Machine Learning Engineer Nanodegree

Capstone Proposal

Evan Farnsworth

January 17, 2018

Do NHL Referees bet on their own games?

Domain Background

The NHL is a sport that is viewed by millions and whose games are attended by thousands to enjoy the competition of the most elite hockey players in the world. The NHL, along with all other professional sports, also have a multi-billion dollar sub-culture of gambling. With billions of dollars being gambled on professional sports, it would be a reasonable, moral belief that these wagers are independent of the games outcome. There have been movies or rumors where some form of organized crime syndicate attempts to “fix” or “rig” the outcome of an event with the intention of profiting off the “sure thing”.

Anecdotally speaking, some sports fans will claim the game is “rigged” when their preferred team receives a bad call or infraction. Is there any validity to this claim? There was a controversial case within the NBA in 2007 where the referee, Tim Donaghy, was found guilty for a betting scandal. If the referee is compromised for a professional sporting event, then there is potential for a game to be swayed in a given teams favor. We can look at it from a less nefarious perspective as well, what if the referee has a bias against a specific team or player (without a financial implication)? We would expect referees to be unbiased and fair for every single game they officiate. However, they are human and capable of error, emotional attachments, and varying perspectives.

Personal Motivation

I chose this subject as I am a sports fan and enjoy office fantasy drafts or fantasy drafts with friends. Thus gaining an edge within my fantasy draft groups would potentially increase my bragging rights as I typically know less about the sport compared to the other competitors. After learning about the case with Tim Donaghy, and the events that happened with FIFA, I was curious to if there is a way to correlate data to suggest that there could be collusion or other ill intentions towards the outcome of sporting events. I chose the NHL rather than the NBA or MLB because I am Canadian.

Problem Statement

The first problem is to see if we can find bias or uneven data in terms of how each NHL referee officiates for a given team and even a given player. This will be a supervised learning problem as I should be able to use Linear Regression or Naive Bayes to predict a bias for upcoming games. Clustering will also be a useful tool in analyzing the data.

- A referee will be classified as bias in the event that there is a large skew of data in a specific direction as compared to other referees, for the officiating of the referees.
    - The referee data to be analyzed will be: *Penalties and Stoppages*.

- This can be measured by the success of adding the bias to our model as compared to a pure linear regression for predicting the Referees stats.

The second step, in the event we do find indication of uneven data, will be to dive further into the given referees operation. If we assume the hypothetical role of an investigator, it would be worthwhile to compare Vegas betting odds with the way the Referee makes calls on penalties and stoppages. If a referee is betting on a game that they officiate, we will assume that Referee has 1 of 2 strategies.

1. Bet on the team/metric with the more profitable (worse) odds in order to maximize profit. This would be riskier for exposure, and should be easier to see in the data.
2. Bet on the team/metric with the safer (better) odds in order to make money but have less chance of getting caught. This would be more difficult to detect and will require more analysis of a team or players current trending performance.

Datasets and Inputs

The dataset will be the JSON data obtained from the live.nhl.com/GameData API for all Play-by-Play data for all games from 2011-2012 season to the most current. The historic Vegas odds will be downloaded as .xlsx for each season and then converted to .csv. This data is free to access.

Input Fields:

- Home Team - the team playing in its home city.
- Away Team - the opposing team playing.
- Referees - the group of linesmen and Referees officiating the game.

Solution Statement

The proposed solution is to use Supervised Learning Algorithms and potential some Unsupervised learning techniques to attempt to correlate Referee bias to the available data on the Referees actions in a given game. This bias will be determined by comparing:

- The average of all referees on the given 2 teams to the officiating referees, along with the officiating referee's data as compared to the average referees data and if that varies significantly for specific teams. If we can find that the officiating referee makes drastically more/fewer calls against a given team compared to all referees, it would suggest a personal bias.
- The total average data of the officiating Referee for all teams compared to the 2 competing teams. If we compare the Referee to all other teams they have officiated, and they seem to make more/less calls against a specific few, that would suggest a bias.
- For the "Are Referees gambling?" research we will compare the game results as per the Vegas betting, with the previous referee bias data. If there is a strong correlation to the number of calls a referee makes and the odds from Las Vegas based on strategy (1), we may be able to assert that these Referees deserve an investigation.
- For attempting to predict betting strategy (2), if the referee data seems to spike on a given team on a regular basis (but not on other occasions), this could warrant an investigation. The deeper analytics of comparing team trends and data to the odds and referee is considered too complex and will be outside the scope of this project.

Benchmark Model

The benchmark model will be a pure Linear Regression for the Referees total officiated games (average) without consideration to specific teams.

Evaluation Metrics

For the general bias, the evaluation will be how well our Bias model fairs against the average data. If we can predict the outcomes of referee calls from a biased subset of teams with greater accuracy than the pure average for the specific Referee, then the scale of that improved accuracy will act as the evaluation metric. For the gambling analysis, we are more so looking for strong correlation than a predicted value. Thus, using Unsupervised learning techniques like PCA or clustering will be used to try to correlate the Vegas odds with the previous bias spikes.

Project Design

The first step will be to create a program that will pull and parse the JSON NHL game data, and compile it into a consumable format for testing. I plan to just parse and manipulate the JSON such that it is organized for each referee in either CSV or JSON format.

The second step will be to setup a pipeline to sample and visualize the data. I have only speculated which algorithms will be most suitable, and thus will need to review the data to assess if another algorithm or learner is required (maybe neural networks could be used).

Next I will create the benchmark and simple models such that I can start comparing data and predicting Referee performance. After which I will create a pipeline such that i can use differing parameters or learners in a quick and optimal manor. At this point i will continually run tests and tweak my models such that I can try to predict a Bias.

Assuming the previous steps show a mild indication of some referee bias, I will compile a csv of all historical Vegas betting odds for all games and develop a pipeline such that I can compare Referee data and (potential) bias with the gambling odds to determine if that Referee is worthy of being under investigation.

References

1. http://www.espn.com/nhl/attendance/_/year/2017
2. https://www.statista.com/topics/960/national-hockey-league/
3. https://www.inc.com/slate/jordan-weissmann-is-illegal-sports-betting-a-400-billion-industry.html
4. https://www.theguardian.com/sport/2015/may/22/ex-nba-ref-tim-donaghy-organized-will-always-have-a-hand-in-sports
5. NHL data : https://live.nhl.com
6. Intro to fetching NHL data: https://stephanefrechette.com/fetching-nhl-play-by-play-v2/#.Wl-4DainGUk
7. Vegas historical odds data: http://www.sportsbookreviewsonline.com/scoresoddsarchives/nhl/nhloddsarchives.htm