Predicting NFL Game Results

Dan Matthews - GADAT3

Problem Statement

Null Hypothesis:

It is impossible to predict the outcome of an NFL game.

Alternative Hypothesis:

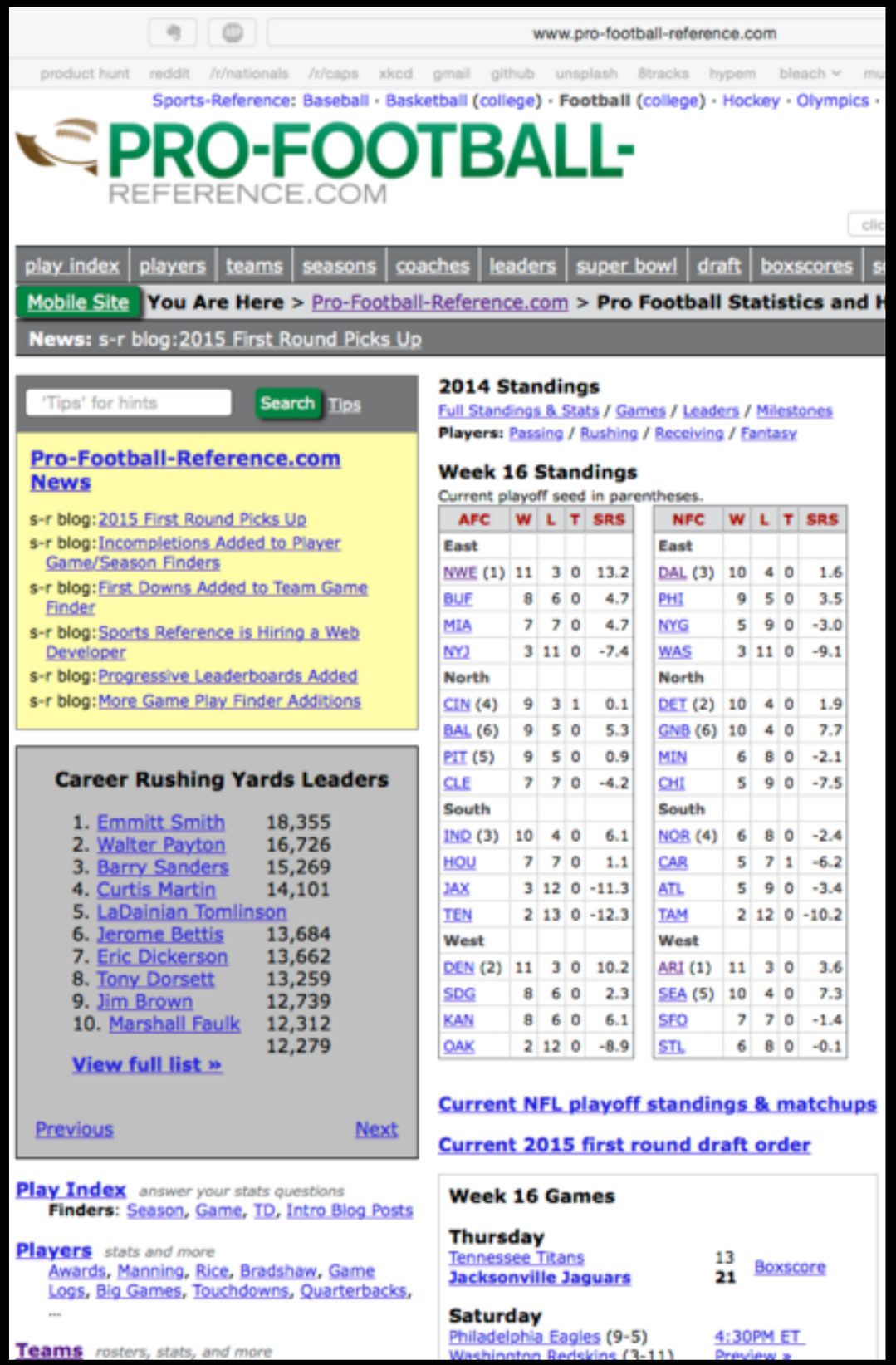
Using game-day statistics, it is possible to predict if a team won a game.

Why this project?

- Dan really likes sports.
- The 2014 Washington Nationals had a ~80% chance of winning a game when they scored four or more runs.
- Curious if certain aspects of a game are more likely to result in a win (IE: first downs lead to wins).

DataSet

Data was obtained from pro-football-reference.com. Downloaded into CSV by year then aggregated into a single CSV.



Raw Game Data

- Win
- Day
- Date
- [link to boxscore]
- Overtime (Y/N)
- Opponent
- Points Scored
- Points Allowed
- First downs gained
- Total yards
- Total passing yards
- Total rushing yards
- Turnovers lost
- First downs given up
- Total yards given up
- Total passing yards given up
- Total rushing yards given up
- Turnovers gained by defense
- Offensive rank
- Defensive rank
- Special Teams Rank

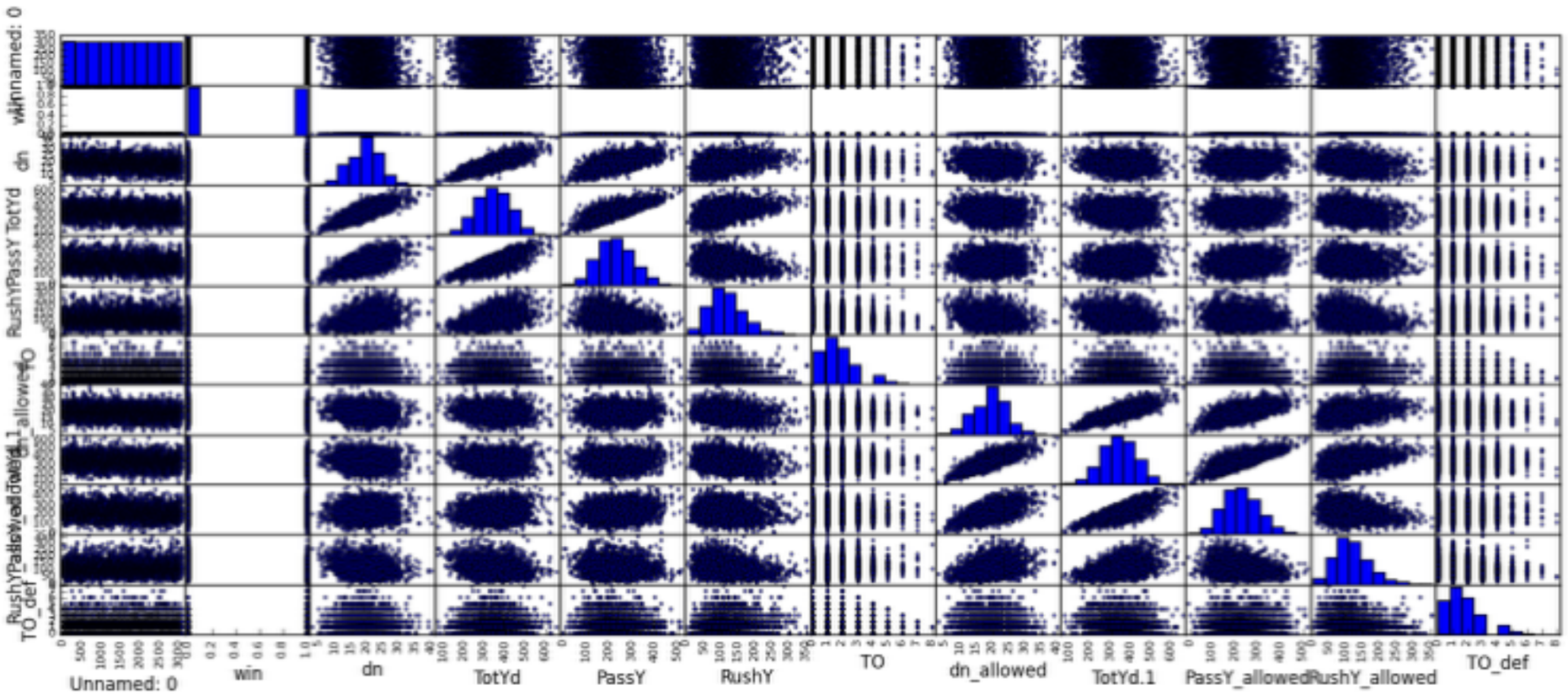
Used Game Data

- **Win (output variable)**
- ~~Day~~
- ~~Date~~
- ~~[link to boxscore]~~
- ~~Overtime (Y/N)~~
- ~~Opponent~~
- ~~Points Scored~~
- ~~Points Allowed~~
- **First downs gained**
- **Total yards**
- **Total passing yards**
- **Total rushing yards**
- **Turnovers lost**
- **First downs given up**
- **Total yards given up**
- **Total passing yards given up**
- **Total rushing yards given up**
- **Turnovers gained by defense**
- ~~Offensive rank~~
- ~~Defensive rank~~
- ~~Special Teams Rank~~

Features

- First downs gained
- Total yards
- Total passing yards
- Total rushing yards
- Turnovers lost
- First downs given up
- Total yards given up
- Total passing yards given up
- Total rushing yards given up
- Turnovers gained by defense

Output: WIN (Y/N) - 1,0

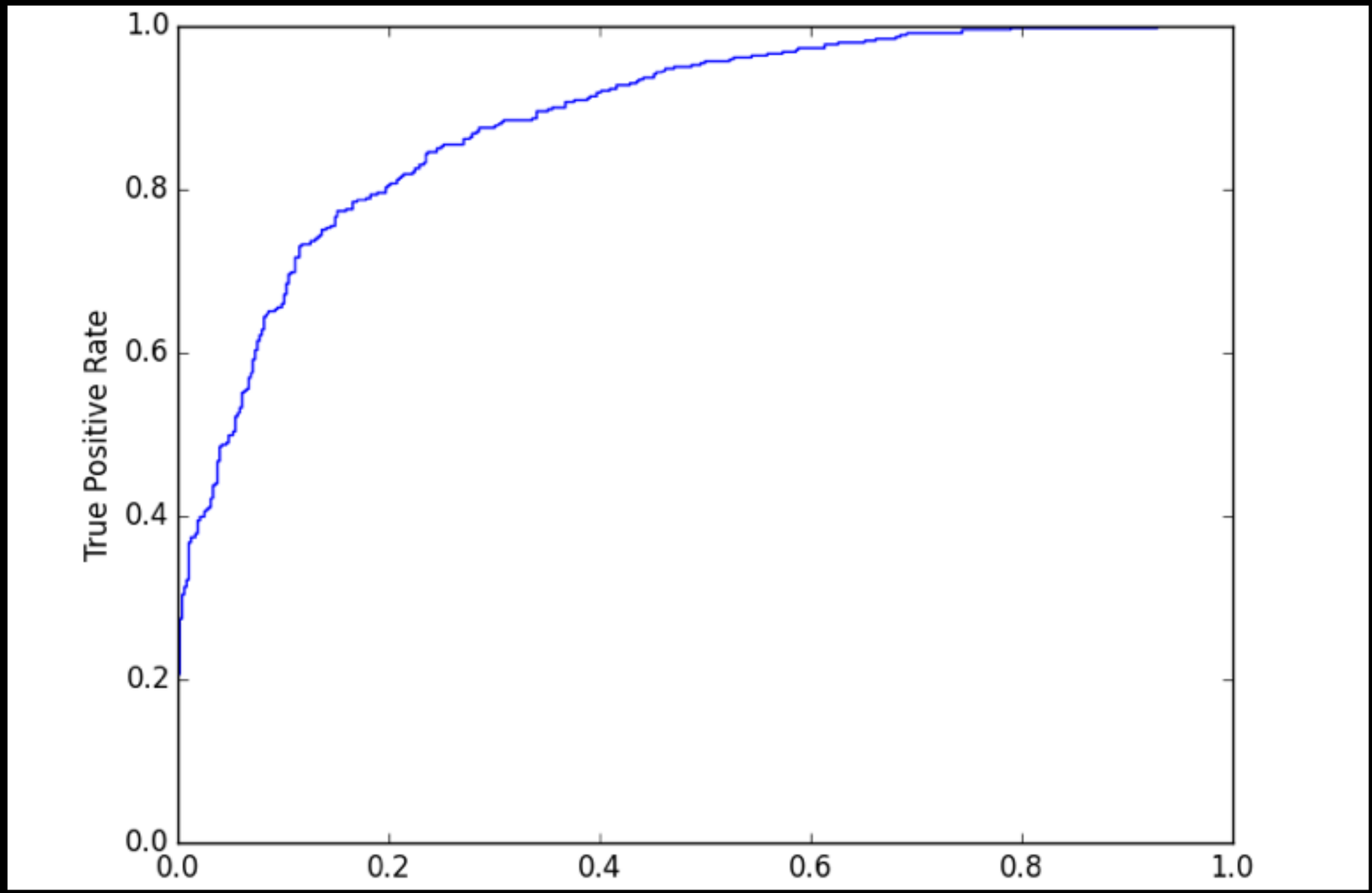


Feature Matrix

Adaboost

- I chose do use a two-response **Adaboost with n_estimators=200** model because individually, none of the factors I chose will be any good at determining a game, but together, they were able to determine the winner of a game with 80% confidence.

- Cross Validation Score: 0.81



Validation

Results

Metric:	dn	TotYd	PassY	RushY	TO	dn allowed	TotYd allowed	PassY allowed	RushY allowed	TO_def
Result:	0.0349	0.1397	0.0067	0.0993	0.2288	0.0532	0.1527	0.0	0.09927	0.2062

Feature values in determining outcome

Results

1. Turnovers given up (TO)
2. Turnovers gained (TO_df)
3. Total yards (TY)
4. Total yards allowed (TY_ald)
5. Rush yards (RushY)
6. Rushing yards allowed (RushY_ald)
7. First downs (dn)
8. First downs allowed (dn_ald)
9. Passing yards allowed (PassY_ald)
10. Passing yards (PassY)

Confusion Matrix

	Predicted Yes	Predicted No	
Actual Yes	349	92	441
Actual No	79	397	476
	428	489	

Challenges

- Determining a project. This is the third project I have attempted.
    - Real estate data: the data was to difficult to clean and I wasn't quite sure what my response variable was supposed to be.
    - NFL vs. Weather: the data was meaningless. Turns out weather has very little impact on an NFL game.
- Adaboost: we never discussed an actual implementation so actually using the model for validation was slightly difficult.

Possible Next Steps

- Expanded data:
  - Completion %
    - Third down conversion rate
- Selling this data to an NFL team so they know how to win ;)
- More data.

Conclusions and Key Learnings

- Turnovers have the largest impact on a game of any of the variables I tested. If a team wins the turnover battle, they have a very good chance of winning the game.

- This makes sense because for each turnover won, that is another shot for your offense to get into the end zone.

Thank you