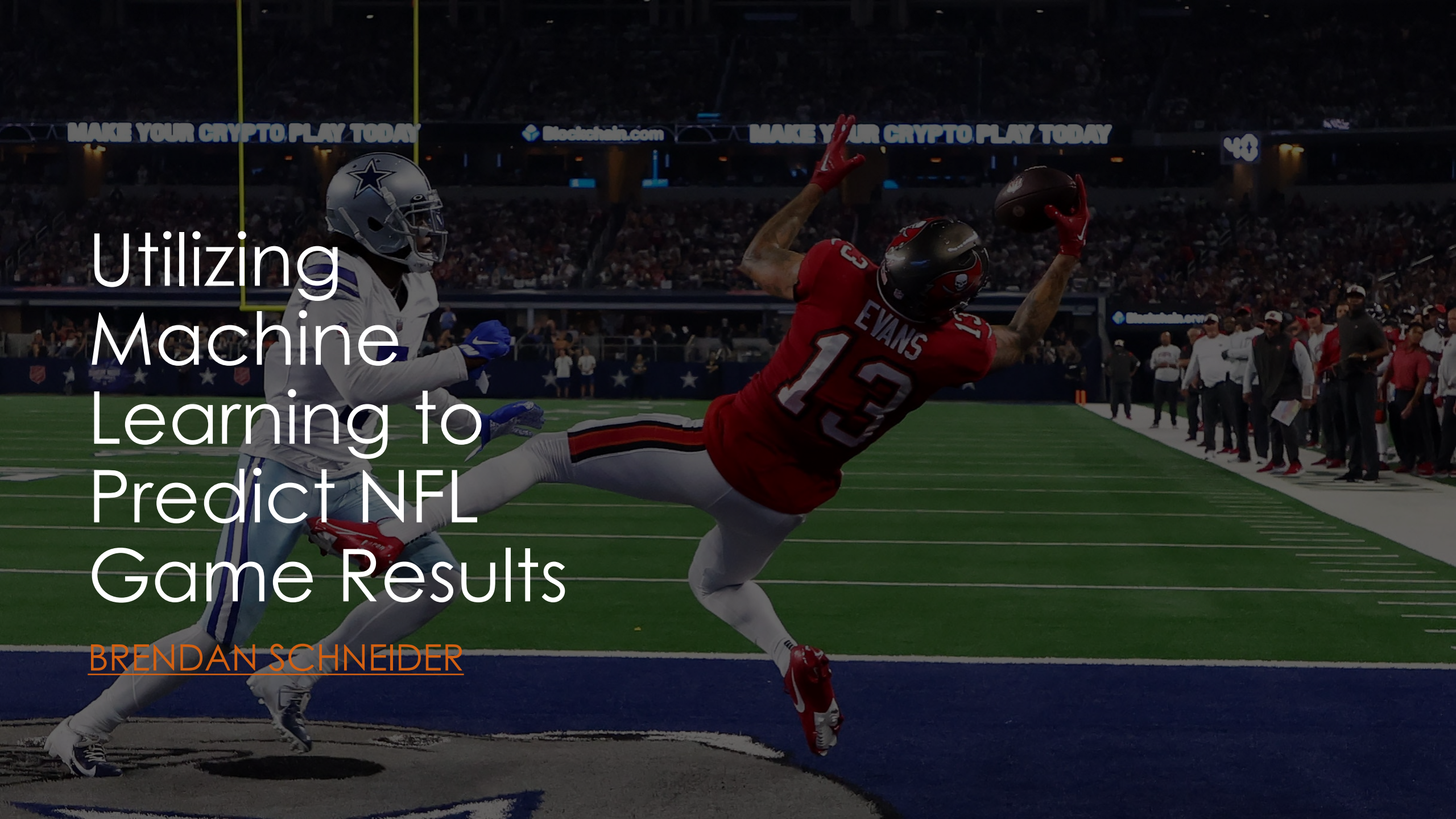


Utilizing Machine Learning to Predict NFL Game Results

BRENDAN SCHNEIDER

Questions

- What predictions can we make about NFL games?
- Can we predict the winner? The final score? The spread?
- How accurately?
- What data is available to us for answering these questions?

Web Scraping

- Pro Football Reference has game statistics for each team

Schedule & Game Results Share & Export ▾ Glossary

									Score		Offense					Defense					Expected Points		
Week	Day	Date				OT	Rec		Opp	Tm	Opp	1stD	TotYd	PassY	RushY	TO	1stD	TotYd	PassY	RushY	TO	Offense	Defense	Sp. Tms
1	Sun	September 10	4:25PM ET	boxscore	W		1-0	@	New England Patriots	25	20	17	251	154	97	1	24	382	306	76	2	-10.42	12.04	4.39
2	Thu	September 14	8:15PM ET	boxscore	W		2-0		Minnesota Vikings	34	28	24	430	171	259	1	19	374	346	28	4	14.53	-7.44	-0.51
3	Mon	September 25	7:15PM ET	boxscore	W		3-0	@	Tampa Bay Buccaneers	25	11	27	472	271	201	2	12	174	133	41	2	14.08	7.28	-1.78
4	Sun	October 1	1:00PM ET	boxscore	W	OT	4-0		Washington Commanders	34	31	23	415	311	104		26	365	258	107		18.21	-17.68	2.47
5	Sun	October 8	4:05PM ET	boxscore	W		5-0	@	Los Angeles Rams	23	14	28	454	295	159	1	17	249	195	54		16.56	-6.25	-2.84
6	Sun	October 15	4:25PM ET	boxscore	L		5-1	@	New York Jets	14	20	24	348	268	80	4	17	244	155	89		-5.91	7.52	-8.89
7	Sun	October 22	8:20PM ET	boxscore	W		6-1		Miami Dolphins	31	17	26	355	256	99	2	12	244	199	45	1	11.06	7.17	-2.90
8	Sun	October 29	1:00PM ET	boxscore	W		7-1	@	Washington Commanders	38	31	21	374	315	59	2	26	472	388	84	1	15.23	-10.32	4.32
9	Sun	November 5	4:25PM ET	boxscore	W		8-1		Dallas Cowboys	28	23	20	292	183	109		27	406	333	73	1	15.00	-10.90	2.75
10									**Bye Week**															
11	Mon	November 20	8:15PM ET	preview				@	Kansas City Chiefs	0														
12	Sun	November 26	4:25PM ET	preview					Buffalo Bills	0														
13	Sun	December 3	4:25PM ET	preview					San Francisco 49ers	0														
14	Sun	December 10	8:20PM ET	preview				@	Dallas Cowboys	0														
15	Sun	December 17	4:25PM ET	preview				@	Seattle Seahawks	0														
16	Mon	December 25	4:30PM ET	preview					New York Giants	0														
17	Sun	December 31	1:00PM ET	preview					Arizona Cardinals	0														
18	Sun	January 7	1:00PM ET	preview				@	New York Giants	0														

Web Scraping

- Matthew Kim built a project to scrape data from Pro Football Reference
- I leveraged his team_game_log.py file to collect this data myself, pulling additional statistics and enabling it to work for the current season
- Scraping 16 teams at a time avoided a Too Many Requests error
- Assumption: regular season data since 2010 is enough for noteworthy findings

Data Cleansing

- Not all franchises used the same name
- Not every franchises had the same home stadium
- Some games were played at a neutral location

Feature Engineering – Average Stats

- Making predictions for a game before it is played means we'll need to utilize data from previous games
- Created columns for average values of statistics for each team over the past 1-8 games
- Averages include stats for the offense and defense of each team, as well as their opponent for that week

Elo Rating System

- Arpad Elo created a system to assign values to chess players to rate their skill; this system can be applied to other games
- Elo ratings factors in quality of the competitor, as opposed to pure numerical statistics in the game
- Defeating opponents with high Elo ratings and by large margins increases a team's Elo score greatly, and vice versa
- Analysis begins in 2012, letting Elo settle

Elo Rating System

- Formulas from a UPenn student project and FiveThirtyEight
- Each team begins at 1500

$$Elo_{n+1} = Elo_n + k \cdot m \cdot (R - E)$$

$$R = \begin{cases} 1 & \text{if Win} \\ 0.5 & \text{if Tie} \\ 0 & \text{if Loss} \end{cases}$$

$$k = 20$$

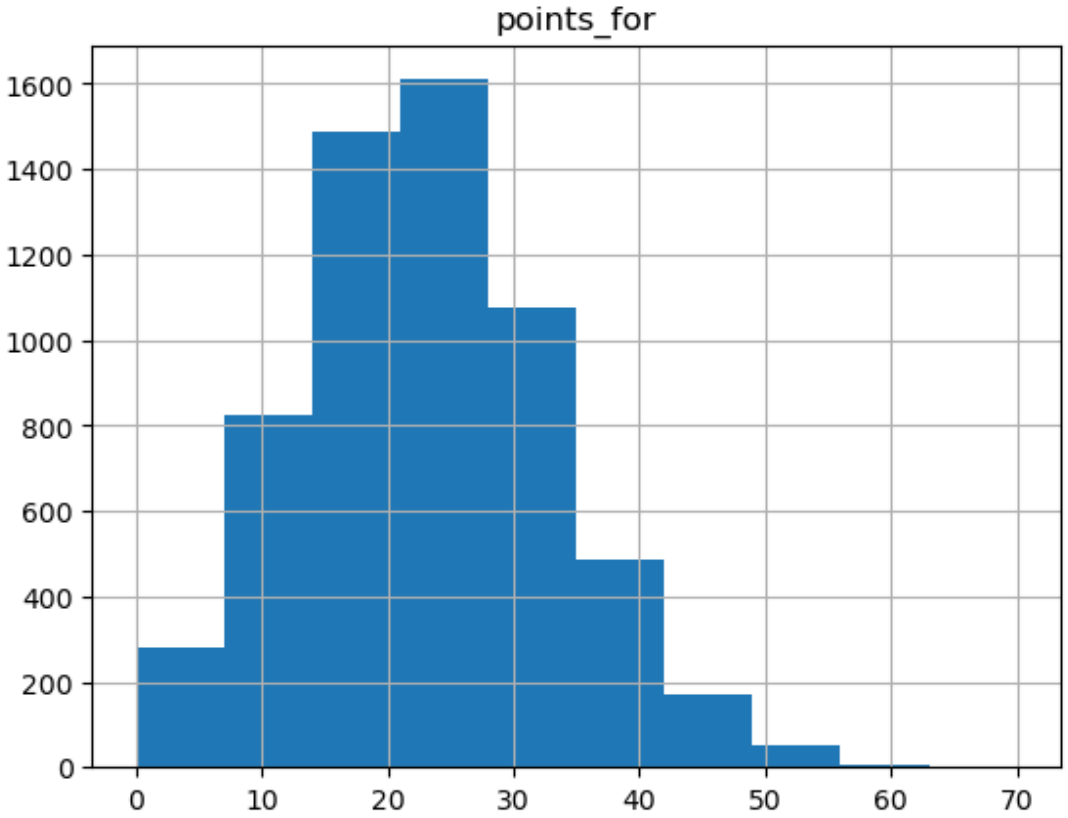
$$E = \frac{1}{10^{\frac{elo_{opp} - elo_{team}}{400}} + 1}$$

$$m = \frac{ln(abs(mov + 1)) \cdot 2.2}{(Elo_W - Elo_L) * 0.001 + 2.2}$$

$\frac{elo_{opp} - elo_{team}}{25}$ is the point spread

Exploratory Data Analysis

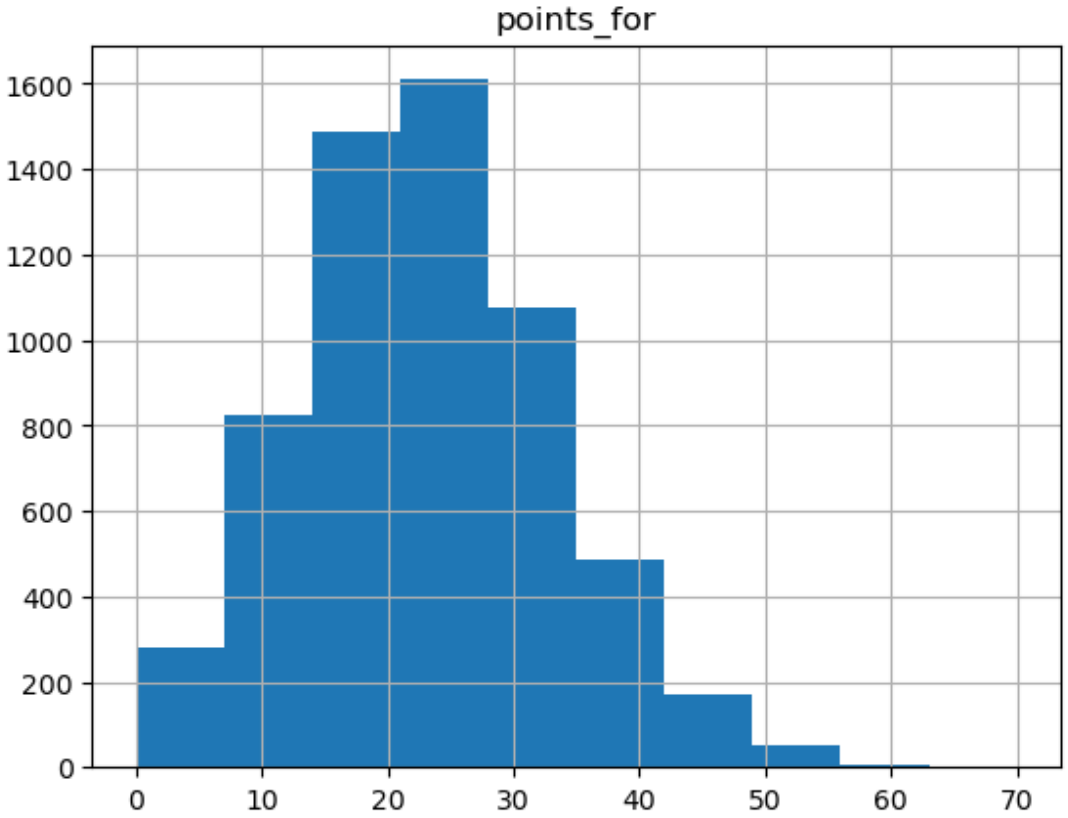
Descriptions of Points Scored by game outcome

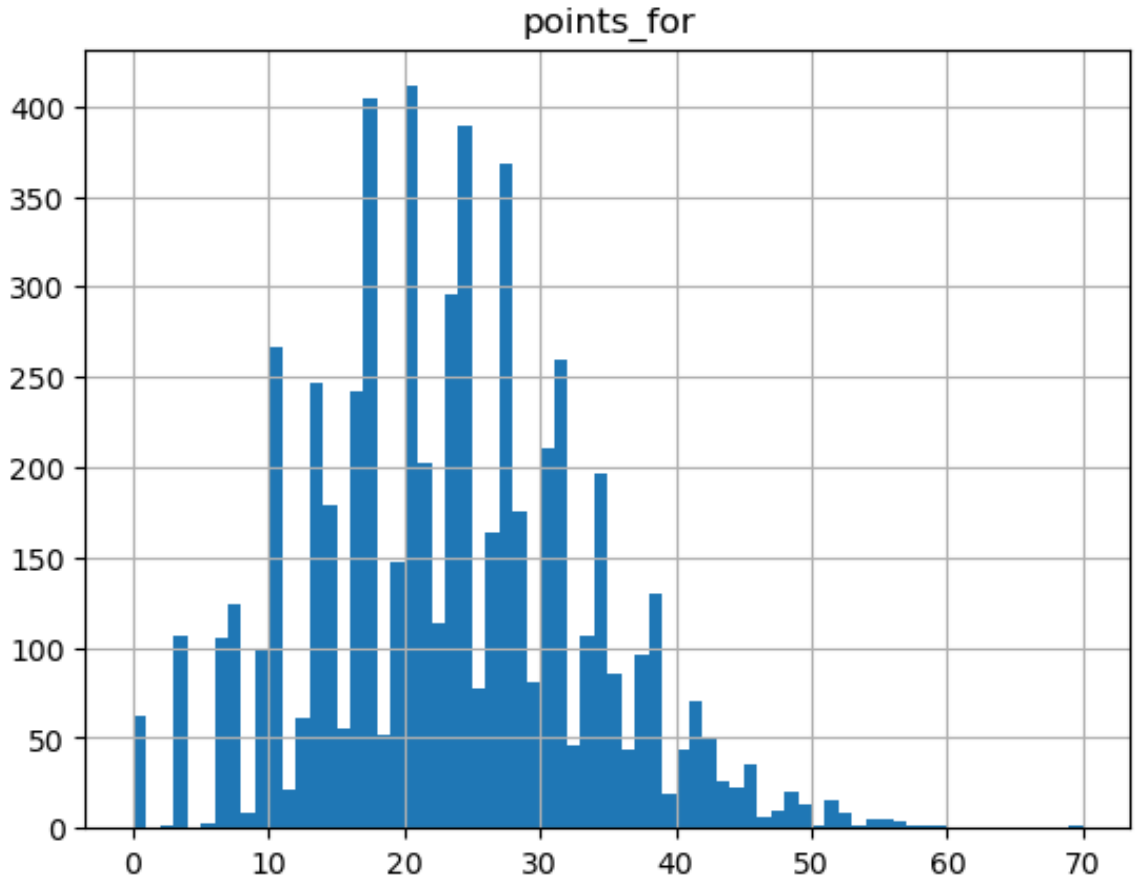


	result	L	T	W
points_for	count	2985.000000	24.000000	2985.000000
	mean	17.137688	23.000000	28.532998
	std	8.139830	7.424987	8.521204
	min	0.000000	6.000000	6.000000
	25%	10.000000	20.000000	23.000000
	50%	17.000000	23.500000	27.000000
	75%	23.000000	27.000000	34.000000
	max	51.000000	37.000000	70.000000

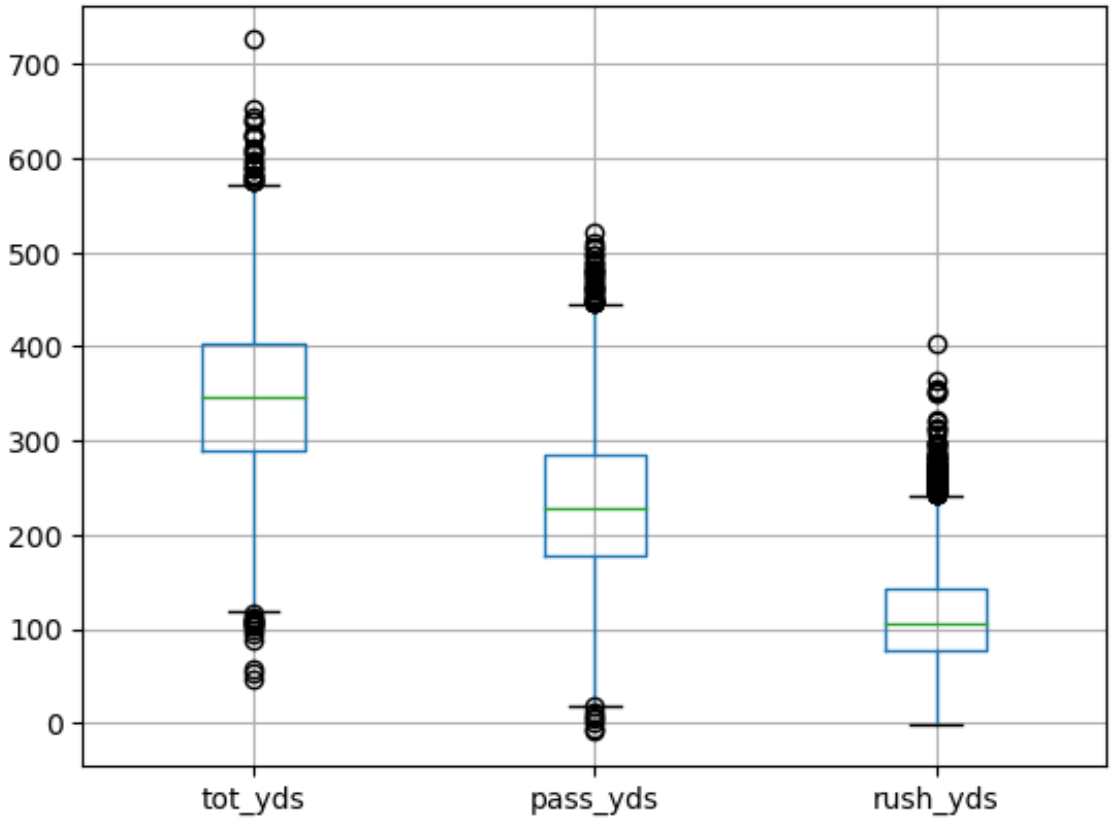
Exploratory Data Analysis

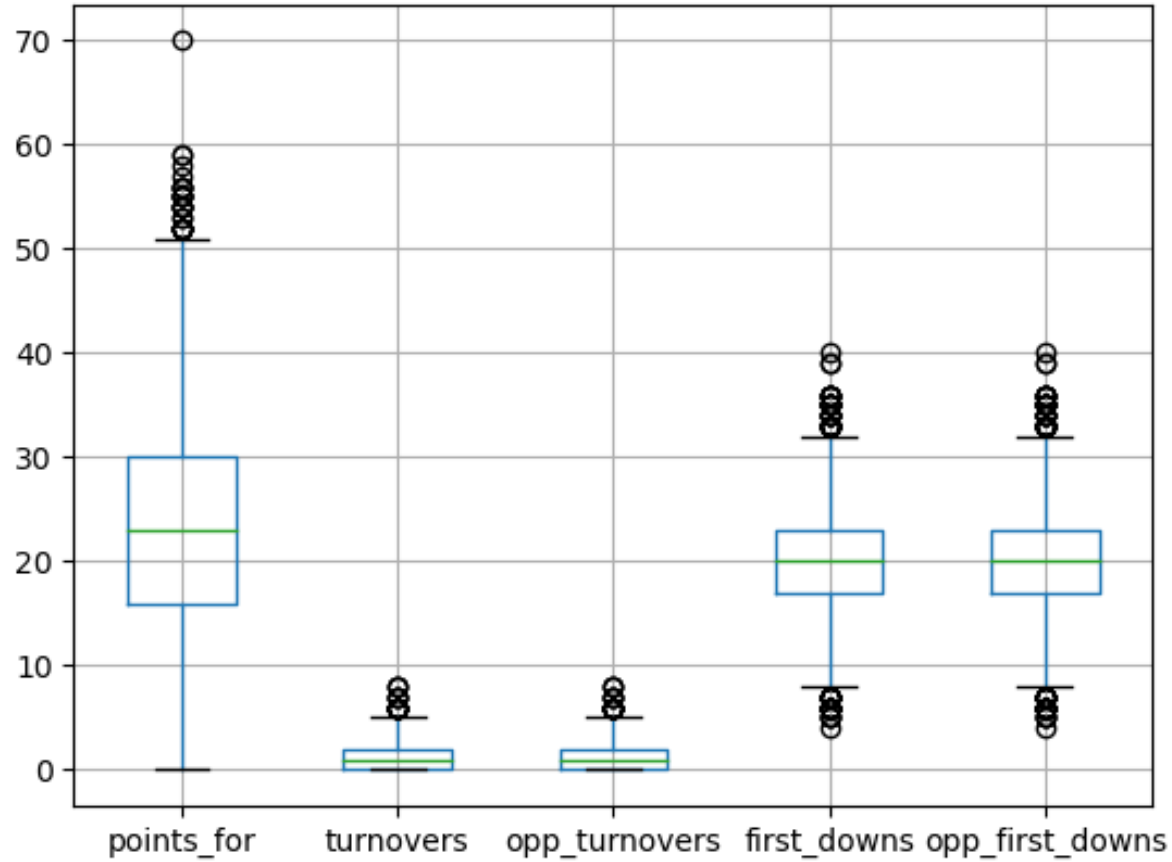
Some scores are more common than others





Exploratory Data Analysis





Exploratory Data Analysis

	team_name	elo_end
BUF-2022-18	Buffalo Bills	1734.919212
KC-2022-18	Kansas City Chiefs	1727.509146
SF-2022-18	San Francisco 49ers	1704.661531
CIN-2022-18	Cincinnati Bengals	1671.018684
PHI-2022-18	Philadelphia Eagles	1618.597923
DAL-2022-18	Dallas Cowboys	1612.895432
MIN-2022-18	Minnesota Vikings	1563.328269
PIT-2022-18	Pittsburgh Steelers	1555.668865
GB-2022-18	Green Bay Packers	1552.728835
BAL-2022-18	Baltimore Ravens	1534.660689
LAC-2022-18	Los Angeles Chargers	1521.444290
SEA-2022-18	Seattle Seahawks	1518.356826
NE-2022-18	New England Patriots	1517.907219
NO-2022-18	New Orleans Saints	1513.355042
DET-2022-18	Detroit Lions	1506.764497
JAC-2022-18	Jacksonville Jaguars	1500.177342
MIA-2022-18	Miami Dolphins	1498.664151
TB-2022-18	Tampa Bay Buccaneers	1495.721416
WAS-2022-18	Washington Commanders	1491.742564
CLE-2022-18	Cleveland Browns	1488.456307
LV-2022-18	Las Vegas Raiders	1471.912069
CAR-2022-18	Carolina Panthers	1467.047644
NYG-2022-18	New York Giants	1453.229181
TEN-2022-18	Tennessee Titans	1449.565974
ATL-2022-18	Atlanta Falcons	1434.653623
LAR-2022-18	Los Angeles Rams	1416.438617
NYJ-2022-18	New York Jets	1390.813865
ARI-2022-18	Arizona Cardinals	1382.056569
DEN-2022-18	Denver Broncos	1371.448257
IND-2022-18	Indianapolis Colts	1366.307937
CHI-2022-18	Chicago Bears	1315.877083
HOU-2022-18	Houston Texans	1307.002723

Can a Linear Regression Predict Points Scored?

- The RMSE for the mean was 10.16 points (per team)
- The best RMSE I could find was 9.43 points (per team)
- The average margin of victory was 11.35 (total)

```
feature_options = ['elo_start',
                   'elo_start_opp',
                   'home_team',
                   'distance_travelled_opp_diff',
                   '7_game_avg_points_for',
                   '7_game_avg_points_allowed',
                   '6_game_avg_tot_yds',
                   '2_game_avg_exp_pts_off',
                   '2_game_avg_exp_pts_def',
                   '7_game_avg_exp_pts_st',
                   '7_game_avg_points_allowed_opp',
                   '7_game_avg_points_for_opp',
                   '6_game_avg_opp_first_downs_opp',
                   '6_game_avg_exp_pts_def_opp']

train_test_rmse(df, feature_options)
```

```
RMSE for Mean: 10.16302675937795
RMSE for Median: 10.179509563004277
RMSE for Mode: 11.593659558625461
RMSE for ['elo_start', 'elo_start_opp', 'home_team', 'distance_travelled_opp_diff', '7_game_avg_points_for', '7_gam
e_avg_points_allowed', '6_game_avg_tot_yds', '2_game_avg_exp_pts_off', '2_game_avg_exp_pts_def', '7_game_avg_exp_pt
s_st', '7_game_avg_points_allowed_opp', '7_game_avg_points_for_opp', '6_game_avg_opp_first_downs_opp', '6_game_avg_
exp_pts_def_opp'] is
9.434421682611026
```

Can a Logistic Regression Predict the Winner of an NFL game?

- The best accuracy I could get was 65.38%
- Compare to the 55% win rate of home teams
- Picking the favorite according to this Elo rating each game has an accuracy of 63.6%

```
logreg = LogisticRegression()

feature_cols = ['elo_opp_diff_team',
                '8_game_avg_points_for',
                '8_game_avg_points_for_opp',
                '4_game_avg_exp_pts_off',
                'home_team'
               ]

X = df[feature_cols]
y = df['result_letter']

X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=123)

LR = LogisticRegression()
LR.fit(X_train, y_train)

pred = LR.predict(X_test)
```

```
LR.score(X, y)
```

```
0.6451451451451451
```

```
LR.score(X_train, y_train)
```

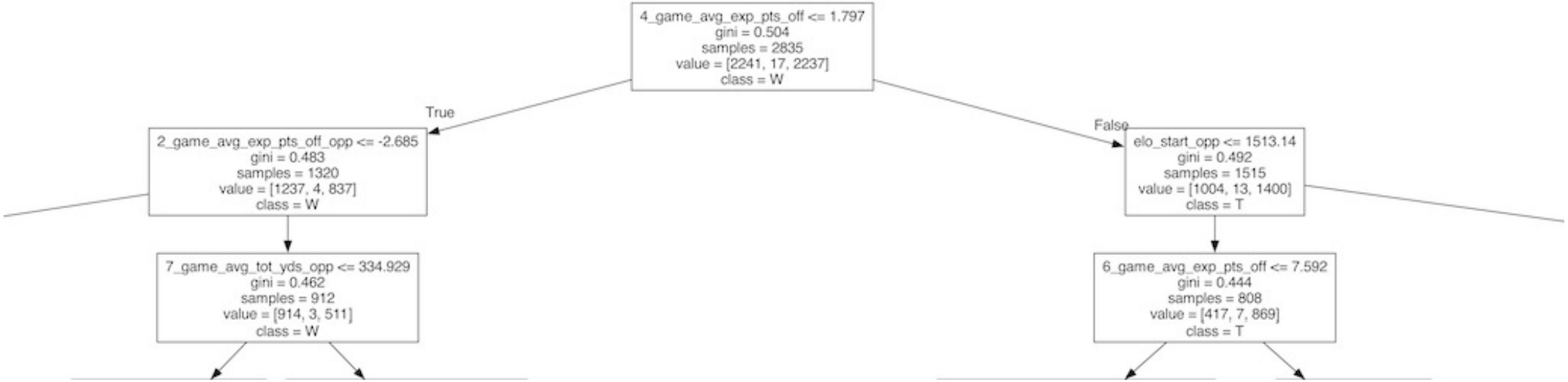
```
0.6422691879866519
```

```
LR.score(X_test, y_test)
```

```
0.6537691794529686
```

Can a Random Forest Predict the Winner of an NFL Game

- Using a Random Forest Classifier, we were able to correctly model 65.58% of games



The beginning of Tree #7 in the Random Forest

How Did We Do?

- Modelling points scored was too variable to predict values
- Logistic Regression and Random Forest got us to around 65% accuracy
- This is about 10% more games correctly picked than picking the home team, and just a bit better than picking the favorite according to Elo
- Compare to experts at ESPN, NFL, CBS, Pickwatch and other cites, it's on par with the best experts

Areas for Improvement

- Fix collinearity between Elo features and window features
  - Look for advanced metrics
- Try exponentially weighting recent games instead of averaging previous games equally
- Improve Elo calculation (FiveThirtyEight considers home field, rest days, and quarterbacks)
- Incorporate player-level data

Bibliography

- Matthew Kim
  - https://pypi.org/project/pro-football-reference-web-scraper/
  - https://pypi.org/user/mjk9/
  - https://github.com/mjk2244/pro-football-reference-web-scraper
- Josh Weiner
  - https://towardsdatascience.com/predicting-the-outcome-of-nba-games-with-machine-learning-a810bb768f20
  - https://github.com/JoshWeiner/NBA_Game_Prediction

Bibliography (Cont.)

- FiveThirtyEight
  - https://fivethirtyeight.com/methodology/how-our-nfl-predictions-work/
- Pro Football Reference
  - https://www.pro-football-reference.com/
- Open Source Football
  - https://opensourcefootball.com/posts/2021-01-21-nfl-game-prediction-using-logistic-regression/
- Will Koehrsen
  - https://towardsdatascience.com/how-to-visualize-a-decision-tree-from-a-random-forest-in-python-using-scikit-learn-38ad2d75f21c
- Wikipedia
  - https://en.wikipedia.org/wiki/Chronology_of_home_stadiums_for_current_National_Football_League_teams
- Google Images