A single basketball game contains many metrics that can be used to predict which team will win. Our goal in this project is to predict the outcome of a (NBA) match using box score statistics. We use various machine learning techniques to predict game results and to understand which features of a team make it successful. We want to figure out which attributes or stats of a team will be critical for winning against the other team.

Through the years a lot of data and statistics have been collected based on NBA and each day the data become more rich and detailed. Therefore it is very complex to analyze and try to predict a game. In order to deal with that complexity and to achieve better predictions rate a lot of Machine Learning methods have been implemented over these data. That is exactly the purpose of this project.

National Basketball Association (NBA ) is the men's professional basketball league in North America. The influence of NBA trespass its borders and have countless fans all around the world. As the league involves a lot of money and fans, a lot of studies have been developed trying to predict its results, to simulate winning teams, to analyze player's performance and to assist coaches.

There are 30 teams and Within each conference, the eight teams with the most wins qualify for the playoffs. The seedings are based on each team's record. Each conference's bracket is fixed; there is no reseeding. All rounds are best-of-seven series; the team that has four wins advances to the next round. the team with the better regular season record takes Home court advantage.

The regular season consists of each team playing 82 games, 41 at home and 41 away. The NBA schedule allows for each team to play every other team in the league at least one time at home and one time away. Between all NBA teams there are a total of 1230 games during the regular season.

Data Sources

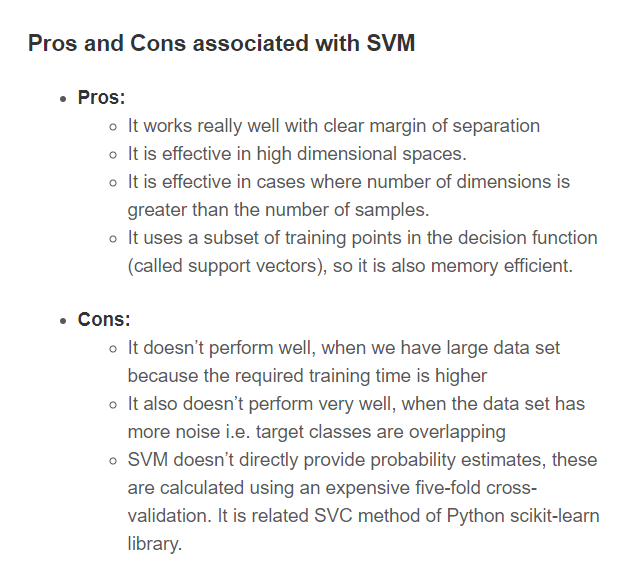
In this project, data used for model training and testing, Basketballreference.com is our major data source. Basketball-reference.com is one of the most famous one for those sports data miners, and has been used in many experiments. It was created in 2003 and data requests are comprehensive, relatively well organized, straightforward and easy to navigate and utilize.

SVM

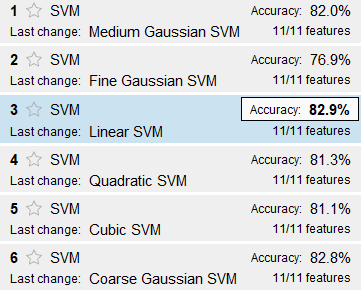
We also modelled our problem with a support vector machine to predict whether or not a streak will occur. To date, few supervised learning algorithms have outperformed SVMs; perhaps the most alluring property of SVMs is that we can utilize their symbiosis with kernels. This lets us create high or infinite dimensional feature vectors, allowing us to capture the subtle interactions between features, which is particularly important for our dataset because it contains so many intertwined and interdependent features. Thus, the SVM presents an ideal classification model that does not limit our ability to utilize a dense, high dimensional feature vector to determine the state of the game.

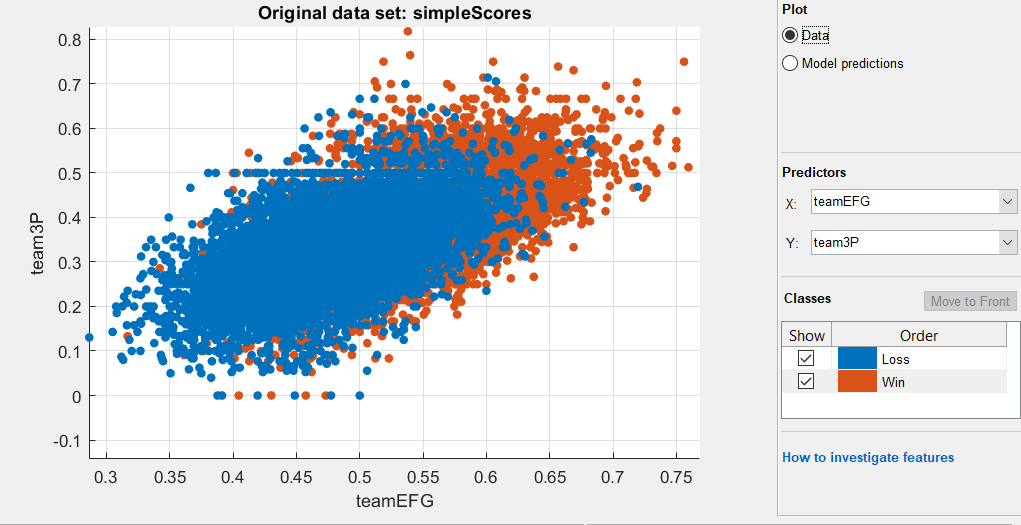
One disadvantage of SVMs, however, is that they are easily susceptible to over fitting because they work in such a high dimensional feature space. Since it is likely that the extracted data will become linearly separable, we run the danger of the model learning hidden attributes of our data rather than the more general trends we are looking for. SVMs constructed in this manner work to maximize the geometric margin for every point by constructing a separating hyper plane in high-dimensional space (as most data sets are not linearly separable) that classifies as many points correctly with the largest margin possible.

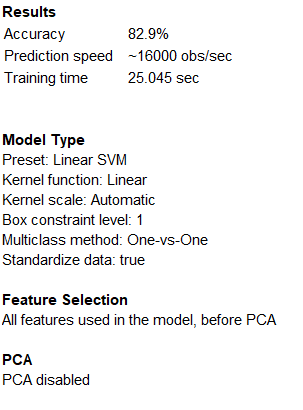
Sometimes it is desirable to select the most important features or also to reduce its number using methods as Principal Components Analysis. After the data is prepared some analysis will be done in order to select the best inputs for the methods. Secondly, the Machine Learning Methods will be implemented. By looking in some articles cited on the references, it seems interesting to start the process with a linear regression, that represents a simple method that has, so far, achieved a good performance.

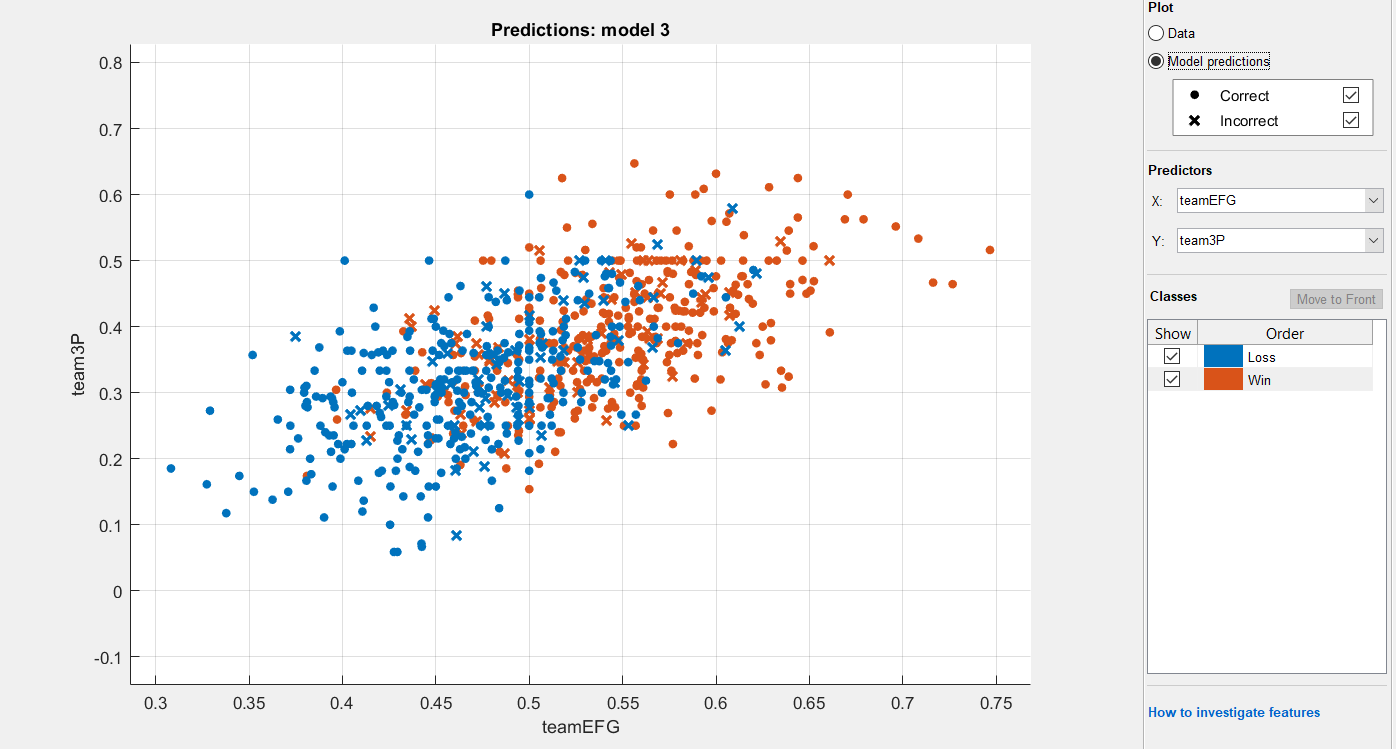


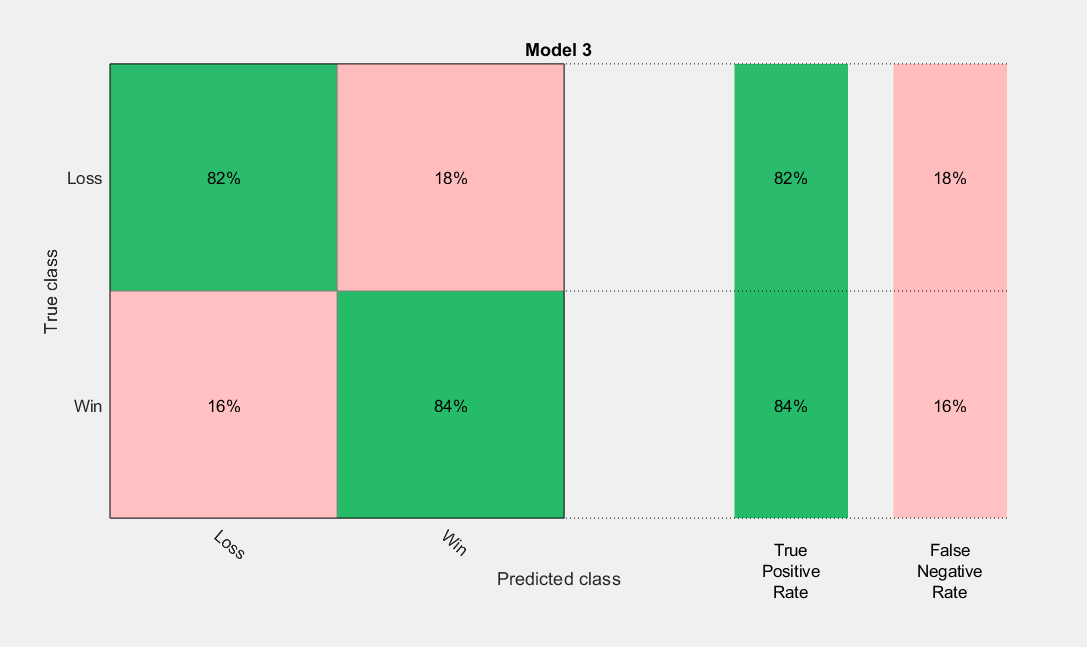
%5











%20

