Akash Alaparthi, Neel Avancha, Aryan Kale,

Aneesh Ponduru, Druv Sarin

Dr. Clachar

Foundations of Data Science

20 November 2022

Phase 3: Model Selection, Training, and Evaluation

The name of our project is **AssistMe**. For our project, we originally wanted to use the stats from the NBA season to predict the winners of certain awards such as MVP (Most Valuable Player), DPOY (Defensive Player of the Year), Sixth Man of the Year (6MOTY), and Most Improved Player (MIP). After taking a look at the machine learning models that we have explored in class and what we have learned so far, we believe that this may be too difficult and have decided to switch directions with our project. We will still be working with NBA players and their stats, but instead, we will be using their major statistical categories to help us, or 'assist' us, in finding players' average assists for the year. Assists is one of the most important statistical categories for NBA players, and our model will attempt to predict this value for many different players using their other statistical categories. 3 algorithms that we will explore are Knn-neighbors, random forest regression, and support vector machines.

Knn-neighbors is an algorithm which can be applied to both classification and regression predictive models. In our project, we would need to utilize a Knn-regression model to predict assists because assists would be considered as continuous data. The Knn regression algorithm works by predicting a new data point's continuous value by returning the average of the k neighbors values. We can tune this model by adjusting the knn-neighbors to adjust the model's ability to predict a certain point. We can do this by experimenting with various neighbors and see

which one is best at predicting the data using an iterative for loop which evaluates different k values in our predictive model.

Random forest regression is a machine learning model which utilizes many different decision trees. Each decision tree compares the values of each of the features and based on whether they are greater or smaller than a certain value (for numerical features), it will either continue down the tree with the other features or make a decision for what the value of the variable should be. Each of these trees is created by the training data. The random forest uses many different regression trees and then averages them to predict the desired value. We can tune this model by changing the number of decision trees in the forest. This can have a great effect on the accuracy of the model, and we would have to experiment with different numbers of trees to see what would be ideal.

The support vector machine algorithm is a unique model which is able to find a hyperplane in an N-dimensional space that distinctly classifies the data points to predict. If the number of input features is two, then the hyperplane is just a line. If the number of input features is three, then the hyperplane becomes a 2-D plane. It becomes difficult to imagine when the number of features exceeds three. We look for a hyperplane which is able to separate the data points into two classes the most effectively. We measure this by finding the hyperplane which has the furthest distance between the two closest points of opposite classes.

The machine learning model that we believe will be best for our project is the Knn-neighbors regression model. For our project, we are using data from the 2021-2022 NBA season for our project. In order to train our model, we will take roughly 70% of the data, which in our case is the NBA players. We will use this percentage of players to train the model and then evaluate the model on the remaining 30% of the players. We can then compare the predicted

number of assists by the model to the actual number of assists for the test data and then see how our model performed. As previously mentioned, the most important hyper-parameter that we need to tune is the number of neighbors that we use to make each prediction. We will try various different numbers of k for neighbors of each new value until we find the one that yields the highest accuracy.

The measure of accuracy that we will be using is mean-squared-error (MSE). This measure of accuracy works well with regression models and we are expecting our MSE to be under 0.10. This would mean that our model is very accurate and we expect this to be the case since we are taking so many different statistical categories into consideration when training the model. This number could be even lower, but there are many other variables other than stats that affect how a player performs during a given season. For example, a player could be coming into a new season off of an injury, in which case they might be slightly out of shape and practice which would likely skew the player assist stats. Additionally, a player can get injured anytime during the season, and if they face a significant injury, then they will have stats that will only represent a fraction of playing time. All things considered, we believe that this is a good starting benchmark and as we begin to evaluate models on the data, we will see whether we are able to stay below 0.10 for the MSE.