Soccer Betting Analysis - How to use betting agencies odds to predict match results?

Chen Trilnik

Posted on Apr 29, 2017

View Posts by Categories

Introduction

As a soccer fan with 3 years of work experience as a live soccer match analyst, I have thousands of soccer game hours in my repertoire. I follow European soccer on a weekly basis and know most of the teams and players in the major leagues of Europe.

Even with all my knowledge and experience, I find it hard to predict soccer match results. Like in any other sport, the best team doesn't always win. There are many parameters that affect the outcomes of any given game. The skills of the players, the tactical formation, and teamwork may be the most important ones. But are meeting all these parameters enough to win all the games played? If so, we should be able to predict match results pretty easily.

The truth is, there are many more factors that affect soccer match results - team motivation and spirit, player injuries, fan support, chemistry between teammates and opponents, reputation, and win history are some of them. The complex interplay between these variables during the fast-paced activity of a game makes every match unique.

Professional betting agencies are making a lot of money from people who want to predict match results. It is safe to say that their betting odds are calculated in a way that maximizes their profits and minimizes their risks.

In this project, I wanted to examine the degree to which betting agencies' odds correlate with actual match results, and to see if there is any way to maximize prediction accuracy. For this, I looked at the betting agencies' reported odds for the basic 3-way bet (home team vs. draw vs. away team) and the level of favored outcomes within each game (high favored, moderate favored, and low favored), and compared

R 243 posts
R SHINY 253 posts
R VISUALIZATION 260 posts
STUDENT WORKS 765 posts
WEB SCRAPING 205 posts

Search For

Our Recent Popular Posts

NYC Data Science Academy Introduces Remote Intensive Bootcamp
by claire.tu
May 10, 2018

How to prepare for a data science interview
by claire.tu
Apr 6, 2018

PandaGo, The Travel Recommendation System
by Andrew Dodd, Wenchang Qian and Roger Ren
Mar 30, 2018

View Posts by Tags

them to actual match outcomes to see if betting agencies' odds had any value. I further categorized the match outcomes by the location of the teams (i.e. hosting vs. visiting), the stage of the season, and the numerical difference between the payouts in order to find patterns that optimize prediction accuracy.

In the end, I found that there are three parameters can help predict the outcomes with up to 80% precision: 1) the agencies' high favored result 2) the location of the team, and 3) the stage of the season. I used R shiny app and ggplot2 to visualize the data. You can find the full results on the app.

Explanation

In most soccer competitions, draws may be the final result of the game, so there are 3 different outcomes to bet on between Team 1 and Team 2:

- First outcome: team 1 wins
- Second outcome: team 2 wins
- Third outcome: team 1 and team 2 draw

The odds are translated into payouts. The result with the minimum odd is the one that is most likely to happen, it has the least risk and therefore offers the lowest payout.

The result with the maximum odd is the one that is the least likely to happen, it has the higher risk and therefore offers highest payout.

For example, let's take the first match in this betting odds chart of the English Premier League and look at the odds for the full time result. In this game, Arsenal is playing against Crystal Palace. For an Arsenal win, any dollar you bet will give you $ 1.29 (a $0.29 profit). For a draw in the match, a dollar will give you $ 4.98

API aws beautiful soup Big Data capstone

Show more

($3.98 profit). And a Crystal Palace win will give a return of $ 8.06 ($7.06 profit) for a dollar bet.

Time	Event	Full Time			Half Time		
		1	X	2	1	X	2
	*ENGLISH PREMIER LEAGUE						
08/16 10:00PM	Arsenal Crystal Palace	1.29	4.98	8.06	1.76	2.60	6.79
08/16 10:00PM	Burnley Chelsea	6.61	4.36	1.39	6.13	2.42	1.91
08/16 10:00PM	Leicester City Everton	2.80	3.15	2.35	3.80	1.97	3.04
08/16 10:00PM	Liverpool Southampton	1.41	4.34	6.24	1.94	2.48	5.52
08/16 10:00PM	Manchester United Swansea City	1.39	4.31	6.73	1.89	2.46	6.08
08/16 10:00PM	Newcastle United Manchester City	4.83	3.79	1.59	5.99	2.27	2.03
08/16 10:00PM	Queens Park Rangers Hull City	2.50	3.23	2.56	3.18	1.96	3.61
08/16 10:00PM	Stoke City Aston Villa	2.03	3.23	3.36	2.66	2.06	4.19
08/16 10:00PM	West Bromwich Albion Sunderland	2.22	3.20	2.97	2.86	2.00	3.99
08/16 10:00PM	West Ham United Tottenham Hotspur	3.07	3.30	2.12	4.02	2.05	2.75

The Data Frame

The data sets were taken from Kaggle, a part of a soccer SQLite data base.

The data sets include data on more than 25,000 matches from 9 different leagues in Europe over 8 seasons (2008/2009 - 2015/2016). The data includes: match results and dates, teams, leagues, and match betting odds from 9 different betting agencies.

The European leagues are: Belgium Jupiler League, England Premier League, France Ligue 1, Germany 1. Bundesliga, Italy Serie A, Netherlands Eredivisie, Portugal Liga ZON Sagresand, Scotland Premier League, and Spain LIGA BBVA.

The betting agencies are: Bet365, Blue Square, Bet&Win, Gamebookers, Interwetten , Ladbrokes, Pinnacle, Sporting Odds, Sportingbet, Stan James, Stanleybet, VC Bet, and William Hill.

It is important to note that there was always consensus between the agencies regarding the probability for each outcome (i.e. they all thought Arsenal had the highest chance to win); the only difference was the magnitude of payout that they

offered. Therefore, I considered the average consensus as a single entity.

For data processing I used RSQLite package for R to convert the different SQL tables to CSV files.

```
install.packages('RSQLite')

library(shinydashboard)
library(shiny)
library(RSQLite)
library(dplyr)
library(ggplot2)

con <- dbConnect(SQLite(), dbname="~/Downloads/database.sqlite")
Match <- tbl_df(dbGetQuery(con,"SELECT * FROM Match"))
League   <- tbl_df(dbGetQuery(con,"SELECT * FROM League"))

```

As part of the data cleaning and preparation, I deleted rows with missing values and ignored data from 2 betting agencies because their betting odds were uploaded to the SQL server as integers rather than exact numeric values. After this process there were 22,434 observations left.

Moreover, I added columns to the data set to include the match winners, the agencies' average minimum, middle, and maximum payout, and agencies' favored result.

For the analysis, I defined the result with the minimum payout as the favored result by the betting agencies. The success rate shown in the charts is calculated as the number of times the favored result was the actual final result of the match, divided by the total matches played.

The favored result level column is a breakdown of the matches to 3 groups using the difference between the payouts (as extrapolations of the odds).
My assumption for this calculated column is that the higher the difference between the payouts, the higher the chance for the minimum payout to be the winning

outcome. Therefore, the groups are categorized in the following way:

A high favored result = max payout - min payout > 2

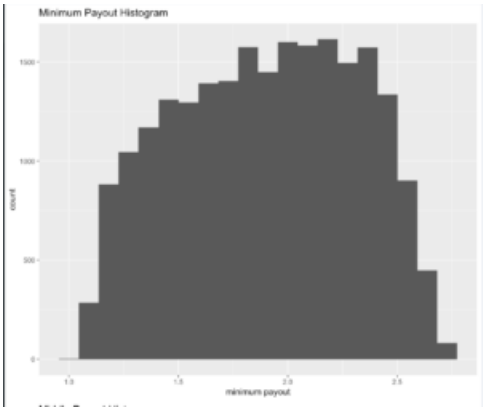
A moderate favored result = 2 >max payout - min payout > 1

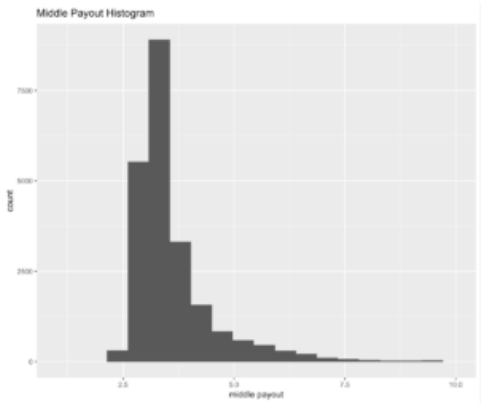
A low favored result = max payout - min payout < 1

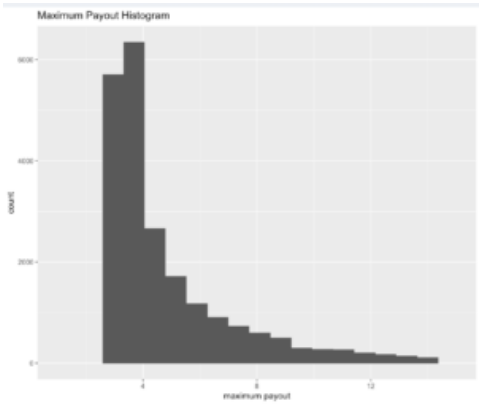
Analysis

The Payout Distributions

The betting payouts have a normal distribution. The maximum and middle payouts are skewed to the right. The minimum payout ranges from a little more than 1 to around 3. The maximum payout ranges from 2.5 to 40. The middle payout distribution looks similar to the maximum payout and range from 1.9 to 10. Below are the histograms of the payouts:

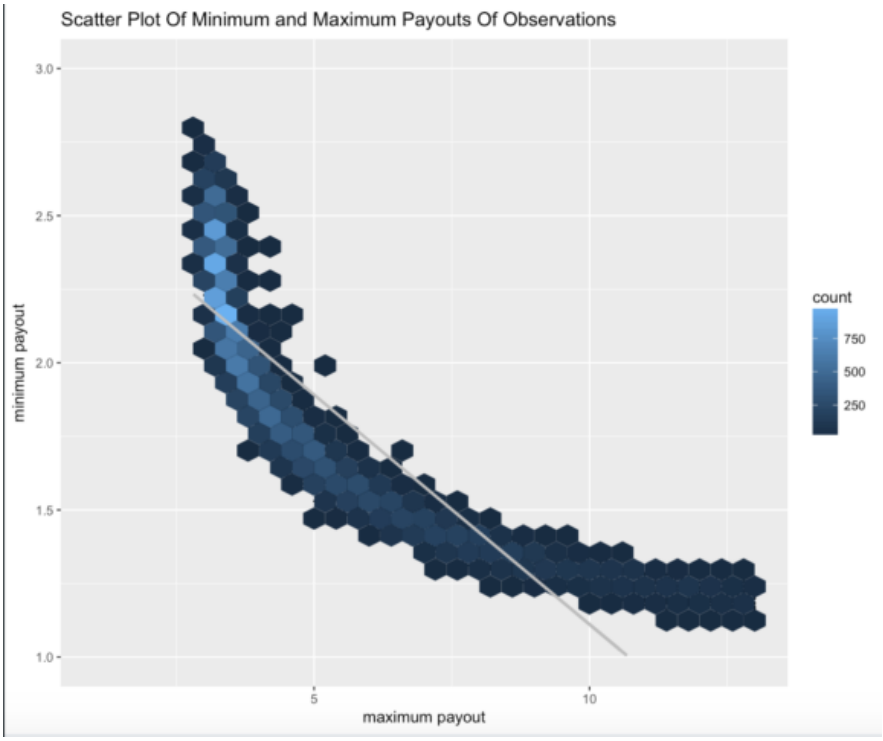






The minimum and maximum payouts are inversely related and the minimum and middle odds are also inversely related. This is explained by the fact that when there is a high favored result (for example, one team has a better track record than the other), its payout will be low and accordingly the other outcomes' payouts will be high. On the other hand, when there is no high favored result (for example, the teams playing have same skill level), the payouts will be quite similar.

This is a scatter plot of the minimum and maximum payouts of the matches observed:

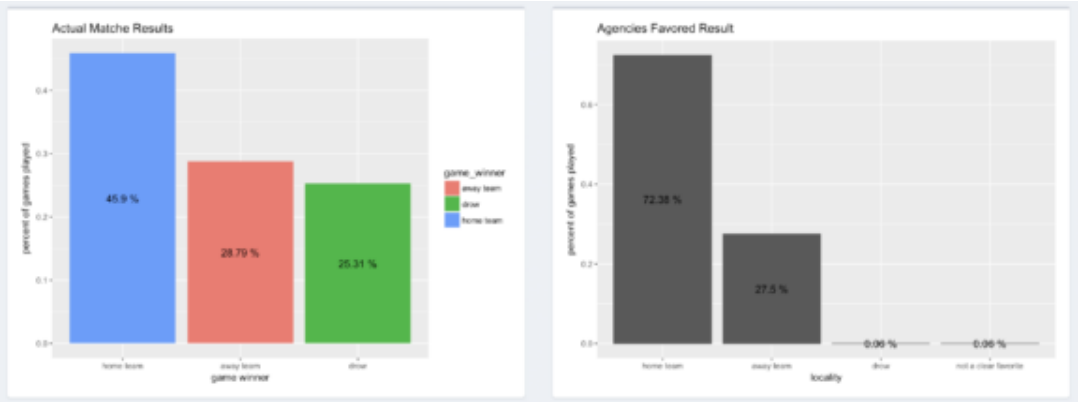


Favored Result Analysis

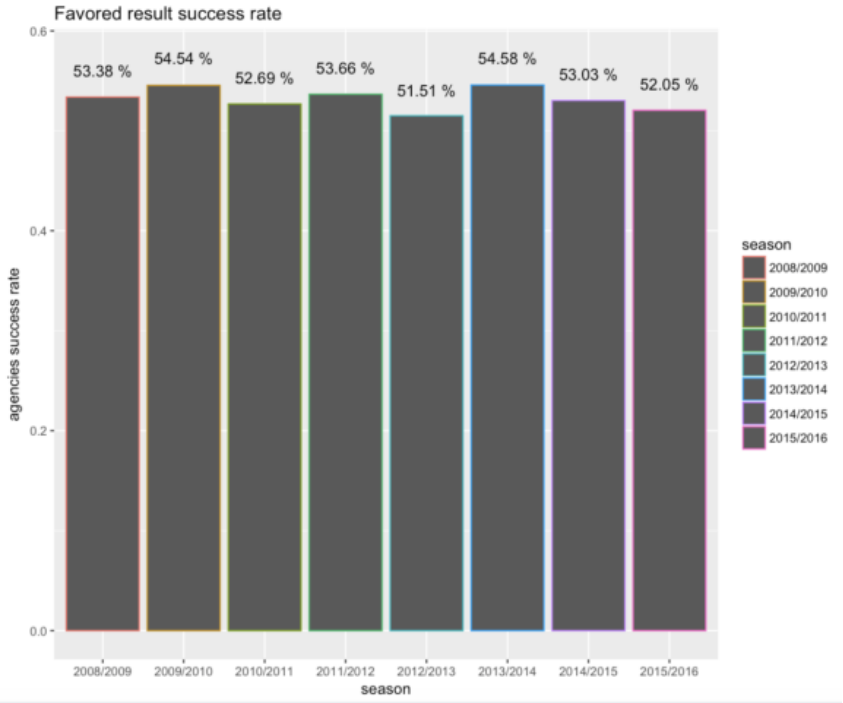
The first chart depicts actual match results. Here, we can see that the home teams wins 46% of the time, the away team wins 29% of the time, and there is a draw 25% of the time.

The next graph shows that the agencies favored the home team 73% of the time and the away team 27% of the time, while they almost never favored a draw (13 out of 22,434 matches). We can see that the agencies favored the home teams in most cases. This emphasizes the importance of location in the competition.

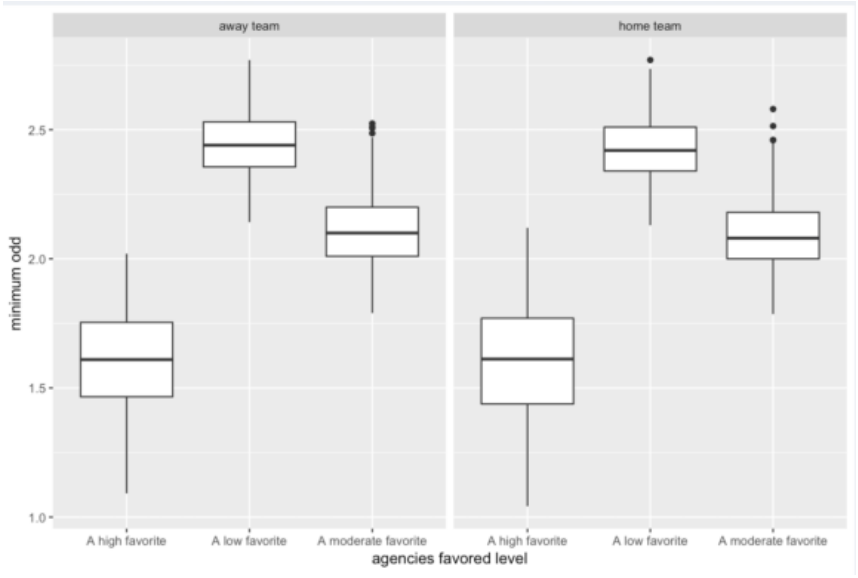
This chart raises an interesting question: why do the agencies never favor a draw result when this outcome occurs in at least 25% of the matches? I did not come across any data explaining how the agencies determine their payouts, however, I believe that agencies prefer to favor one team over the other because it's easier for them to promote the bet among gamblers. It's just more interesting to have a face-off.



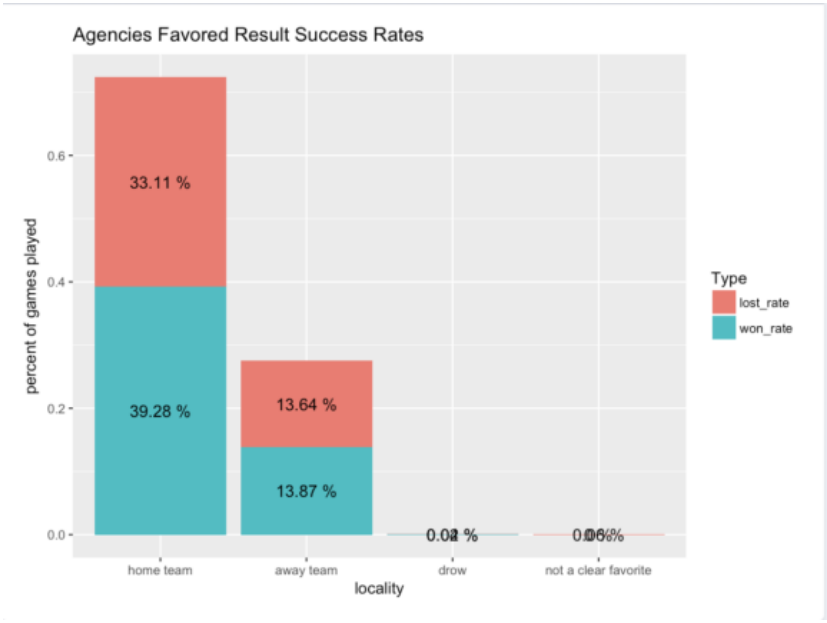
Next, I wanted to check how accurate the favored result was in terms of predicting the match outcome. I found that the agencies' favored result (represented by the minimum payout) had an average success rate of 53%. This was consistent for each of the seasons in the data frame.



I also wanted to examine to what degree the location made a difference in the payouts given to favored teams. For instance, I would expect the payout for a favored home team to be lower than a favored away team. After all, it is widely believed that the home team has the higher advantage. I used two box plots, which demonstrated that there is *almost no significant difference in the minimum payouts that can be attributed to location.*

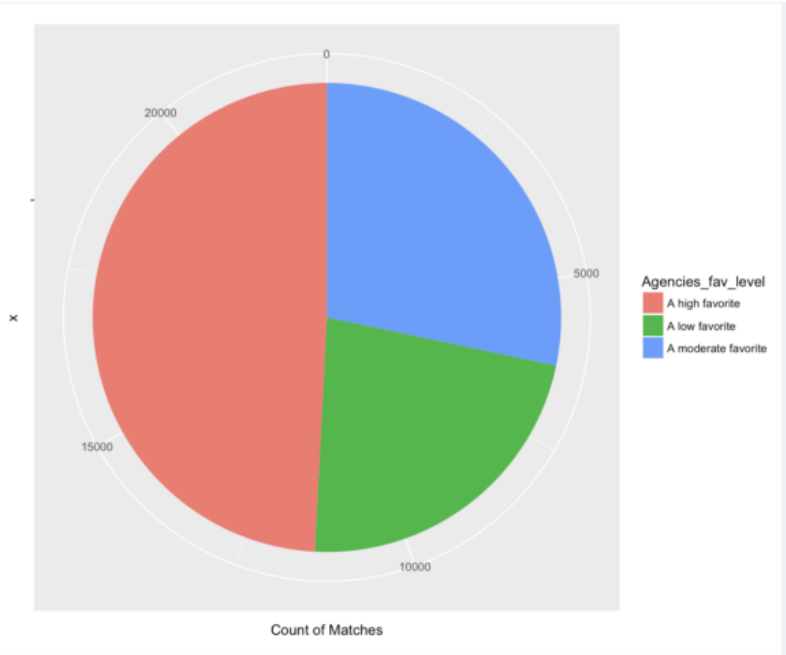


If the minimum payout for a favored home team and a favored visiting team are almost identical, does this mean that the rates of winning them are the same?

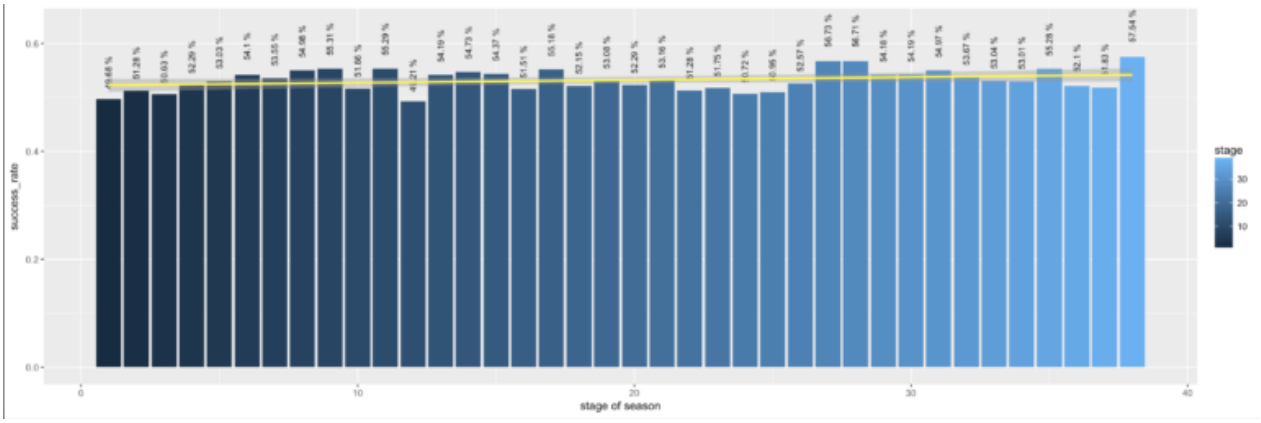


In fact, no, the rates of winning are not the same. As we can see from this bar chart, the favored home team had a 55% chance of winning while the favored away team had a 50.5% success rate. Although these are not earth-shattering numbers, could this be an opportunity for the betting agencies to attract more gamblers? By raising the payout for the favored away team, they are promising larger compensation when in fact the probability of the favored away team winning is actually quite low.

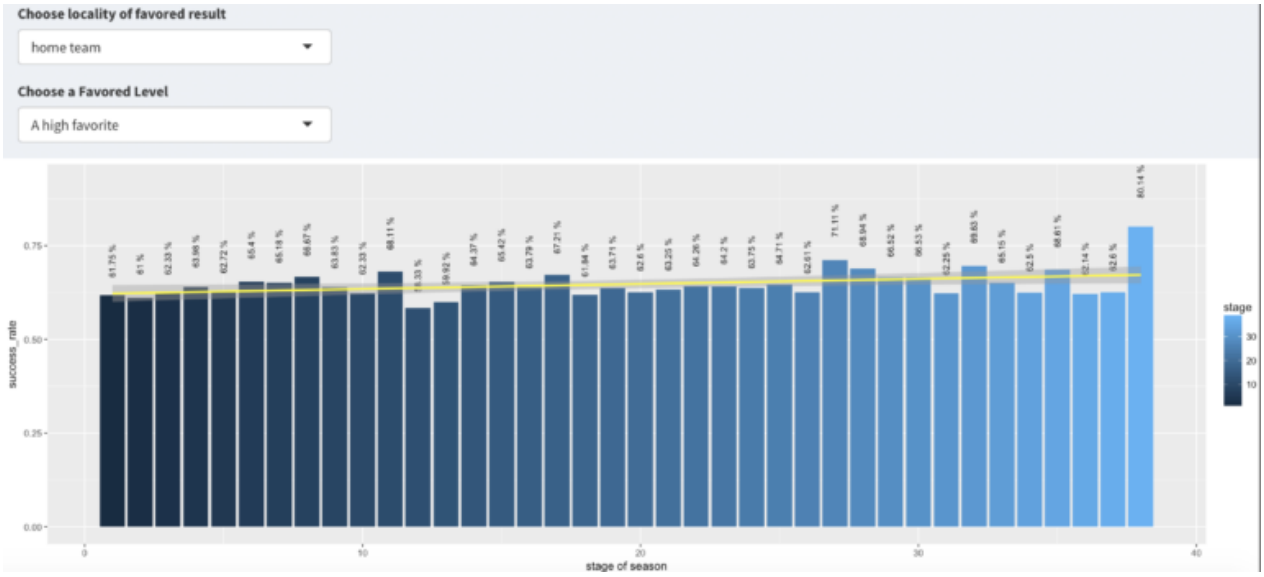
I wanted to further visualize the proportion of games that can be cataloged as high favored, moderate favored, and low favored odds. Because of the league structure, like number of games and arrangement of opposition, there are more games where the difference in team skills and strengths are large. Therefore, we see that almost 50% of the matches were categorized as high favored, and more than a quarter that is moderate favored.

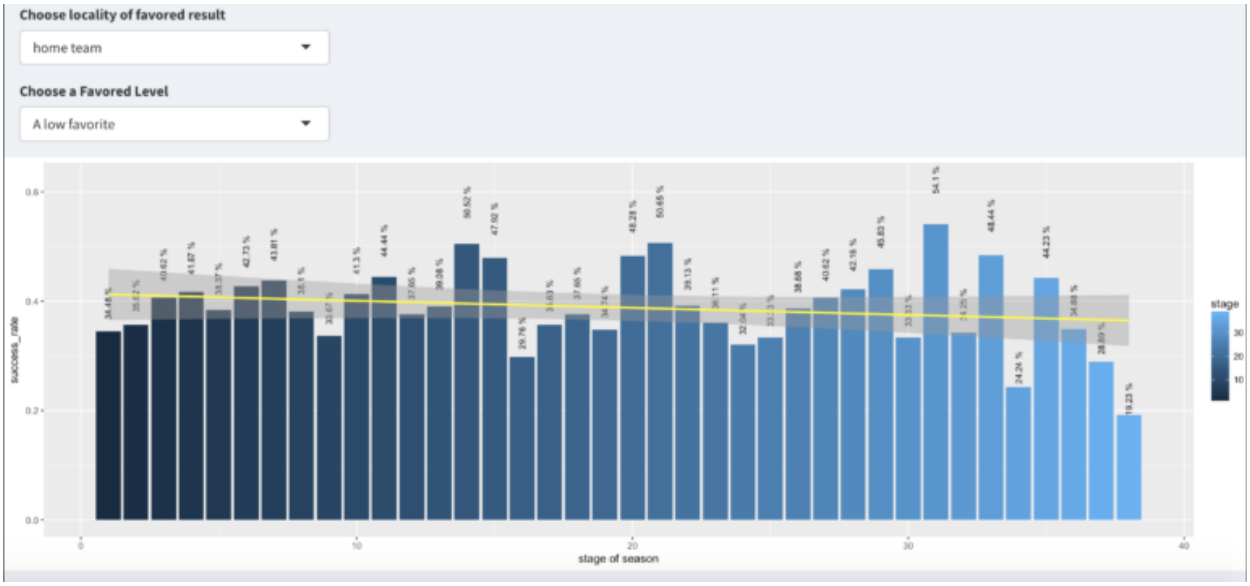


One of my more interesting findings was that the accuracy of the predicted wins increased during the later stages of the season. The late stages of the season carry higher stakes than the early stages, as this when the champion and the relegations are decided. In the chart below, we can see a slight improvement in the success rate over the span of a season.

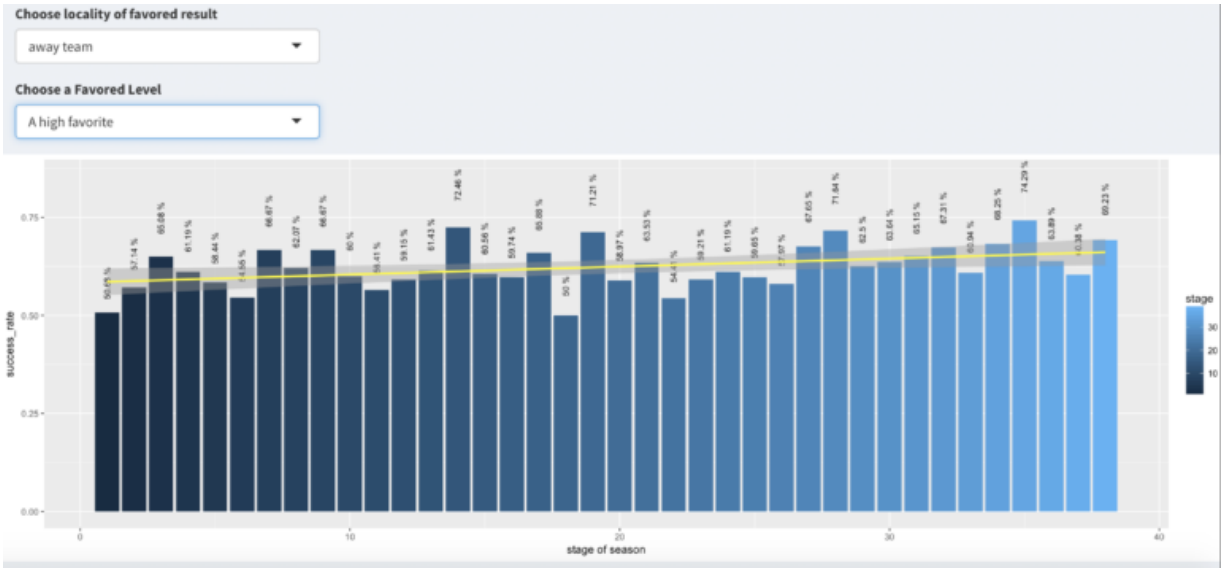


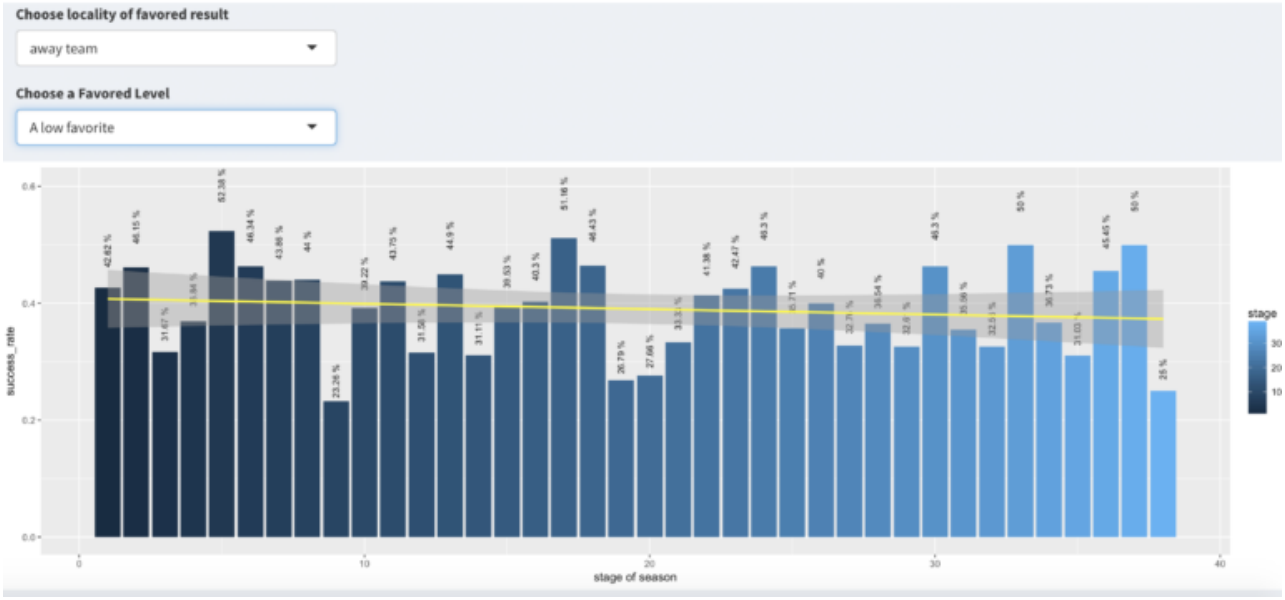
Next, I checked the stage of the season against the location of the team as well as the favored level, and I found that the success rate increases over the span of a season when the high favored team plays in their home arena. It seems as if the high favored teams are capable of winning in the important stages of the season, while on the contrary, the low favored team success rate decreases over time, as if the pressure of the last stages of the season and the presence of their fans have a negative effect on their performance.





The results of the away team demonstrate a different trend. While the high favored teams' success rate increases over time, the low favored teams show inconsistent success rates throughout the season. It seems that there is no clear effect of the season stage on their performance.





Notes

Moreover, in my shiny app, you can explore the success rates broken down by the location and the favored level. When I did this, I found a clear pattern: matches with a high favored result have a success rate of around 65% and matches with a low favored result have a success rate that ranges from 33%-43%. Moreover, when cross-examining the data, I found that there were three parameters that carried the most weight when determining a probability of a match outcome:1) the agencies' high favored result 2) the location of the team (home vs. away), and 3) the stage of the season. Ultimately, choosing a high favored team in their home arena, in the late stages of a season can raise the probability of winning the bet by 80%.

Email*

Submit

About Author

Chen Trilnik

View all posts by Chen Trilnik >

Related Articles

R SHINY

Uber Optimization: Finding Passengers Faster

by Bennett Gelly

May 8, 2018

Motivation Ridesharing drivers and companies, such as Uber and Lyft, can make more money by filling their cars...

R

The Changing Landscape in the NBA

by Emanuel Kamali

May 7, 2018

Motivation As an avid fan of the NBA and NBA technology, one can argue that the landscape of...

Continue Reading

Continue Reading

Leave a Comment

Your email address will not be published. Required fields are marked *

Write a response...

Name *

Email *

fanatik April 19, 2018

I just like the helpful information you supply in your articles. I will bookmark your blog and take a look at again here frequently. I'm quite certain I will be told many new stuff proper right here! Best of luck for the next!

Online Training

Careers

Corporate Offerings

Hiring Partners

Social Media