Aidan Chandrasekaran, Casey Ng, Brandon Tat, Eric Tran

29 November 2020

DATA 301

Dr. Stanchev

DATA 301 Project Write up

We decided to build models to determine if a player will be an all-NBA player. This is an award that goes to the 15 best players in the league, 5 per position. We wanted to see how big of a statistical difference there was between All-NBA and non All-NBA players because they are voted for by broadcasters and sports journalists, meaning that more than just how well they play is important.

To start out, we pulled from two different datasets: one had detailed player stats from 1977-2020 and one was a list of all the players who received All-NBA awards. Since older data might not be as accurate and there were changes in how basketball in the NBA is played, we decided to cap our data to be from 2000-2020 so that our model was best at determining if a player is All-NBA in the "modern" era of the game. After data cleaning and merging the two datasets, we had a dataset that was prepared for our models. In addition, all data was normalized using the z-score. The stats given to both models to train on were P/36, MP%, %Pos, Min, PIE% and USG%. Each of these statistics were hand selected based on their correlation with All-NBA. These variables were among the most highly correlated with whether or not a player was All-NBA. Furthermore, both models were given the datasets, then randomly split them to ensure the models were not being tuned to static datasets and giving inaccurate results.

We decided on 4 different methods: KNN trained on unbalanced data, KNN trained on balanced data, NN trained on unbalanced data, and NN trained on balanced data. This would

allow us to see which model performs the best and to see if the balancing of the data has a role in how accurate our models become.

The unbalanced neural network was trained on 80% of the entire cleaned dataset, and tested on the remaining 20%. The network itself was composed of three hidden layers with optimizer 'rmsprop' and loss 'binary_crossentropy.' In addition, class weights were calculated using sklearn; in effect these weights forced the neural network to pay more attention to the minority data class, in this case the players labeled as All-NBA. The neural network trained on the balanced dataset performed similarly to the unbalanced neural network. However when we tested both models on balanced data, the balanced NN did significantly better at 94% accuracy vs unbalanced which was 54% accurate.

For our KNN models, we did a similar set up like our neural networks where we tested both balanced and unbalanced data. We used 5 nearest neighbors as our parameters because anything less was giving too much variance in our results and anything more had no noticeable effect on our results. We had a balanced and unbalanced trained model and we tested both on the balanced dataset. Along with testing both on the balanced dataset, we tested the unbalanced trained model on the unbalanced dataset to see how it would perform. The unbalanced model on unbalanced data performed really well at 98% accuracy. For the balanced data, the balanced trained model was 94% accurate while the unbalanced trained model was 56% accurate.

After seeing the results from our models, we determined that both models on balanced data performed similarly well so either model would work well to predict an all-NBA player. However, both unbalanced models performed poorly when tested on balanced data. We suspect that the accuracy of the unbalanced models may have been superficially increased by the large majority group, which the accuracy score cannot account for. The upside to using unbalanced

data is that the dataset is several times larger than our balanced dataset. This is due to the small minority and a supermajority between the two classes within the data. In a situation where we had too little data to create a balanced dataset, a solution would be to artificially increase the size of the minority by generating data points that are close to the minority class. However if given the choice, our results indicate that it is best to use balanced data to ensure an accurate model for either KNN or neural networks.