Vince Laird

SPORTS BETTING LINE MOVEMENT MILESTONE REPORT: FIRST DATASET (MLB, APRIL 2016)

DATA SCIENCE - CAREER TRACK

Introduction

Sports betting lines and odds move and change throughout the day, and it would be useful to be able to predict this line movement. This will help people betting on sports make an informed decision on the optimal time of when to place their wager.

Initial dataset

Two datasets were initially going to be used: one containing all data from 2016, and one containing Spring Training data from 2017. However, scraping the 2016 data became more time-consuming than anticipated, so only April was used. April 2016 MLB baseball data contains betting line movements from 354 games.

Spring Training betting line movement data was collected, but not used. Spring Training games are not bet on heavily; out of over 160,000 JSON documents collected during one week, which contained line movements for every MLB baseball contest every minute, the average number of line movements per contest was less than four, compared to over four times that amount from the data collected from April 2016.

The script that was used to collect Spring Training data will continue to run throughout the regular 2017 season. The resulting JSON documents can be copied locally weekly, and these will be saved for later analysis.

Key data contained in the initial dataset included: time-based betting lines (favorite and underdog), odds, and teams. This data was cleaned to determine opening and closing odds, which were used for comparison purposes.

Two limitations of the dataset were that it did not contain which team was the home team, and the time that the contest actually started. This data can be found elsewhere, but it would have been nice to have it all in one place.

There will always be data that is desired but not necessarily easily found: what were the starting lineups for each team, who was the starting pitcher for each team, were there any last-minute scratches to the lineups due to injury, rest, etc. The tradeoff between time to collect data and value of data collected will have to be evaluated as the project continues.

Data cleaning

Several steps were taken to clean the data. Whitespace was removed so that the columns of data were consistent. Columns were added with timestamps that could be compared and used in calculations. Duplicate rows were removed. Data was sorted and subsets were created as necessary for data exploration purposes.

Data exploration

After the data was cleaned, I wanted to determine when most changes in odds occurred, in the event that data collection for future contests was limited.

As was expected, the majority of changes occurred 15 minutes before betting on the contest was closed (which corresponds to when the contest begins). There were also a few points in the day when line changes seemed to occur a bit more frequently, and this could be during times in the morning / afternoon hours (due to gambling volume), and when MLB lineups are announced. A table showing frequency in odds changes in 15 minute increments leading up to the start of MLB baseball games in April can be seen here:

https://github.com/vincelaird/springboard/blob/master/capstone/Data%20story.ipynb

The second piece of data I was interested in was seeing how often odds for favorites became worse, vs. odds for underdogs becoming worse. It turned out that underdogs became worse bets in 187 out of 338 cases (while the original dataset contained 354 games, 16 were ignored because the odds were the same at open and close). The

probability of this being due to chance is 2.5% - there may be something to underdogs receiving worse odds, when comparing opening and closing numbers.

After discovering this, I was interested to see which teams had the most disparity between when they should be bet. Out of 30 MLB teams, the three that had the largest number of differences between when they should be bet were:

Baltimore, odds were better when they opened 16 out of 22 times.
Colorado, odds were better when they opened 17 out of 23 times.
Texas, odds were better when they closed 17 out of 22 times.

Because this is only one month of data, the probability of seeing three (or more) out of 30 teams with this sort of disparity (difference in absolute value of odds improving at close or getting worse) is not that unusual - about a 14% chance. However, it's still worth looking at these teams in terms of season expectations and how they played the first month of the 2016 season.

I found two articles (referenced at the end of this document) from prior to the 2016 season which talked about expectations for all 30 teams. Looking only at these three, both Baltimore and Colorado played much better in April than what was expected.

Baltimore was expected to be a below-average team, however they started the year by winning their first seven games and they ended April with a 14-9 win-loss record. This may have been a factor in why their odds were so often better when they opened - people may have been betting on them, thinking they were undervalued and that collective expectations were wrong.

Colorado was in a similar situation. Forecasted to be one of the worst teams in baseball (two projected season records for this team had them finishing the year at 70-92 and 68-94 respectively), they played only slightly below-average in April and finished the month with 11 wins and 12 losses. Like Baltimore, if you were going to bet on Colorado, you were better off doing so early as their odds were worse before their contests began.

For Texas, this logic fails. They were projected to be either slightly above average or significantly above average, and they seemed to play to their expectations, finishing the month of April 14-9 (though they started 6-6). If you were going to bet on Texas, their

opening odds were bad - you were better off waiting, which seems to imply that people were betting against this team.

Next steps

Now that the data has been wrangled and cleaned, and some data exploration has occurred, we will use time series analysis to solve the original problem of determining future betting line movements. The techniques learned in section 14 (advanced topics in machine learning) will be utilized to generate forecasts in betting lines for MLB baseball games for the 2017 season.

Conclusion

Overall, I'm satisfied with how the project is going. It would be nice if the 2016 data were a little easier to wrangle, but the data collection for the current year is going well and I've found some interesting observational things about the data that I do have. I'm looking forward to diving into the 2017 dataset and seeing if any of the (potential) patterns found from last year hold for this year as well.

References

Two articles on 2016 MLB season projections for every team:

http://bleacherreport.com/articles/2615555-mlb-predictions-2016-projecting-the-final-standings

http://www.espn.com/blog/sweetspot/post/_/id/69345/read-em-and-weep-your-final-2016-standings-and-world-series-winner