SakaiAtMidnight 521 Final Project Proposal

Leonardo Shu
Department of Statistics
Duke University
Durham, NC
leonardo.shu@duke.edu

Mengrun Li
Department of Statistics
Duke University
Durham, NC
mengrun.li@duke.edu

Yaqian Cheng
Department of Statistics
Duke University
Durham, NC
yaqian.cheng@duke.edu

Wei (Emily) Shao
Department of Statistics
Duke University
Durham, NC
wei.shao@duke.edu

Abstract

This article is a proposal with preliminary ideas and plans for our final project. We choose a data set of NCAA basketball and hope to predict the champion of the 2015 NCAA tournament. We have come up with some motivation questions about selecting variables and considering time effect. According to the details of the data set and what we have learned from the course, we plan to use regression with bayesian approaches and do classification or clustering to achieve our goal.

1 Introduction

The topic of our team final project is predicting the results of the 2015 NCAA tournament. We chose this topic because all the members of our team are interested in sports and two of them are very familiar with basketball. Everyone in Duke feels proud of our team's achievements, so do we. Therefore, when we found the data set of NCAA basketball that would allows us to work on this topic, we decided on it immediately. For our final project we want to apply what we have learned to our field of interest.

Since we know what happened in the 2015 tournament then we have a complete set of results with which to compare the prediction our model produces and hopefully we can also take a stab at predicting next year's edition as well.

The data set we will be using is provided by a past competition on Kaggle.com, which ensures that the data is both reliable and rich enough for doing our predictions. Specifically, there are more than 5000 pieces of game data for each year and 34 columns of many other kinds of information for each game including 26 technical skill indexes, which will be useful for us to compare the performance of each team. Except for the detailed data of regular season for 2015, there is also detailed data of both regular season and tournaments from 2003-2014, which can be helpful if we want to consider or compare the performance of past years. With this data set, we can use many methods that we have learned from the class, such as regression and LASSO. Therefore, it is an appropriate data set for the final project.

2 Motivating Questions

2.1 Question 1

Teams in the NCAA change rosters very quickly which means that historical data of past seasons/tournaments has diminishing returns. For example, how useful will be the results of a season that happened 12 years ago in predicting this year's results when all the teams players (possibly coaches) may be completely different. Therefore, we want to know how useful and to what extent should we use historical data in our analysis.

2.2 Question 2

Another question is how useful are each individual game metric to our model? Are number of overtimes in a game really significant for our predictions? One way we might want to tune our model is by using backwards-step methods to identify the best model to use.

2.3 Question 3

As we use the model to predict the historical results of tournament in 2015 and test how well our model performs, we also want to know whether this model can potentially predict results of tournament in 2016. If the model we build is robust and works for future prediction, once regular season results in 2016 comes out in future, we are interested to know whether the prediction will still be valid, or whether we need to revise our model to adjust for any variability in the data.

3 Methods

For the first motivation question, we may use some Bayesian methods. For example, we can use the historical data from 2003-2014 as our prior information and use the data in 2015 regular season as our observations. Or we can put different weights on the historical data we use. For example, put higher weights on data collected from more recent seasons.

For the second motivation question, we may build a simple model with fewer variables and an advanced model which includes a mixture of more variables. The model we choose can be complexed regression models. Since we include different variables in each model, it imposes the necessity to apply more robust variable selection methods in the analysis, e.g. backward-stetp methods, LASSO.

For the third motivation question, we simply use the same model to predict tournament results in 2016 once regular season results comes out.

4 Data Links

https://www.kaggle.com/c/march-machine-learning-mania-2015