SENG 474 Project Proposal

NBA Shot Expected Point Value Regression

Michael Reiter
Faculty of Engineering
University of Victoria
Victoria, Canada
mreiter@uvic.ca

Spencer Vatrt-Watts
Faculty of Engineering
University of Victoria
Victoria, Canada
asvw@uvic.ca

Erik Reppel
Faculty of Engineering
University of Victoria
Victoria, Canada
ereppeld@uvic.ca

***Abstract*—Given a basketball shot in an NBA game of known features, use regression to predict the expected point value (0 - 3). Data of shots taken during the 2014-2015 NBA season is sourced from Kaggle and includes 120,000+ data points.**

Keywords—software engineering; data mining; machine learning; basketball;

I. Project Description

The aim is to predict the expected value of a given NBA player's shot based on historic data. The dataset does not include free throws because they are far more straightforward to predict and depend on fewer variables. Since shots can be worth either 2 or 3 points, the regression will produce a value in the continuous interval between 0 and 3 inclusive. Rather than classifying a shot as statistically favourable or not, a continuous value is a better metric because it allows for an absolute expression of shot quality. For example, a contested 3 point shot, while having a higher potential outcome, is worse than a 2 point shot near the hoop. The expected value allows clear comparison between shot classes taking into account the quality of the shot.

II. Related Work

On the Kaggle dataset page [1], other individuals have explored the data in an effort to draw various conclusions. Some examples include classifying defenders based on average shot percentage from a distance with a given nearby defender, and expected expected point values for a shot by a given player. The latter only takes into account which player was shooting, ignoring relevant features such as distance to defenders, and distance the hoop. Thus it cannot be generalized for any given shot. This project aims to develop a model that is better suited a general shot case, not one from a specific player.

IV. Proposed Project

A. Data

A 128,000 data point dataset of shots taken during the 2014 - 2015 season has been sourced from Kaggle.com. It was originally scraped from the NBA's REST API [2]. It contains relevant features such as the player who took the shot, the distance to the hoop, the distance to the nearest defender, the time on the shot clock, etc. It is possible that further data acquisition will be deemed necessary as potential algorithms are explored.

B. Algorithm

Since the goal is to predict a value, rather than classify a shot, the project will make use of some form of regression. Linear regression using Gradient Descent to optimize the function could be employed [3]. Alternatively, a decision tree or random forest regressor could be used. Based on preliminary evaluations, linear regression will likely produce the best result.

C. Evaluation

The dataset will be split into testing, training and validation sets. Various regression algorithms will be tested and evaluated for accuracy.

IV. Estimated Timeline

Acquire initial dataset - September 30

Acquire any necessary supplementary data - October 7

Clean and validate total dataset for completeness - October 14

Explore potential features of interest and investigate potential regression algorithms - October 21

Train algorithms and evaluate prediction performance - November 7

Tune features and parameters for tested algorithms - November 14

Decide on final algorithm and perform any necessary tweaks and implementations - November 21

Finalize and submit report - December 2

IV. Distribution of Tasks

Responsibilities will be divided as equally as possible with specific members leading various tasks. Spencer will focus on data acquisition, cleaning, and validation. Michael will explore features and algorithms of interest. Erik will lead the training of prospective algorithms. The group will collectively decide on a final algorithm. Once that has been decided, any remaining implementations and tuning will be completed. The production and editing of the final report will be split three ways.

REFERENCES

1. "NBA shot logs | Kaggle", Kaggle.com, 2016. [Online]. Available: https://www.kaggle.com/edwardyun/d/dansbecker/nba-shot-logs/the-best-and-worst-defenders. [Accessed: 04- Oct- 2016].
2. "NBA.com/Stats", Stats.nba.com, 2016. [Online]. Available: http://stats.nba.com. [Accessed: 04- Oct- 2016].
3. "sklearn.linear_model.LinearRegression — scikit-learn 0.18 documentation", Scikit-learn.org, 2016. [Online]. Available: http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html. [Accessed: 04- Oct- 2016].