Matt Russo

8/6/20

Dr. Clachar

Final Project - Phase 1

As we have learned throughout the course, Machine Learning can help identify meaningful features in data sets. Having the ability to understand the features that are most influential towards future performance/past performance is what we all strive for when working within the Data Science field, specifically Machine Learning Algorithms. For this Final Project, I am going to choose to work in the sports industry and use college basketball team statistics to predict which team attributes were most influential in correlation to the number of wins a team had. The dataset will involve the last five college basketball seasons (2015-2020) for every team in the NCAA Division 1. After writing my research paper on the use of the double team in basketball and the evaluation of the performance metrics in the NHL, I thought this would be an interesting follow up. The number of wins a team has goes much deeper than just scoring more points that the opposition. There are many team statistics that are taken account throughout the course of a game/season and I want to go in depth to see which can generate a high accuracy model compared to the wins they had. Applying algorithms like Decisions Trees, Logistic Regression, Select K best and Support Vector Machine (SVM) are some of the models that I am thinking of experimenting with. I look forward to working with the dataset I found and hope to understand the team statistics that seem to most influence the wins a team has.

The motivation for choosing this topic is simple and it is due to me being a passionate sports fan. I thought this would be a unique opportunity to use Machine Learning algorithms to build my case and even learn more about the relevance of each statistic that is recorded throughout games. I know that working throughout the various tasks would be enjoyable if I chose a topic that piqued my interest. In addition, I would definitely be open and would like to enter the sports industry in a future job position. Providing this type of valuable information to an organization would be really unique and would add to the positive trend of analytics sports are continuing to see. I can envision the positive feeling that would come from being able to supply critically significant information to an organization. I know this project will provide useful exposure and experience into some of the tasks I would be working on in a future position in the sports industry.

The goal of the project is to find a Machine Learning model that achieves a high level of accuracy on the data set. Being able to generate a high accuracy model will allow us to feel confident in understanding which features (team statistics) are most influential in predicting the number of wins a team had. Similar to what we have done in the programming assignments we want to understand the importance of each feature in correlation to the given model that is being used. I can anticipate that there will be certain statistics (features) that I would not have expected to be so critical in predicting wins for a team. The first objective in order to solve this problem is to understand the data set you have. Without comprehending the attributes, statistics used and other various elements in the data set, we will not be able to make the right decisions when it comes to our model. Luckily, since I am a fervent sports fan, I know almost, if not all statistics and so it will be easier for me to narrow down which features will play the biggest role in the chosen algorithm. Data preprocessing as well as handling missing data will play a key role like it has for us all semester. I look forward to collecting data and start working on this project and walk you through the steps I took to see the stats (features) that ended up being the most relevant in predicting the actual number of wins these teams had.