an executive summary and proper documentation of any steps taken to preprocess/clean or explore the training data. The next 5 sections should include analysis of five different models, each proposed, estimated, and testes/validated by a *different* group member (clearly identifiable in the document), appropriately commented in the R code chunks and with a properly constructed narrative between code chunks where applicable. The final section should include a discussion of how you chose the final candidate model as well as the output of the predictive algorithm when you run the model on the test data.
- You should push your RMarkdown document with all of your analysis and results to one of your group member's repositories on GitHub and provide the link to slevkoff@sandiego.edu. The deadline to submit the project is next Sunday, 5/13/18, before midnight.