DATS 6202 Final Project Proposal - Group 4

What problem did you select, and why did you select it?

We selected a dataset containing shots taken by former NBA player Kobe Bryant who is regarded as one of the greatest basketball players of all time. Bryant won five NBA championships, was an 18-time All-Star. We found this dataset on Kaggle, and we found it interesting as sports fans ourselves. Kobe Bryant transcended the game of basketball and became a worldwide icon, and he tragically passed away in 2020.

Kobe Bryant led the NBA in scoring twice, and ranks fourth in league all-time regular season and postseason scoring. We can apply our curiosity and passion, along with our neural network skills, to find insights in this data and try to find Kobe Brayant shooting techniques and patterns. We will also do an opponent analysis, diving into his best and worst performances against specific teams. We will try to understand why he performed well or poorly. This could have been helpful to opposing teams as they created a game plan to stop Bryant.

The dataset contains categorical variables (such as shot zone), and numeric variables (shot distance). This allows for us to practice pre-processing and scaling that comes with most neural network use cases. Furthermore, the categorical variables will provide the basis for our exploratory data analysis and help us select features to put into the model.

What database will you use? Is it large enough to train a machine learning network or different algorithms?

Each row corresponds to one shot taken by Kobe Bryant, with the target variable being whether he made the shot or not. The dataset has approximately 30,000 rows, with 5000 held out for testing in the Kaggle competition. We will use the 25,000 labelled rows to train and test our model. This is enough sample for most neural network techniques.

What neural network will you use? Will it be a standard form of the network, or will you have to customize it? What algorithms will you use?

As we have a binary classifier, and we believe we can take advantage of multiple decision boundaries, we plan to use a multi-layer perceptron classifier. We plan to use the standard form

that comes from the sklearn library in Python, although we will alter parameters in the base library to see what approaches improve model performance.

We plan to use a conventional model, such as a random forest to compare our neural network to.

What software will you use to implement the neural network or different algorithms? Why?

We will use python, sklearn for pre-processing and modeling, as sklearn is covered in our class. We are aware of libraries such as keras, but those are out of the scope of this course.

What reference materials will you use to obtain sufficient background on applying the chosen network or algorithm to the specific problem that you selected?

We will read about multi-level perceptrons to understand the technique and its limitations. We will read about NBA shooting, and what factors may influence whether a player makes or misses a shot. Combining the technical and domain knowledge will allow us to take on this problem. We referenced code for our shot chart visualizations, such as here and here.

How will you judge the performance of the network? What metrics will you use?

We will use mean squared error for our neural network, as compared to accuracy/precision/recall for our random forest.

Provide a rough schedule for completing the project.

June 16: Complete exploratory data analysis
June 19: Complete both neural network and random forest
June 20: Start compiling graphs and findings
June 22: Put all findings together into a written report and create slides
June 23: Review slides and prepare for final presentation