Michael Dettmar

Capstone Proposal

October 19th, 2019

Fantasy Football Performance: Adding a Layer with Sentiment Analysis

Overview and Problem Statement

Fantasy Sports is a nearly 14 billion dollar industry[1] indulged in by 21% of the American populace.[2] Whether for money or bragging rights, all of those 60 million players are interested in finding an edge in fantasy.

In daily and season-long fantasy formats, participants draft combinations of players with the hope that their combination will perform better than others'. The core task, then, is predicting the performance of individual players. The widespread availability of sports data for free online enables us to apply machine learning to this problem. Indeed, many have built models to attempt to beat traditional predictions, which are usually based on expert opinion.

Among the most sophisticated modeling approaches to fantasy performance is outlined in a 2018 article by Christopher Zita.[3] He applied different types of neural networks to the task of predicting the fantasy performance of two very different players – consistent quarterback Tom Brady and inconsistent running back Todd Gurley – during the 2018 season. Using multivariable RNNs customized to each player, Zita was able to predict Brady and Gurley's performance with mean absolute deviations of 5.3 and 5.68, respectively – better than all other models built in studies we reviewed and very close to or better than the best expert projections.

---

[1] Ausick, Paul. "How Much the $7 Billion Fantasy Football Business Costs Other Employers." *247wallst.Com*, 15 Aug. 2019, 247wallst.com/economy/2019/08/15/how-much-the-7-billion-fantasy-football-business-costs-other-employers/.

[2] "Industry Demographics • Fantasy Sports & Gaming Association." *Fantasy Sports & Gaming Association*, 2017, thefsga.org/industry-demographics/.

[3] Zita, Christopher. "Predicting Player Performance, Using Neural Networks." *Medium*, Medium, 29 July 2019, medium.com/@christopherzita/predicting-player-performance-using-neural-networks-f6142784b681.

Of course, until a model achieves perfection, we can still improve it. Here, we will take on the problem of improving upon Zita's model by modeling the 2013 performance of Brady and running back LeSean McCoy (we use 2013, and McCoy in place of Gurley, for reason of data availability). We will roughly replicate Zita's approach by building a multivariable RNN for each player, and we'll improve performance though the inclusion of an innovative feature class in the form of Twitter sentiment features.

Domain Background

Zita's article is just that – an article rather than a study. He outlines the broad contours of his methodology but reveals few details. We know only that his RNNs considers fantasy points scored by the player in the last game, defensive statistics of the week's opposition, and whether or not the game was played at home or on the road.

It seems safe to assume any other features were based on player/teammate/opposition statistics and game context, as these are standard in fantasy modeling. Lutz's seminal 2015 paper considers passing and rushing statistics as well as quality of the opposing defense (based on turnovers forced, points allowed, and yards allowed) in an artificial neural network that generalizes to all players of a given position.[4] The network achieved a mean absolute deviation of 6.25 for quarterbacks in the 2015 season – good, but not as good as Zita's customized RNN approach. It is likely that Zita used some of these proposed statistics, and we will use them, as well.

We can also imagine a number of other feature classes that might considerably improve performance. One of particular interest is social media sentiment analysis, which has the potential to harness the 'wisdom of crowds.'

The only expert prediction able to beat Zita's RNN for Tom Brady's performance was FantasyPros.com, who 'project their fantasy points by using advice from over 100+ experts.' If these experts are selected with some degree of randomness so as to avoid groupthink, 'wisdom of crowds' theory tells us that they will be able to predict player

[4] Lutz, Roman. "Fantasy Football Prediction." *ArXiv.org*, 26 May 2015, arxiv.org/abs/1505.06918.

performance with a high degree of accuracy.[5] This same theory tells us that a highly diverse crowd non-experts and experts alike is even better than a smaller group of experts for such a problem. There is reason to believe, then, that we will be able to improve model performance by including features that capture mass opinion of the players at the time. Thankfully, in the modern age, we have a suite of tools at our disposal that can collect such mass opinion: social media.

Yanchen Hong and Steven Skiena's 2010 study[6] showed that the ratio of positive to negative sentiment associated with a given NFL team as expressed in the news in the week before a game correlated with game outcomes, and that a betting model based on a simple sentiment score and home field advantage could achieve a win rate of 60% between 2006 and 2008 – well above the 53% needed to profit (note that a 60% win rate does not mean 60% accuracy, because the authors were examining the point spread). Sinha, et. al.[7] built upon this study in 2013, finding that Twitter rate (volume) features could greatly improve the performance of a model trained on traditional statistical features (though they found fairly unsophisticated Twitter unigram features, focused on the content of Tweets, less useful). Their best model, trained on the 2010 and 2011 seasons and incorporating both Twitter rate and classic statistical features, was able to choose the winner under two different betting schemes over 58% of the time in the 2012 season. Finally, Schumaker, et. al's 2017 paper[8] introduced a metric based on the change in the ratio of positive to negative sentiment Tweets associated with a team in in the three days before kickoff that was more strongly correlated with actual game outcomes than raw sentiment features alone. This 'swing' metric, as they call it, was able to correctly call the same number of games in the 2015 season as raw odds alone, but was much more profitable because it was more successful on

---

[5] Surowiecki, James. 2005. *The Wisdom of Crowds*. New York: Anchor Books.

[6] Hong, Yancheng & Skiena, Steven. 2010). "The Wisdom of Bookies? Sentiment Analysis Versus. the NFL Point Spread." ICWSM 2010 - Proceedings of the 4th International AAAI Conference on Weblogs and Social Media.

[7] Sinha, et al. "Predicting the NFL Using Twitter." *ArXiv.org*, 25 Oct. 2013, arxiv.org/abs/1310.6998.

[8] Schumaker, R.P., Brown, L.L., Labedz, C.S., & Jarmoszko, A.T. 2017. Using Financial Analysis Techniques on Twitter Sentiment to Improve NFL Predictions.

underdogs than the odds-based approach (incidentally, this also suggests some orthogonality in the two feature classes).

All of this research shows the strength of Twitter sentiment in predicting NFL game outcomes. As far as we can tell, however, it has not yet been applied to individual player performance. Nonetheless, given the strong correlation between a player's performance and his team's chance of success, we suspect that the usefulness of these features will transfer.

Data

Though both sports and Twitter data is readily available online, much of it requires a license. We managed nonetheless to compile a dataset from which we can derive all features and targets required during the 2010 – 2012 NFL seasons using a variety of free sources. The five principal sources are detailed below.

- **brady.csv and mccoy.csv – fantasy stats for Tom Brady and LeSean McCoy for the 2010 – 2012 seasons**.
  - Two .csv files with the following notable columns:
    - `G#` (int). The game number for the given season.
    - `Date` (datetime). The date on which the game was played.
    - `home/away` (Boolean). Whether or not the game was played at home or on the road.
    - `Opp` (str). The abbreviated name of the opposing team.
    - `FanPt` (float). The number of fantasy points scored by the player in the game according to the standard fantasy football scoring scheme.
  - This data contains the target we are modeling for: fantasy points. It is in time series format, which will be good for our RNN application. It also has a few interesting features, notably `home/away` and opposing team.
  - To get the data, we downloaded a separate .csv for each player/season combination was from Pro-Football-Reference.com.[9] We then combined

9 "Tom Brady NFL Fantasy Football Statistics". Pro-Football-Reference.com. Retrieved October 19, 2019; "LeSean McCoy NFL Fantasy Football Statistics". Pro-Football-Reference.com. Retrieved October 19, 2019.

all the data into a single .csv and removed fields unnecessary fields. Notably, we dropped the players' actual football statistics for this project, because fantasy score is directly determined by those statistics. We can therefore use it as a stand-in.

- **game_metadata.csv – data on the location, weather conditions, betting lines, and results of all NFL games played since 1966.**
  - Notable columns:
    - `schedule_date` (datetime). The date on which the game was played. Good for joins.
    - `team_home` (string). The home team's name.
    - `team_away` (string). The away team's name.
    - `spread_favorite` (float). The point spread for the team favored to win.
    - `over_under_line` (int). The over/under line for number of points scored.
    - `weather_temperature` (int). The temperature in Fahrenheit.
    - `weather_wind_mph` (int). Windspeeds in miles per hour.
    - `weather_humidity` (int). The humidity on a scale of 1 to 100.
    - `weather_detail` (str). A freeform field to add additional weather details. In practice, it is either blank or used to indicate that the game was played under a dome.
  - We will use the betting and weather features in our model, as the Schumaker paper suggested the former is orthogonal to sentiment features while the latter likely influences player performance.
  - The data was obtained from Kaggle.[10]
- **defensive_stats.csv – defensive statistics for each team during each game of the 2010 – 2012 seasons.**
  - Notable columns:
    - `team` (str). The team whose defensive stats are listed.

[10] spreadspoke (2019, August). NFL Scores and Betting Data. Retrieved October 19, 2019.

        - `date` (datetime). The date of the game played.
        - `pass_yards_allowed` (int). The number of passing yards allowed by the defense.
        - `interceptions` (int). The number of interceptions made.
        - `fumbles_forced` (int). The number of fumbles forced.
        - `rush_yards_allowed` (int). The number of rushing yards allowed.
        - `sacks` (int). The number of sacks made.
        - `points_allowed` (int). The number of points allowed.
    - Because both of the players we are modeling are offensive players, the quality of the defense they are up against in each game will be important. We will make features based on three day moving averages of each statistic in the dataset.
    - This data was also procured from Pro-Football-Reference.com, this time using the sportsreference Python package.
- **tweets.nfl.weekly.csv – Tweets associated with each team during the 2010 – 2012 NFL seasons**.
    - Here we have three .csvs: one for each season containing tweets that mention an NFL team in the hashtag. The tweets occur between twelve hours after the start of a previous game and one hour before the start of a following game. Other data was available looking at the tweets occurring just before and after each game, but for simplicity's sake we decided to narrow our analysis to these three .csvs.
    - Notable columns:
        - tweet_id (int). the ID of the tweet.
        - Tweet_UTCtime (datetime). The time the tweet was sent out, in UTC.
        - team (str). The team associated with the tweet.
    - Per Twitter's terms of service, the dataset does not contain the content of the tweets – only the IDs, which can be used to retrieve the content. I will use the content in my project but will not include it here to comply with those terms of service. I will, however, include the script I used to pull the

content. Note that we can only pull the content of tweets that still exist, which is only ~70% of tweets in this dataset. This may bias the data, as negative tweets are more likely to be deleted than positive tweets. Nonetheless, if we assume this phenomenon is distributed evenly across teams, there should not be a problem. We will make this assumption.

- As stated above, we hope to perform sentiment analysis on each tweet. We will then construct eight features for each game: number of tweets in a week for the player's team and opponent, change in number of tweets three days before game to one day before game for the player's team and opponent, positive-to-negative sentiment ratio for the player's team and opponent, and change in positive-to-negative sentiment ratio three days before game to one day before game for the player's team and opponent. Note that these tweets are on the team level, not the player level. We will assume that sentiment for the team as a whole translates to sentiment for the players we are modeling. This does not seem like a crazy assumption, because both Brady and McCoy were very important for their teams' overall prospects in the seasons examined. It would nonetheless be interesting in a future study to look at tweets directed specifically at the players.
- This is the same data used in the Sinha, et. al study cited above.[11] The availability of this data, graciously provided by the authors, is the reason that we have chosen to focus on the 2010 – 2012 seasons. We could have attempted to gather a new Twitter dataset, but this is nearly impossible to do in a short period of time without an enterprise license.

Proposed Solution

To predict the fantasy performance of Brady and McCoy in the 2012 season, we will build a multivariable RNN with number of fantasy points as the target. We will train

---

[11] **Predicting the NFL using Twitter.** Shiladitya Sinha, Chris Dyer, Kevin Gimpel, and Noah A. Smith. In *Proceedings of the ECML/PKDD 2013 Workshop on Machine Learning and Data Mining for Sports Analytics*, Prague, Czech Republic, September 2013.

on all data from the 2010 and 2011 seasons and predict the outcome in fantasy points of all games in the 2012 season. Specifically, we will actually retrain the model for each game in the 2012 season to predict the following game, so that we are incorporating all available information. We'll use the following broad feature classes:

- Player fantasy points
- Opposing defense quality
- Twitter sentiment and volume for the player's team and opponent
- Home/away
- Weather conditions

Baseline Model

Unfortunately, there is no perfect apples-to-apples comparison for our model. We wanted to replicate the models built by Zita with some new features, but the data available to us was constrained to the 2010 – 2012 seasons.

Nonetheless, we believe that we can still use Zita's multivariable RNNs as baseline models. He modeled the Tom Brady's 2018 season by training on seven years of past data and achieved a mean absolute error of 5.3. While we will be modeling Brady's 2012 season with only two seasons of previous data, we believe the task and methodology is similar enough that we can benchmark our model's performance to Zita's. Similarly, we selected LeSean McCoy as our second player to model because he has a similar profile to Zita's second subject, Todd Gurley. Both were early-career running backs in their respective seasons who performed inconsistently game-to-game. Zita's model of Gurley achieved a MAE of 5.68, which we will use to benchmark our McCoy model.

Evaluation Metrics

To preserve our ability to assess our results against Zita's, we will use the same evaluation metric as he did: mean absolute error between actual and predicted fantasy points for an entire season of football. This is also the metric we will optimize for in our loss functions.

Project Design

Because of the ease of implementation of multivariable RNNs, we will use Keras to train our models. The project will proceed in four phases:

1. Feature engineering. Our data is still messy and needs to be cleaned before we can model off of it. Phase one of the project will involve wrangling the data into multivariable time series format and engineering our desired features. In particular, we will need to look at a moving average of defensive performance for each of our players' opponents and we will need to engineer the Twitter features outlined in the "Data" section. Time allowing, we would also love to train a quick XGBoost model on the 2010 and '11 seasons for each player examine the Shapley values of each feature to get a general idea of feature importance and to further benchmark our RNN.
2. First model. In phase two, we will normalize and tokenize our data so it is compatible with an RNN and train our first multivariable RNN on the 2010 and 2011 seasons. We will then deploy this model using SageMaker and predict the outcome of the first games of the 2012 season.
3. Model updates. In the final phase, we will iteratively update our model for each game of the 2012 season, retraining it to incorporate the most recent game before predicting the next one. For example, the first model will consider only games from the 2010 and 2011 seasons to predict the first game of the 2012 season. We will then retrain this model, adding the first game of the 2012 season as an input to predict the second game of the 2012 season. We will repeat this until we have a prediction for each game.
4. Data compilation and analysis. Having made all of our predictions, we will assess our model's performance versus the baseline and reflect on things that went well, things that did not go well, and future areas of research.