Project Midterm Report.

Andrew Kozma
Akash Nadan
Friday, Oct. 28, 2016

Description of the Dataset
Our dataset consists of every shot taken by every NBA player from October 28, 2014 (opening day) to March 4, 2015. Each row of the dataset is a single shot, but the features of the dataset contain data from that shot, and from the game in which that shot was taken. The shot data is compared to the game data in a summarized table below.

Data from each game	**Data from each shot**
The game number	Shot number
The opposing team	Period number
Home or away	Game clock
Win or loss	Shot clock
Final score of the game	Number of dribbles before shooting
*Number of shots made	Time held onto the ball before shooting
*Number of shots attempted	Distance of the shot
*Number of points made	2 point or 3 point shot
*Number of points attempted	Distance of closest defender
	Who is the closest defender
	Was the shot made
	Which player took the shot

* This feature was added to the data

Avoiding Over and Under Fitting
To avoid overfitting, we will not use a complicated model. Instead of creating one single model that uses all the features of our data, we will create multiple models with fewer features. Furthermore, we will limit the dimension of the input space to small values. For example, in creating a linear model, we will look for quadratic or cubic linear models, instead of 10th-order linear models, and utilize regularization.

Testing the Effectiveness of Our Models
To test the effectiveness of the models, we will test our models on a test set. We will use a training set to pick the best model and then validate our model on a test set. This ensures the model isn't trained on the training set, and can generalize to any future NBA shot taken. In order to create our training set and our test set, we have split up the data. The test set consists of all the shots taken from every fifth game of the season, and the training set has data from the rest of the games. By doing this are training set contains 80% of the data and our test set contains 20% of the data.

Histograms and Descriptive Statistics of the Data
The average shooting percentage over all players in the dataset is 45% with 48% shooting percentage on two pointers and 35% shooting percentage on three pointers.
The farthest shot was taken from 47.2 feet from the hoop and the closest was taken from 0.0 feet away, a dunk. These are general statistics that we used during our analysis.

Features and Examples
Our dataset consists of 128,070 shots taken in every NBA game from October 28, 2014 (opening day) to March 4, 2015. This means there are 128,070 examples in our dataset. Furthermore, there are 17 + (the ones we added) features to our dataset.

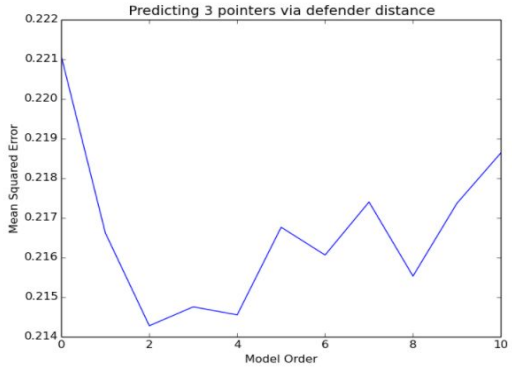
Corrupted and Missing Data
The only data that is missing or corrupted comes from the shot clock column. We know that the data is corrupted because the mean of that column returns a value of "NA". But, less than 5% of the shots taken in our dataset come when the shot clock doesn't have a value. This "NA" value occurs when a team gains possession of the ball with less than 24 seconds to go in the period, which eliminates the need for a shot clock.

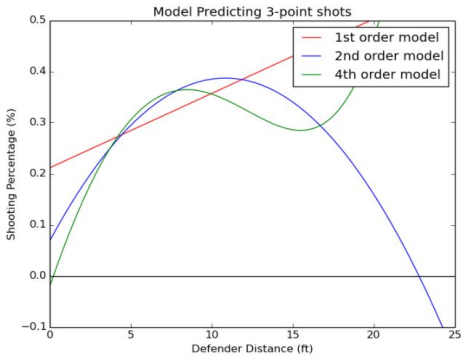
We also found that some of the data is messy, but it seems that the data was recorded incorrectly. Some of the data that was recorded as being a potential 3-pointer is from under 10 feet from the hoop. This does not make sense because the 3 point arc is 22 feet away from the hoop. This could potentially be due to the situation where a player is fouled while shooting a 2-point shot and gets to shoot a free-throw to try and gain one more point. This feature would have to be taken into consideration when we run our models.

Preliminary Analyses

First, we performed some feature transformations of the data by adding multiple columns to the dataset. We recorded the boolean value of both the location and the result of the game: 1 for home and for win and 0 for away and for loss. We added a column that recorded the shots made and attempted, the points made and attempted, and for the player's shooting percentage that game.

The first model that we ran on the data was fitting a polynomial model using one feature of the data in order to predict another feature of the data. We chose to predict if a three-point shot is made based on how far away the defender is. A model was fit using the least squares for every polynomial order from 0 to 10, and the mean squared error was calculated on the test set. The mean squared error for each model order is plotted below in the left graph.

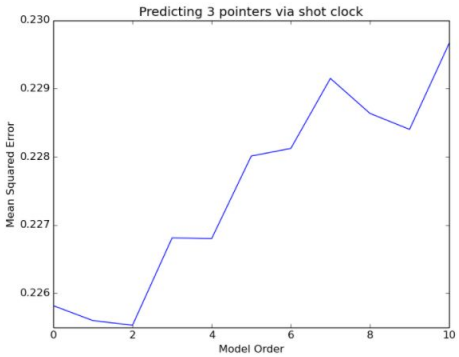


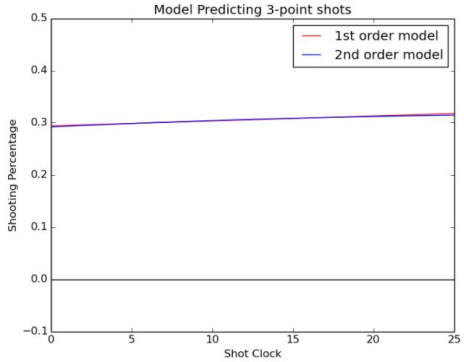


The models with the lowest error on the test set were second and fourth order models. These 2 models, along with a linear model are plotted in the graph above on the right.

The linear model is interesting because it predicts around a 20% 3-point shooting percentage for a defender 0 feet away. The downside to this model is it increases forever, and thus would eventually predict over a 100% shooting percentage. The 2nd order model had the least mean squared error on the test set. This model provides a better explanation for the actual scenario of taking a 3-point shot. The marginal increase in shooting percentage for every extra foot the defender is from the shooter decreases, and no matter how far the defender is, one's shooting percentage will never be above 50%.

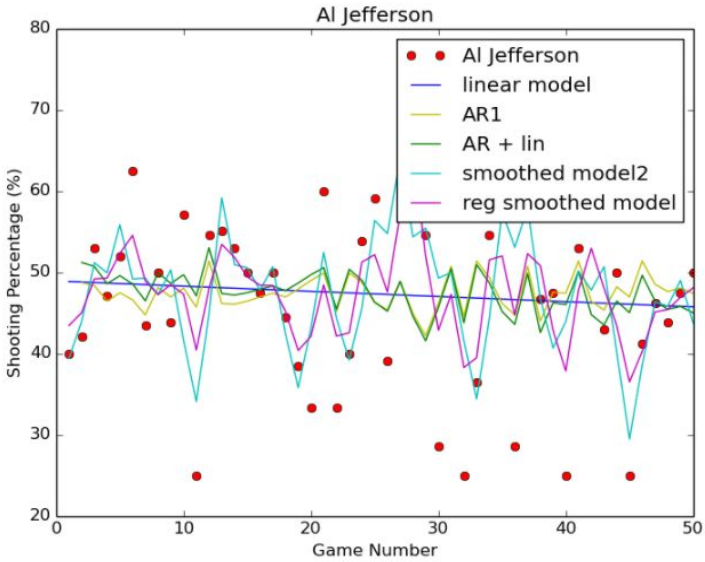
We also computed a model to test how the shot clock affects 3 point shooting, of increasing polynomial order. It is expected that with more time on the shot clock, the shooter doesn't feel as pressured into shooting and will have a better chance of making the shot. The mean squared error of each model can be seen in the plot below on the left.





The model with the least mean squared error is 2nd and 1st order, so those were plotted for values of the shot clock, as seen above on the right.From this graph, there doesn't seem to be much relation between the value of the shot clock and the likelihood a player is to make a 3-point shot.

We also ran a model to predict the shooting percentage of a player over the course of the season. It would be useful for this model to take into account the ups and downs that a player experiences over the course of the season. Unlike the previous models we ran, this one was using game data, specifically the shooting percentage that game, instead of shot data, as in the previous model. Below is a plot modeling one player, Al Jefferson's, shooting percentage over the course of the season.
This model is very similar to the analyses we did on the homework and during class. The linear model



uses just the game number to predict the shooting percentage. The autoregressive model (of lag 1) uses the shooting percentage of the last game to predict the shooting percentage in the next game. Included is also a combination of the two: autoregressive plus a linear model. Also, a smoothed and a regularized smoothed model (using Ridge Regression) is included in the plot.

Further Steps
For the rest of the semester, we plan on incorporating the different skill level of each player in the model. There are different ways we could go about doing this. One way is to find online rankings of the player, use their seasonal shooting percentage to rank them, or add a value to indicate if they were selected to the All-Star team this season. This feature is important because all players are not identical in skill level. Some players are simply more likely to make shots from all over the court and in the same scenario as another player.

We plan on running models to predict 2 pointers as well as 3 pointers. So far, all the analysis run has been predicting the latter, but most shots taken during a game are 2-point shots.

A feature we have still yet to include in our models is the amount of time a player had the ball before shooting. We believe this is a useful feature in predicting if the shot was made in a combination of other features. An example of this situation is if the player had the ball for a long time and the shot clock was very close to zero then this could indicate that they are rushed into taking the shot. Along with this, we plan on using other reasonable feature transformations in 2 dimensional space.

We are eager to implement other tools in our model. As we learn new concepts in class, we will implement those in our dataset in order to create more models for NBA shots. Included in this is the loss functions and different ways of regularizing the data. The different loss functions will be useful because there is some messiness to our data, as discussed in a previous section.