Kevin Chan, William Parsons, Conor McGeehan
EECS 349: Machine Learning
Professor Doug Downey

Final Project Report

Title:

America's Pastime: Gambling

Task:

On an episode of Last Week Tonight with John Oliver, Oliver states that the two largest daily fantasy sports companies, DraftKings & FanDuel, are multi-*billion* dollar companies. Forbes reports that the fantasy football industry is a 70 billion dollar industry. So, it should be no surprise that sports gambling and sports betting are a huge deal, whether it be just a small group of friends or a online pool through an app. The NBA postseason is coming to an end, and the NFL season hasn't started yet, but the MLB season is happening now. Therefore, we wanted to create a machine learning algorithm that could make better predictions on a MLB team's chances at making the playoffs. Since one could easily tell if a team would make the playoffs simply based on a team's win/loss record at the end of the season, we set out to predict a team's journey to the playoffs based on their statistics at the end of month of the regular season: March/April, May, June, July, August, September/October (March & April and September & October are grouped together due to the small number of games in March and October that arise from scheduling). Taking a team's batting and pitching stats at the end of each month, we trained and tested an algorithm to predict if a team would make it to the postseason from these various points. We hoped this approach would enable users to make more educated guesses in their own sports gambling ventures by providing numerical metrics from which they could better predict the current 2017 MLB season, which is still under way.

Data:

We sampled five statistics for each MLB at the end of each month for the last five years since before then the playoff structure was different which would result in skewed data. After the World Series in 2011, the MLB announced that they would be adding a second wildcard team to the postseason to play in wildcard "play-in" games both for the National League and the American League.

The five statistics we used were: BABIP, tOPS+, WHIP, SO/9, and RD, which we picked after much discussion and consulting with our baseball-loving friends. BABIP, or batting average on balls in play, is important in determining whether a team's batting average will increase or decrease as the season goes - a high BABIP indicates a decrease in production going forwards as hitters tend to reach base on 30% of balls put in play, and significantly different rates are hard to maintain. tOPS+ (OPS split relative to player's career production) takes a player's or

team's OPS (on-base percentage + slugging percentage) for a given range of time and compares it to his or their career OPS statistic. This is considered a good determinant as to whether a team's hitting will regress (normalize) or not as the season continues. WHIP calculates a pitching staff's (wins + hits) / innings pitched. This is an indicator of how many runners a staff, or starting rotation, lets on base per inning, and is thought to be one of the most indicative predictors of strong performance for a pitcher. SO/9 (strikeouts per 9 innings) takes the number of strikeouts from a staff and normalizes for 9 innings of play, or a full game. Strikeouts, coupled with walks and hits allowed, encompasses the majority of important statistics for a pitcher or staff. RD, or run differential, is the number of runs scored by a team minus the number of runs allowed. Generally a large RD is indicative of a team batting well and fielding well, as a positive number means more runs are being scored than are being allowed. A team that scores more runs and allows fewer is more likely to win games.

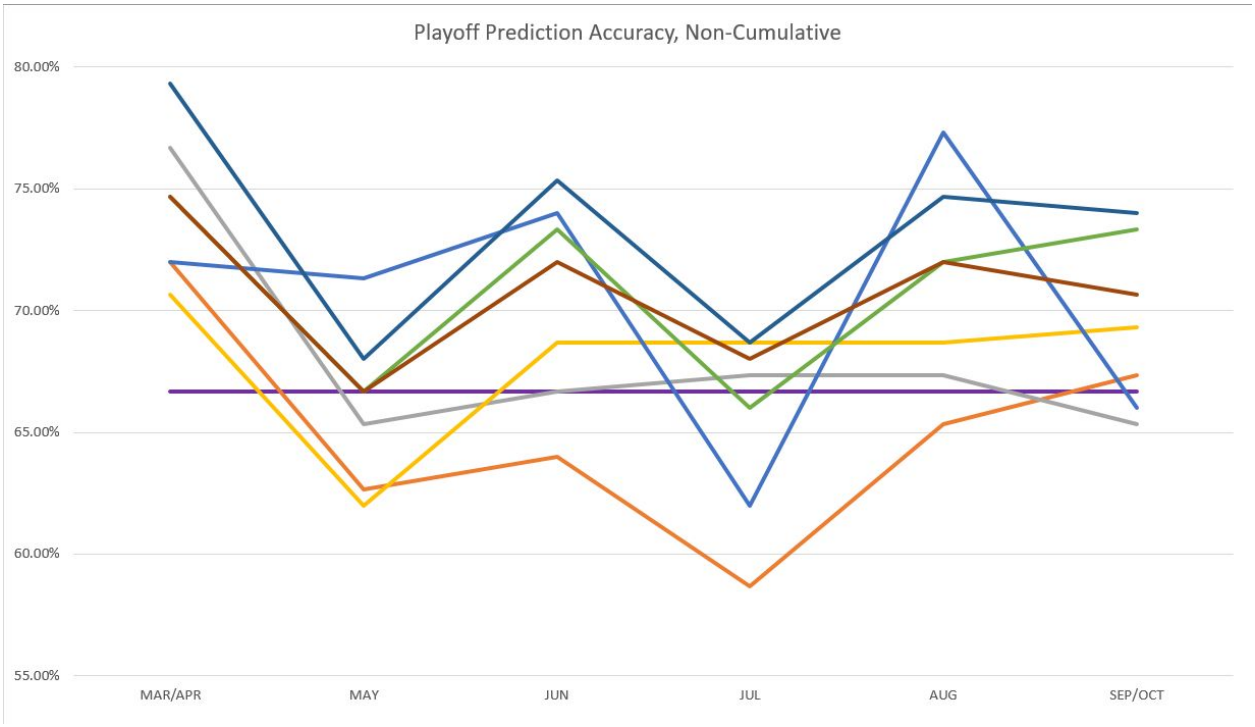
All of our statistics were taken from baseball-reference.com, and we kept two different data sets. The first set (6 subsets total, one for each month), contains numbers corresponding directly to the month in question. The attribute value of a team in the May data set is strictly for the month of May. The second set (6 subsets total, one for each month), contains cumulative stats up to the designated point in time. For example, the numbers in the May dataset are averaged over March/April and May, and the numbers in the June dataset are averaged over March/April, May, and June, etc. Run differential was simply summed instead of averaged.

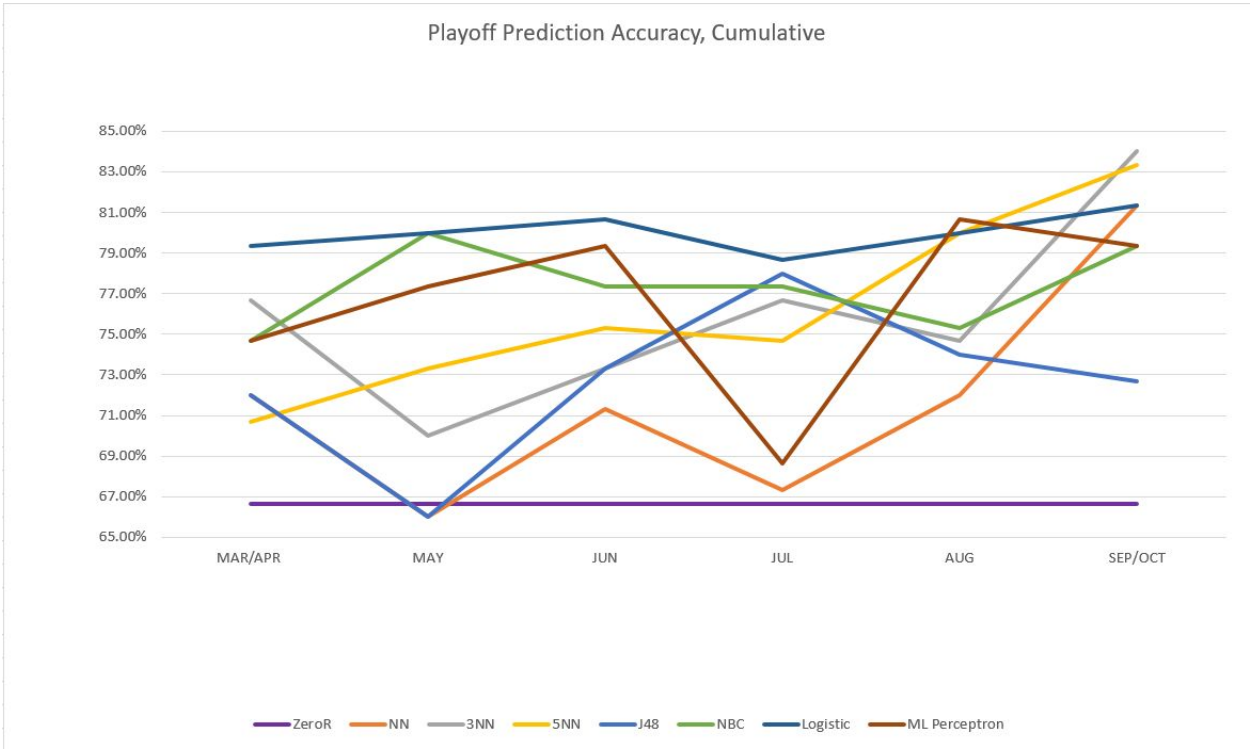
Methods:

We used the data above to accomplish two main tasks. First, we ran our 12 total data sets (6 months non-cumulative, 6 months cumulative) through WEKA using 10-fold cross-validation to check our accuracy. We tried 8 different algorithms on our data set through WEKA, and our results are displayed below in the results section. The 8 algorithms we used were: ZeroR, 1NN, 3NN, 5NN, J48 DT, Naïve Bayes Classifier, Logistic Regression, and Multilayer Perceptron. After identifying the best option, which appeared to be Logistic Regression, we moved onto our second part, which was to predict the current playoff outcomes of the current 2017 season. We passed the 2017 stats of each of the 30 teams to the Logistic algorithm in using cumulative stats up to the end of May, and used the output to predict if a team would make it to the postseason. The results of our prediction are listed below.

Results:

		Algorithm:	*All accuracies reported from Weka are from 10-fold cross validation						
	Month	ZeroR	NN	3NN	5NN	J48	NBC	Logistic	ML Perceptron
Regular	MAR/APR	66.66%	72.00%	76.67%	70.67%	72.00%	74.67%	79.33%	74.67%
Regular	MAY	66.66%	62.67%	65.33%	62.00%	71.33%	66.67%	68.00%	66.67%
Regular	JUN	66.66%	64.00%	66.67%	68.67%	74.00%	73.33%	75.33%	72.00%
Regular	JUL	66.66%	58.67%	67.33%	68.67%	62.00%	66.00%	68.67%	68.00%
Regular	AUG	66.66%	65.33%	67.33%	68.67%	77.33%	72.00%	74.67%	72.00%
Regular	SEP/OCT	66.66%	67.33%	65.33%	69.33%	66.00%	73.33%	74.00%	70.67%
Cumulative	MAR/APR	66.66%	72.00%	76.67%	70.67%	72.00%	74.67%	79.33%	74.67%
Cumulative	MAY	66.66%	66.00%	70.00%	73.33%	66.00%	80.00%	80.00%	77.33%
Cumulative	JUN	66.66%	71.33%	73.33%	75.33%	73.33%	77.33%	80.67%	79.33%
Cumulative	JUL	66.66%	67.33%	76.67%	74.67%	78.00%	77.33%	78.67%	68.67%
Cumulative	AUG	66.66%	72.00%	74.67%	80.00%	74.00%	75.33%	80.00%	80.67%
Cumulative	SEP/OCT	66.66%	81.33%	84.00%	83.33%	72.67%	79.33%	81.33%	79.33%





Using a Logistic Regression to predict 2017 outcomes, we predict that the following teams are going to make the playoffs at the end of this season: Astros, Brewers, Diamondbacks, Dodgers, Indians, Nationals, Rays, Red Sox, Rockies, White Sox, and Yankees.

Analysis:

Above we have graphed the accuracy in prediction method for all classifiers we tested for each month. If you note the change in vertical axis scaling between the two graphs, it is clear the cumulative statistics produced a much tighter classification than non-cumulative, with all classifiers accurately predicting over 65% of teams during cross-validation. Additionally, in both graphs, logistic regression appears to outperform nearly all other classifiers for each month. 5NN also appears to be a good classifier, as it drastically improves as the season goes on and more indicative statistics are created for the team's yearly performance, which aligns with our expectations of what was supposed to happen. The other classifiers show this trend somewhat, but 5NN is the most consistent. Even so, we decided Logistic Regression was the best classifier as it outperformed the other classifiers in almost all months.

Future Work:

In the future, we would want to explore more features and potentially viable attributes to acquire stronger predictions. Given the tedious task of entering in data by hand into the spreadsheet, we chose to stick to 5 descriptive statistics. The time constraint prevented us from using more data, since the data online was formatted differently than how the Weka software used accepts inputs. This forced us to enter thousands of data points by hand, which was very time consuming. If we had sufficient time, we would choose to analyze far more stats to create more accurate combinations of statistics so we could hopefully find surprising correlations. One option would be to pass the data to analyzing software such as STATA, where we could find p values and determine if certain attributes were statistically significant. Using this information, we could numerically determine the most important stats. Unfortunately, STATA costs money on a subscription basis, so we did not have the financial resources to perform this task.

Another important future step would be to predict outcomes of the 30 teams this season as time goes on. At the time of this project, only the months of March, April, and May have elapsed, giving only early stats to make predictions on for those 3 months. As described above, our predictions get more accurate as time progresses, so we would want to report our predictions after more months have passed in the season.

Yet another future alteration to this project would be to set restrictions regarding the number of teams that can make the playoffs from each division and league. Our current classifier does not restrict this in any way, and it is quite surprising the prediction for this year managed to select a feasible playoff picture. In further work, we would divide the data further and make predictions based on the constraints that only certain numbers of teams can come out of certain leagues. Coincidentally, our 2017 predictions match a feasible postseason lineup, but this is merely a coincidence, and was not intentional.

Division of Labor:

Kevin: Kevin helped compile a large portion of the data for this project, both in the training data sets and the dataset containing the 2017 team stats. In addition to writing and editing chunks of the final report, he also was the one to take all the data into Weka and test the different algorithms on it. He compiled a spreadsheet of all the results, which enabled Conor to generates graphs for our website.

Will: During the project Will helped with a large portion of compiling statistics into the excel spreadsheets for all the data. Additionally he helped with some minor website design and aesthetic input. He helped with writing and editing the final report both in the write-up and how it appeared in the website. Finally, he determined how to generate predictions for the current season based on the stats that we were able to compile so far, allowing us to acquire a list of teams our classifier believes will make the playoffs.

Conor: Conor was instrumental in constructing our website. He designed and coded our website so that we could display all the information from our project. He is the biggest baseball fan among us, so he provided key advice to the team about which baseball stats were the best to use. He also helped compile data, and edited our final report. He also generated graphs using our output from Weka.