Random Football

Predicting NFL Game Winners using Random Forests

•••

July 8, 2019

Data Overview

- **NFL Scores and Stadium Data**
  https://www.kaggle.com/tobycrabtree/nfl-scores-and-betting-data
  - Game info going back to 1966 including scores, weather, & betting spread and over/under

- **FiveThirtyEight Elo Data**
  https://projects.fivethirtyeight.com/nfl-api/nfl_elo.csv
  - FiveThirtyEight Elo ratings for each game back to 1920

Understanding the problem

- **Can we use the data to predict the winner of NFL games?**
  - Difficulty -
    - Football games are pretty random
    - Many non-random aspects of a game are still difficult to quantify (ex: injuries)
    - For these reasons, we can't expect to be able to predict games 100%
  - Realistic Goal -
    - Beat Vegas in choosing the winner (based on the point spread favorite team)

Understanding the Data

Data (220 KB)

Data Sources

- nfl_stadiums.csv 100 x 15
- nfl_teams.csv 41 x 8
- spreadspoke_scores.csv 12.4k x 17
- spreadspoke.R

About this file

NFL football games since the 1966 season with game results and descriptive info including if a playoff game, played at a neutral site, and weather information if available. Data set was built from publicly available NFL data, weather provided by the NOAA, and betting data from a variety of sources but cross referenced with Pro Football Reference.

Columns

- schedule_date
- schedule_season
- schedule_week
- schedule_playoff
- team_home
- score_home
- score_away
- team_away
- team_favorite_id
- spread_favorite
- over_under_line
- stadium
- stadium_neutral
- weather_temperature
- weather_wind_mph
- weather_humidity
- weather_detail

Source: https://www.kaggle.com/tobycrabtree/nfl-scores-and-betting-data

Features

- Home team/away team
- Season/Year
- Week of game
- Weather
- Elevation of stadium

Added Features

- For each team in each game:
  - Distance travelled since last game
  - Days elapsed since last game
  - Season win percentage prior to current game
  - Season average score prior to current game
- From 538
  - Elo probability of home team win

Elo?

Via Wikipedia: Elo ratings are a system developed by Hungarian-American physics professor Arpad Elo to rate relative skill levels between chess players.

Players/teams are ranked with points after each win/loss based on their previous ranking.

Modeling

Random Forest Regressor

Grid Search

Strategies

Model Features

From Kaggle Data:

1. Season (Year)
2. Week of Season
3. Elevation
4. Home Team
5. Away Team
6. Weather Conditions

From FiveThirtyEight:

7. Elo Probability of Home Team Win

Added Features:

8. Home / Away (H/A) Season Win %
9. H/A Average Season Score
10. H/A Travel Distance
11. H/A Days since last game

Random Forest

- Initial testing performed on 2017 NFL Season
- Training data used the 2013-2016 NFL Seasons
- Grid Search performed to find optimal parameters

Random Forest

- Accuracy: 64%
  1. Naive Baseline (Always Pick Home) - 57%
  2. Target Baseline (Vegas Favorite) - 67%
  3. Elo Baseline (Elo odds only) - 64%

- Most Important Features (Importance Percentage):
  1. Elo Probability Home - 22%
  2. Away Team Average Score - 9%
  3. Home Team Season Win % - 8%
  4. Away Team Season Win % - 7%
  5. Away Team Travel Distance - 7%

Alternative Strategies

- We define a range of predicted probabilities between 40%-60%, which we will call “close-calls”
- Strategy 1:
  - Pick the Home Team if the game is a close-call
  - Initial model (2017) Accuracy: 62%
- Strategy 2:
  - Don’t bet on the game at all if it is a close-call
  - Initial model (2017) Accuracy: 68%

Results

Test Season	Training Season(s)	Baseline All-Home Accuracy	Target Vegas Accuracy	Random Forest Accuracy	Close-call Home Accuracy	Close-call Abstain Accuracy
2012	2011	56.55%	63.67%	62.17%	61.42%	69.43%
2013	2011-2012	59.18%	64.42%	60.30%	58.43%	69.60%
2014	2011-2013	57.30%	68.54%	66.29%	62.55%	70.00%
2015	2011-2014	54.31%	64.04%	63.30%	59.18%	70.27%
2016	2012-2015	57.68%	67.04%	64.04%	58.80%	66.87%
2017	2013-2016	56.93%	60.30%	64.04%	62.17%	68.39%
2018	2014-2017	58.80%	67.42%	62.17%	63.67%	64.33%

Conclusions

- Random Forest model accuracy (62%-66%) regularly beats a naive, home-team-wins strategy (54%-59%).

- With Elos included and using a strategy of not betting on close-calls, Random Forest Regressor model accuracy (64%-70%) beats out Vegas odds (60%-68%).

The Team

Berry Brooks, CEO

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor

Ronny Reader, CFO

Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat