DS 301 Project Proposal

Future ESPN Analysts

Ryan McNally: rmcnally
Shiv Neelakantan : sneel
Saketh Jonnadula : venkataj
Ryan Scehovic : scehovic

I. Data Set Intro/Overview

Data Set Link: Kaggle CBB data

Data Set Overview: This dataset is Division 1 men's college basketball data collected from 2013-2019. There are 24 unique columns in this data set. Some are identifiers like team name and conference, and some are statistics collected from the team's season like 2-point and 3-point shooting percentages, the number of free throws, and offensive and defensive rating and efficiency. The data set also includes information about the team's postseason performance and seed in the March Madness tournament. This will be a focus of our project.

II. Questions & Exploration

II.a

Question: Can we predict if a team will make the postseason tournament based on their regular-season performance? *(Classification)*

Insights into the question from data exploration: Efficiency metrics for offense and defense, power ranking, and conference, all play a role in determining what seed a team gets. We will continue to explore this and find which predictors have the most impactful relationship in predicting what seed a team will be in the tournament.

Methods We'll Use:

1. Logistic Regression
2. QDA

II.b

Question: Which predictors are most significant for a linear model that predicts the number of total wins a team will have? *(Regression)*

Insights into the question from data exploration: We learned what some of the best predictors were in winning teams. The predictors for power ranking and offensive/defensive 2pt and 3pt efficiencies were shown to be tied to winning teams.

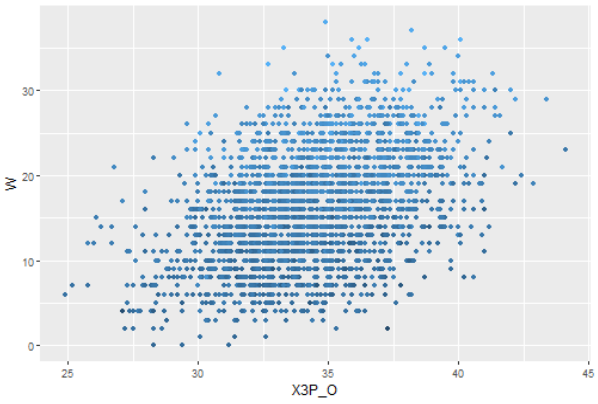
Methods We'll Use:

1. Lasso/Ridge Regression
2. Subset Selection

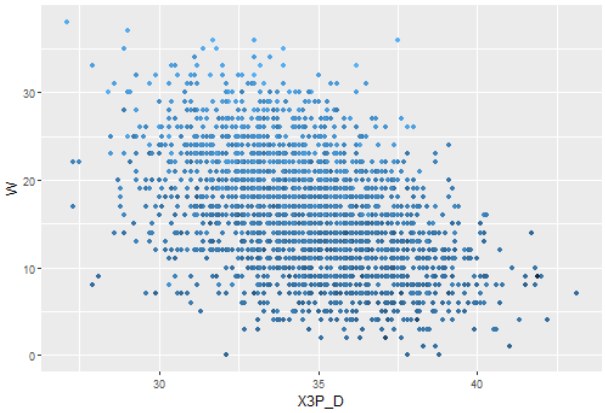
Graphs

Below are graphs for the number of games a team won and the team's offensive/defensive 3pt percentages.

Offensive 3pt Percentage



Defensive 3pt Percentage



For offensive 3pt percentage, there is a general trend up and to the right, meaning as teams make more 3-pointers, they are more likely to win. The defensive 3pt percentage shows a general trend of up and to the left, meaning that the fewer 3-pointers a team allows then the more games they should win.

Potential Problems

To wrap up our exploration we looked for potential problems with the data. The most obvious potential problem is how the postseason column uses letter and number combinations so which makes it useless for any regression models because it takes each different column term (Champion, 2nd, F4, E8 S16, R32, R64) and treats it as it's own predictor when really it'd be better off if each of those were just a numerical value associated with how the team finished. This is a change we can implement ourselves and allow us to see if the column is a valuable predictor in determining the number of games a team wins.