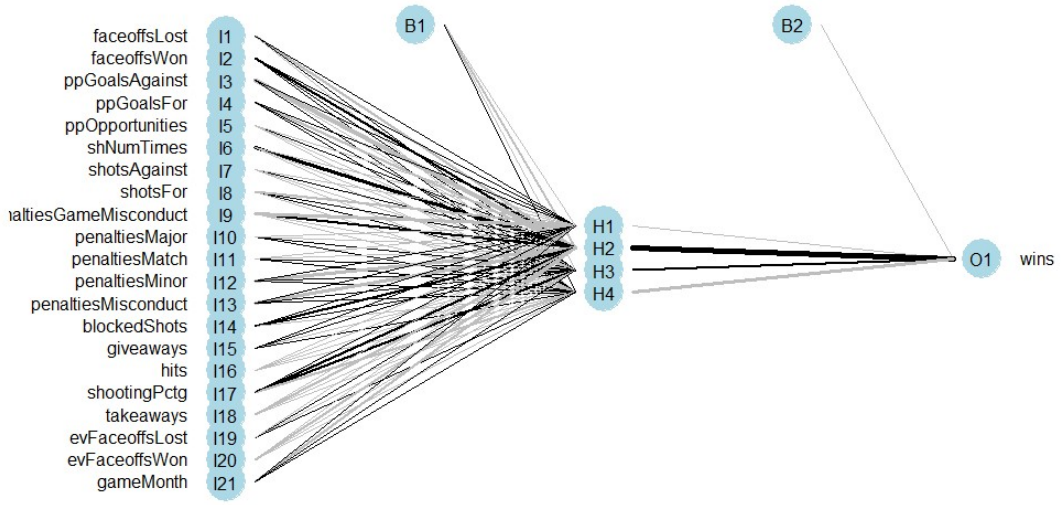
From the start of the 2008-09 season, there have been over 10,000 games played in the National Hockey League (NHL) and for each of those games a standard set of statistics has been reported along with the score in what is called the "box score". This is what would be published in newspapers and in the past, it would be used by the public to see how games they weren't able to watch played out. But are the standard box scores useful? Do they really tell the story of the game? Can you look at the statistics alone and determine who won the game? That is what I set out to do in this project. This can be a fun project to see if we are looking at the right statistics but also useful for NHL clubs to see if there are areas where they should focus.

To begin, I needed box score data from a bunch of games to get a large enough sample size to be able to train and test a model. I discovered quickly that NHL.com has the most complete data I've ever seen. I scraped data from 2008 through the end of the 2017-2018 season and was pleased to discover that there were no missing values which increases my confidence in the model. The cleanliness of the data also allowed me more time to focus on changing parameters and trying to figure out the best ways to measure success of the models.

R was utilized in this project to import, clean and manipulate the data as well as to build and test the models. The dummies package was utilized to one-hot encode values and the tidyverse and lubridate packages were utilized in exploratory data analysis. The e1071 package was utilized in building the SVM model while the nnet package was used for the neural network model. To visualize the neural network, the plot.nnet function from Marcus W Beck (https://gist.github.com/fawda123) and also his gar.fun function was utilized for finding the importance of variables in the nnet model.

More details of my project are available in the attached presentation. In the end, I had some success predicting the outcome of the games. The accuracy for each of my models was around 80% with the best performing model predicting 81.1% of the test games correctly. Some interesting data

was found when looking into the importance of each variable with power play goals against, shooting percentage and game misconduct penalties having the largest impact on the outcome and faceoffs lost, game month and hits having the lowest impact. The visualization of the final neural network is shown below.



References

Beck, M.W. (August 12, 2013) Variable Importance in neural networks. R-bloggers. Retrieved from https://www.r-bloggers.com/variable-importance-in-neural-networks/

Koshy, J. (Nov 11, 2016) Web scraping: Best practices to follow. Prompt Cloud. Retrieved from https://www.promptcloud.com/blog/web-scraping-best-practices

NHL.com (2018) NHL.com/stats. Teams game by game results. Retrieved from http://www.nhl.com/stats/team?reportType=game&dateFrom=2008-10-01&dateTo=2018-04-30&gameType=2&gameLocation=H&filter=gamesPlayed,gte,1&sort=points,wins

ProgrammingR (2018) "Web scraping R data from JSON". Retrieved from http://www.programmingr.com/examples/reading-json-data/