Sports Analytics

Machine Learning to Predict Sporting Events

About Me

https://www.kaggle.com/mconway

https://www.linkedin.com/in/markrconway

The Sports Book

The Odds Board

Online Odds Board

Sports Betting 101

- You are not just predicting the outcome of a game, known as Straight Up (**SU**). You are betting Against The Spread (**ATS**).

- The oddsmakers define a **point spread**, or **line**. It is a median number between two teams set to generate bets for *both* teams.

- In each game, you have a favorite (**chalk**) and an underdog (**dog**), although some games are known as **pick 'ems**, where the teams are equally matched.

Line Examples

TEAMS	LINE	BETS	TOTAL	BETS
New England	-3.5	66%	44o	59%
Denver	+3.5	34%	44u	41%
» Matchup	Sports Authority Field at Mile High		Jan 24@ 3:05 PM	
Arizona	+3	42%	47o	63%
Carolina	-3	58%	47u	37%
» Matchup	Bank of America Stadium		Jan 24@ 6:40 PM	

- Here, New England is currently favored by 3.5 points (**-** sign) over Denver. Arizona is an underdog by 3 points (**+** sign)

- The Over/Under is 44, meaning that the total points scored for both teams is estimated to be 44. If the final score of the game is 31-17, then the game would be "over" (48 total points). If the score is 24-17, then the game would be "under" (41 total points). A score of 24-20 would be a "push", i.e., no winners.

Money Lines

- A **Moneyline** is premised on a $100 bet. For example, if the Yankees are favored at -185, then you must bet $185 to win $100. Conversely, if the Red Sox are underdogs at +165, then you bet $100 to win $165.

- Unlike point spreads, moneylines are asymmetric and can be converted to probabilities that either team will win.

- Point spreads are most appropriate for games with a relatively large number of points: basketball and football.

- For sports such as baseball, soccer, and hockey, money lines are generally used, although you will see spreads.

Strategies

- You need to win at least 56% of the time to make a living because of the "vig" or "juice". Think of it as a commission for every bet you place.

- **Fading the Public**: "*In the NFL over the past eight seasons, games in which 75% of the public is on one side lost roughly 53-54% of the time, obviously meaning that fading them has resulted in more wins than losses. Likewise, large underdogs were among the best bets during this stretch with the underdog covering the spread 55% of the time when 70-75% of the public was on a favorite of 7+ points, which is one of the key numbers in football betting.*"

Strategies, continued

- Create your own line based on team statistics, player statistics, and handicapping personnel moves, e.g., key injuries. Compare your line against the established line to compute your edge.

- There is an opening line and closing line. Some professionals observe the direction and magnitude between the two to spot any patterns.

- **Finding Anomalies**: If the line is not as expected, then someone probably knows something. Please refer to:

  http://www.buzzfeed.com/heidiblake/the-tennis-racket#.pwIapMVMO

Enter Machine Learning

- Keeping general strategy in mind, apply machine learning algorithms to predict game outcomes using supervised learning, i.e., classification.

- We will create binary features to determine whether or not a team will win the game or cover the spread.

- We can also try to predict whether or not the total score will be *over* or *under*.

Classification
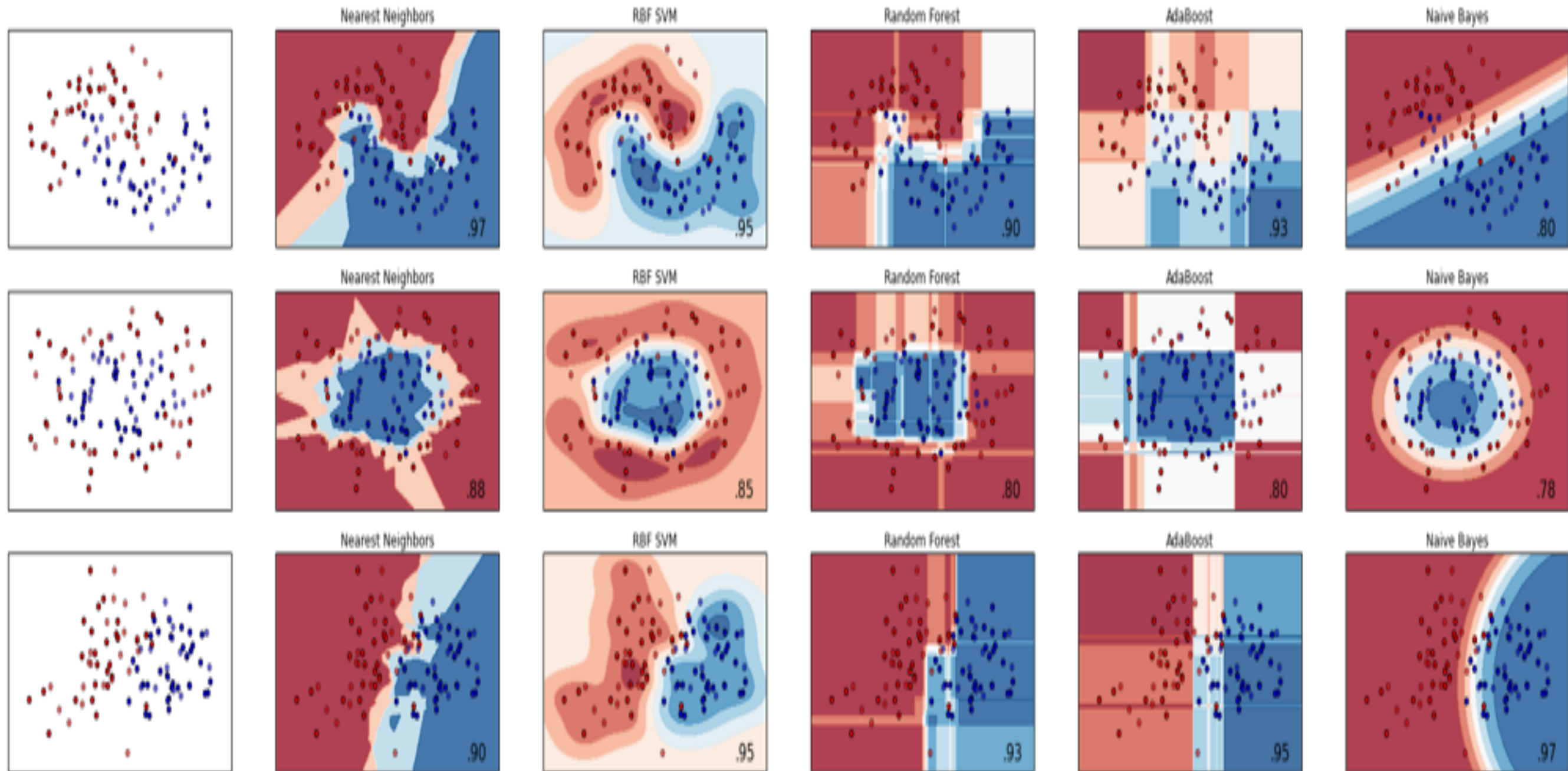
Data

	A	B	C	D	E	F	G	H
1	season	date	away.team	away.score	home.team	home.score	line	over_under
515	2015	2016-01-03	TEN	24	IND	30	-4.0	39.0
516	2015	2016-01-03	WSH	34	DAL	23	-3.0	41.0
517	2015	2016-01-03	PHI	35	NYG	30	-3.5	51.0
518	2015	2016-01-03	DET	24	CHI	20	2.5	45.0
519	2015	2016-01-03	TB	10	CAR	38	-10.5	44.5
520	2015	2016-01-03	OAK	17	KC	23	-7.0	44.0
521	2015	2016-01-03	SD	20	DEN	27	-10.0	42.0
522	2015	2016-01-03	SEA	36	ARI	6	-6.0	47.0
523	2015	2016-01-03	STL	16	SF	19	3.0	38.5
524	2015	2016-01-03	MIN	20	GB	13	-3.0	43.5
525	2015	2016-01-09	KC	30	HOU	0	3.0	39.5
526	2015	2016-01-09	PIT	18	CIN	16	2.0	44.5
527	2015	2016-01-10	SEA	10	MIN	9	4.0	40.0
528	2015	2016-01-10	GB	35	WSH	18	-1.0	46.0
529	2015	2016-01-16	KC	20	NE	27	-5.0	43.5
530	2015	2016-01-16	GB	20	ARI	26	-7.5	49.5
531	2015	2016-01-17	SEA	24	CAR	31	-2.5	42.5
532	2015	2016-01-17	PIT	16	DEN	23	-7.5	41.0

Features [nflgame 1.2.19]

Here's a quick teaser to find the top 5 running backs by rushing yards in the first week of the 2013 season:
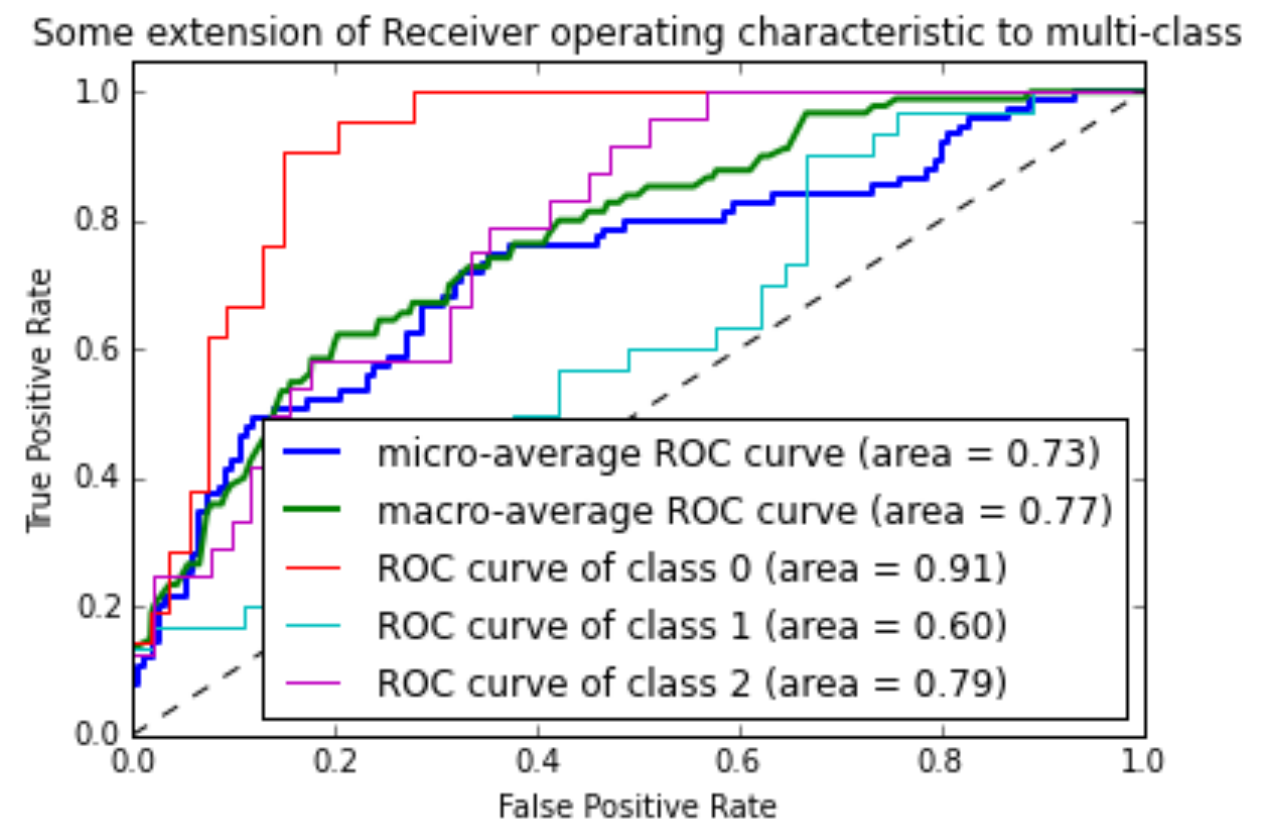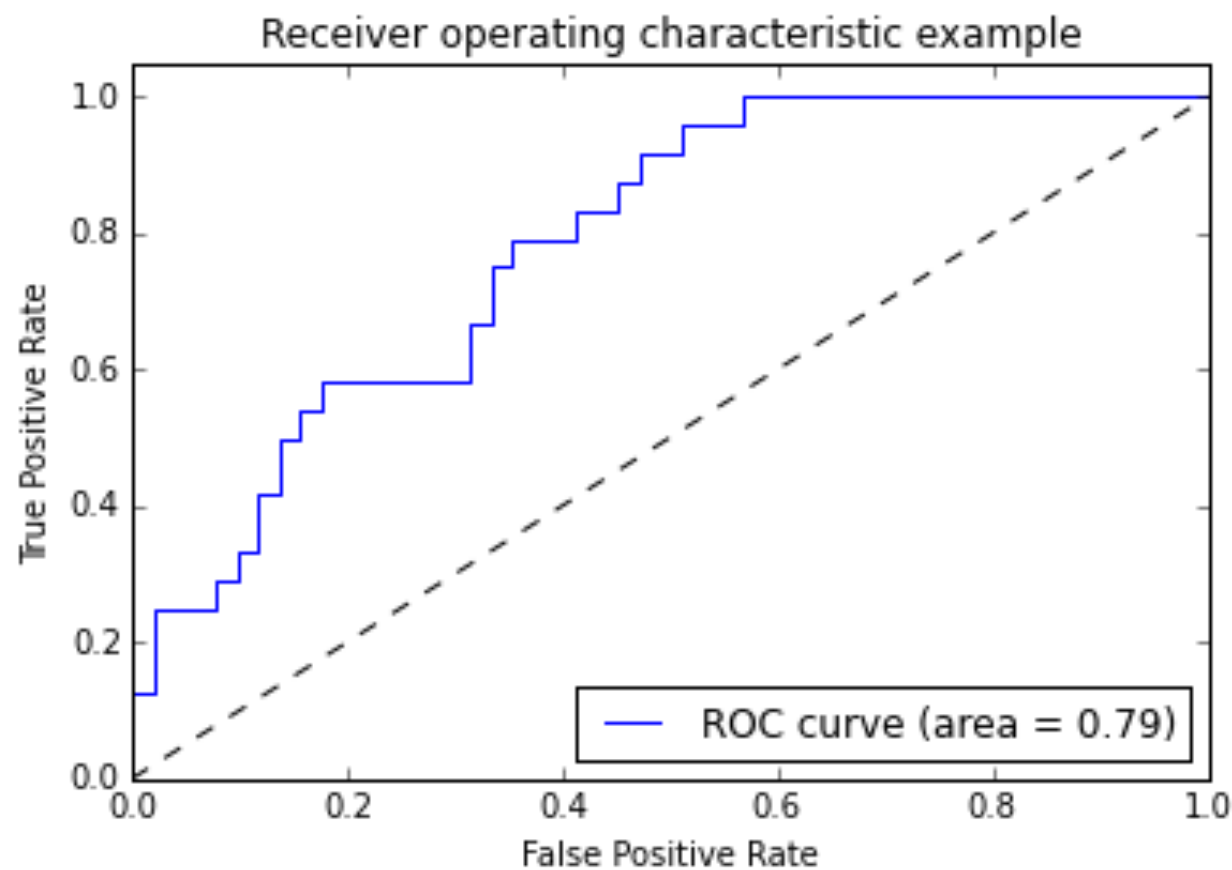
```python
#!python
import nflgame

games = nflgame.games(2013, week=1)
players = nflgame.combine_game_stats(games)
for p in players.rushing().sort('rushing_yds').limit(5):
    msg = '%s %d carries for %d yards and %d TDs'
    print msg % (p, p.rushing_att, p.rushing_yds, p.rushing_tds)
```
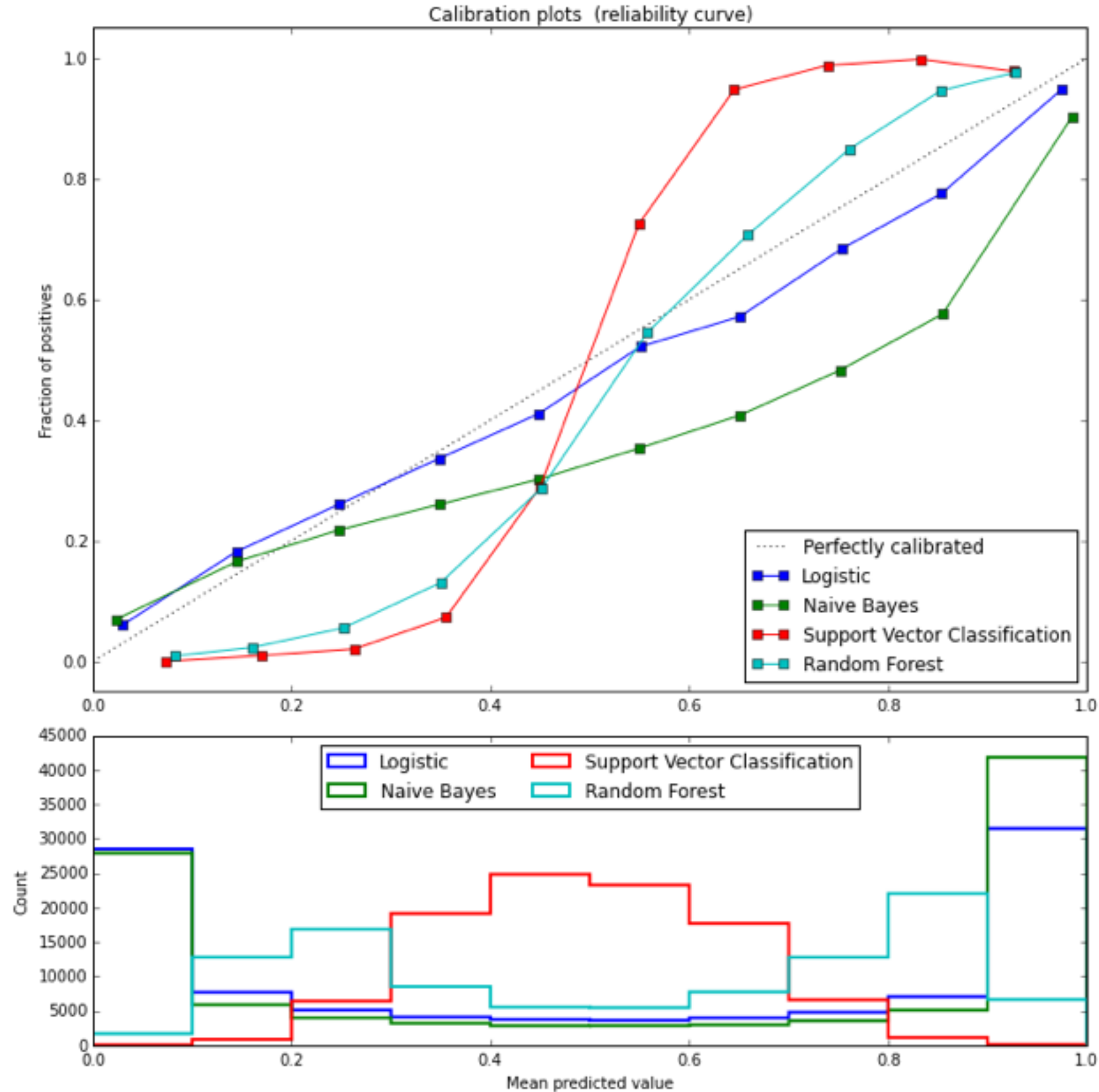
And the output is:

```
L.McCoy 31 carries for 184 yards and 1 TDs
T.Pryor 13 carries for 112 yards and 0 TDs
S.Vereen 14 carries for 101 yards and 0 TDs
A.Peterson 18 carries for 93 yards and 2 TDs
R.Bush 21 carries for 90 yards and 0 TDs
```

ROC Curves

Receiver operating characteristic example

Some extension of Receiver operating characteristic to multi-class

- micro-average ROC curve (area = 0.73)
- macro-average ROC curve (area = 0.77)
- ROC curve of class 0 (area = 0.91)
- ROC curve of class 1 (area = 0.60)
- ROC curve of class 2 (area = 0.79)
- ROC curve (area = 0.79)

Calibration Plot

Odds Resources

- http://www.bettingexpert.com/blog/how-to-convert-odds

- http://espn.go.com/nba/lines

- http://www.oddsshark.com/

- http://www.oddsshark.com/sports-betting/betting-against-public

- https://www.sportsbook.ag/sbk/sportsbook4/home.sbk

- http://www.donbest.com/

- http://www.foxsports.com/college-basketball/odds

- http://espn.go.com/espn/feature/story/_/id/12280555/how-billy-walters-became-sports-most-successful-controversial-bettor

Kaggle Competitions

Kaggle Resources

- https://www.kaggle.com/c/march-machine-learning-mania

- https://www.kaggle.com/c/march-machine-learning-mania-2015

- https://www.kaggle.com/c/worldcupconf

- https://www.kaggle.com/c/finding-elo

- https://www.kaggle.com/c/poker-rule-induction

- https://www.kaggle.com/c/ChessRatings2

What about Fantasy?

- Fantasy sports are generally based on lineups, where you select a team based on a salary cap, and individual players have performance-based market values.

- When competing head-to-head, you have to consider the opponents for each player, e.g., a Steelers running back against a Broncos safety.

- Modeling fantasy sports is an optimization problem, where you are picking the "best" player for the position against the "worst" opponent, subject to the salary cap constraints.

- If you dabble in DK or FD, you will be competing against heavily capitalized whales who submit hundreds of lineups with automated scripts that make last-second adjustments.

Fantasy Resources

- https://www.draftkings.com/

- https://www.fanduel.com/

- http://sports.yahoo.com/fantasy/

- http://fsta.org/

- http://www.newyorker.com/magazine/2015/04/13/dream-teams

- http://www.nytimes.com/2016/01/06/magazine/how-the-daily-fantasy-sports-industry-turns-fans-into-suckers.html?_r=0

- http://www.wired.com/2015/10/daily-fantasy-sports-scandal-fanduel-draftkings/

"The journey is the reward."

- Steve Jobs