MA5851 Assessment 3 Part 1:

Overview

James Watts

James Cook University

April 2021

This project set out to use Natural Language Processing (NLP) to investigate the single issue of how one may incorporate human discussions about NBA player performances as a consumable data feed that makes up part of a pricing model for NBA betting markets.

The intended production model use a web scraper to retrieve lists of active NBA players and their headline advanced season statistics then retrieve discussions about these players from online forums. The NLP part of the model will perform two primary tasks, firstly a sentiment analysis to highlight players that fans have strong feelings either positive or negative about recent performance on the court. The second is a classification task where texts are grouped into five topics, the intention is to break out talks about injury, trades ect. Future development of the model out of the proto stage would connect into pricing model to weight selection pricing predictions and build warning systems when topics like injuries or trades are beginning to be discussed.

The impetus for considering this topic for this project is to explore ways to exploit inefficiencies in betting markets and capture the collective wisdom of the crowds that is not easily born out in recent on court statistics. Betting markets and corporate bookmakers odds are largely driven by statistical models to price odds for games and player props. These models use the rich volume of historical statistics available such as team scores, shooting percentage, assists, player time on court to form the basis of probabilistic predictions. However, odds offered by bookmakers are not fixed to a knowable statistical likelihood like that found at a casino game table, ie the price of red at roulette does not change when a player places a large bet on a selection. But instead sports betting odds have elements of stock market prices whereby prices move relative to collective sentiment and weight of money of the participants involved.

It is this relationship between sentiment and price that may create a value opportunity irrespective of if fans sentiment accurately predicts future player performance. If fans positive sentiment that is not part of a recent statistics trend is positively correlated with actual performance then this information can be used to bet into early markets on such players before sentiment weight of money moves the market. Conversely if it turns out that sentiment is in fact just hype that does not get born out in statistically meaningful trends in performance this information can also be take advantage of as it would be prudent to allow the market to be influenced by the hype before betting against the sentiment.

The second part of the NLP part of the project is to classify comments into groups, discussions about players being injured, traded or coaching decisions would like to be isolated out of the discussions for separate analysis. This kind of news can have significant impact on a team's performance and is hard for current models to incorporate into pricing as the models are built around historical (recent) on court performance.