PREDICTING WIN PERCENTAGE FOR MLB TEAMS

LEAH GAETA

DATA SCIENCE FINAL PROJECT

BACKGROUND

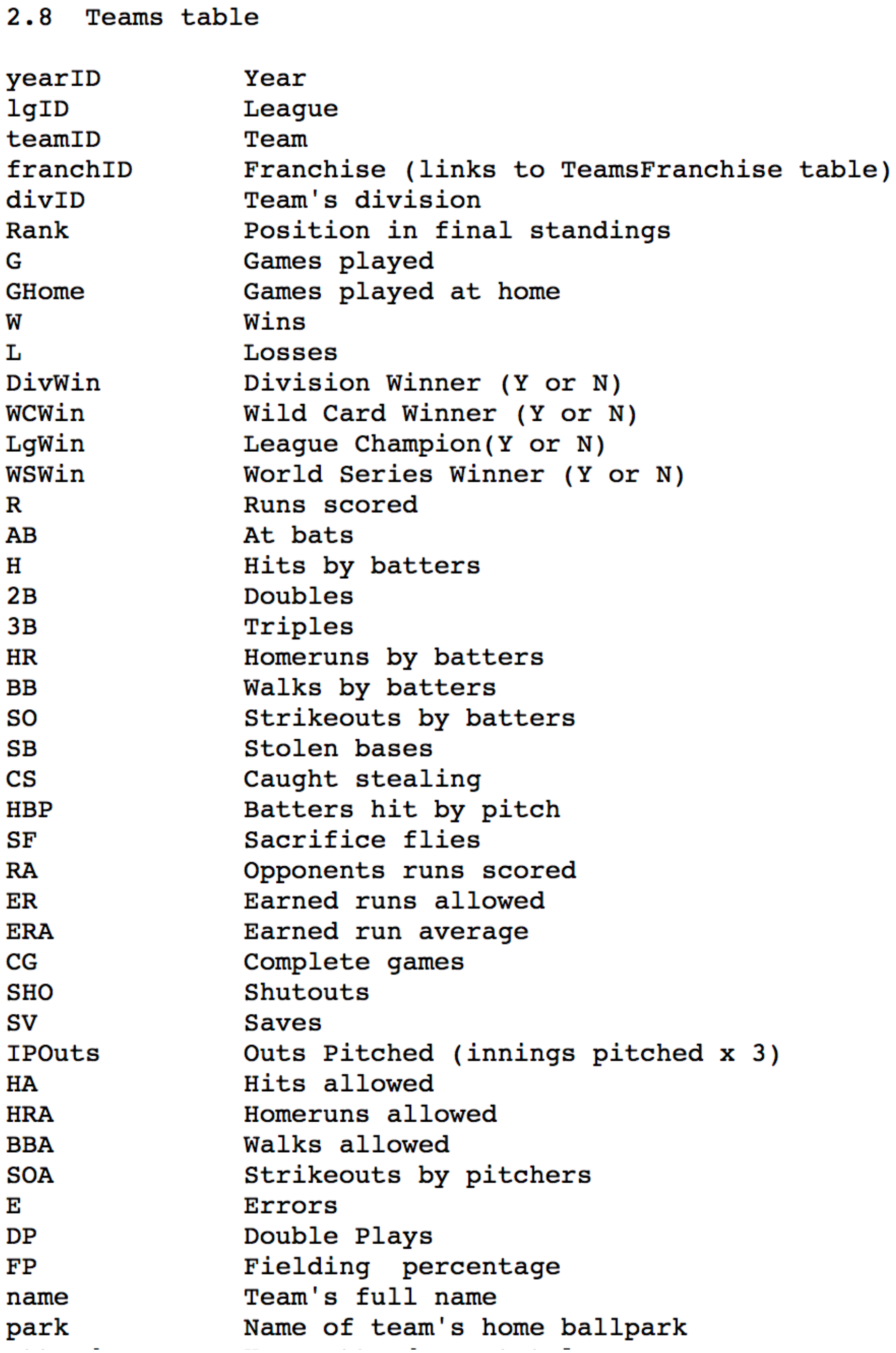
- Bill James – spearhead of the field of sabermetrics
  - Sabermetrics – analysis of baseball using statistics
- James' "Pythagorean Theorem" predicts Expected Win Percentage, using only Runs Scored (RS) and Runs Against (RA)
  - $EXP(W\%) = (RS)^2/[(RS)^2 + (RA)^2]$
  - Exponents have gone through changes but 2 is still a good approximation
- But **HOW** are teams winning?

DATA

- Team statistics for every team from 1960 - 2015 seasons
    - From Sean Lahman, http://seanlahman.com/baseball-archive/statistics
    - The History of Baseball, https://www.kaggle.com/seanlahman/the-history-of-baseball
    - 1960 - now is considered after the “Post-War Era of Baseball”
        - Free agency, divisional play, and the “Steroid Era”
- Statistics include home runs (hr), strikeouts (so), fielding percentage (fp), walks (bb), at bats (ab), wins (w), games (g), etc.

PREDICTING WIN PERCENTAGE FOR MLB TEAMS

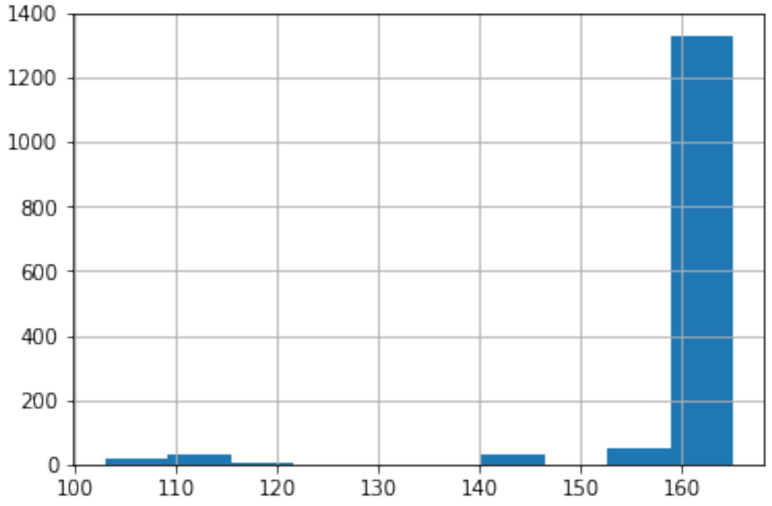
DATA



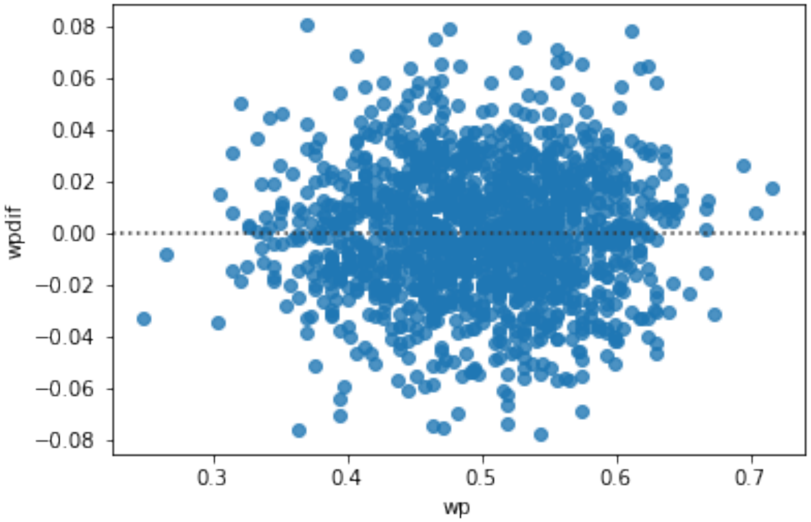
2.8 Teams table

yearID	Year
lgID	League
teamID	Team
franchID	Franchise (links to TeamsFranchise table)
divID	Team's division
Rank	Position in final standings
G	Games played
GHome	Games played at home
W	Wins
L	Losses
DivWin	Division Winner (Y or N)
WCWin	Wild Card Winner (Y or N)
LgWin	League Champion(Y or N)
WSWin	World Series Winner (Y or N)
R	Runs scored
AB	At bats
H	Hits by batters
2B	Doubles
3B	Triples
HR	Homeruns by batters
BB	Walks by batters
SO	Strikeouts by batters
SB	Stolen bases
CS	Caught stealing
HBP	Batters hit by pitch
SF	Sacrifice flies
RA	Opponents runs scored
ER	Earned runs allowed
ERA	Earned run average
CG	Complete games
SHO	Shutouts
SV	Saves
IPOuts	Outs Pitched (innings pitched x 3)
HA	Hits allowed
HRA	Homeruns allowed
BBA	Walks allowed
SOA	Strikeouts by pitchers
E	Errors
DP	Double Plays
FP	Fielding percentage
name	Team's full name
park	Name of team's home ballpark

DATA



- Teams do not play the same number of games
  - Rainouts, tie-breakers, strike-shortened seasons



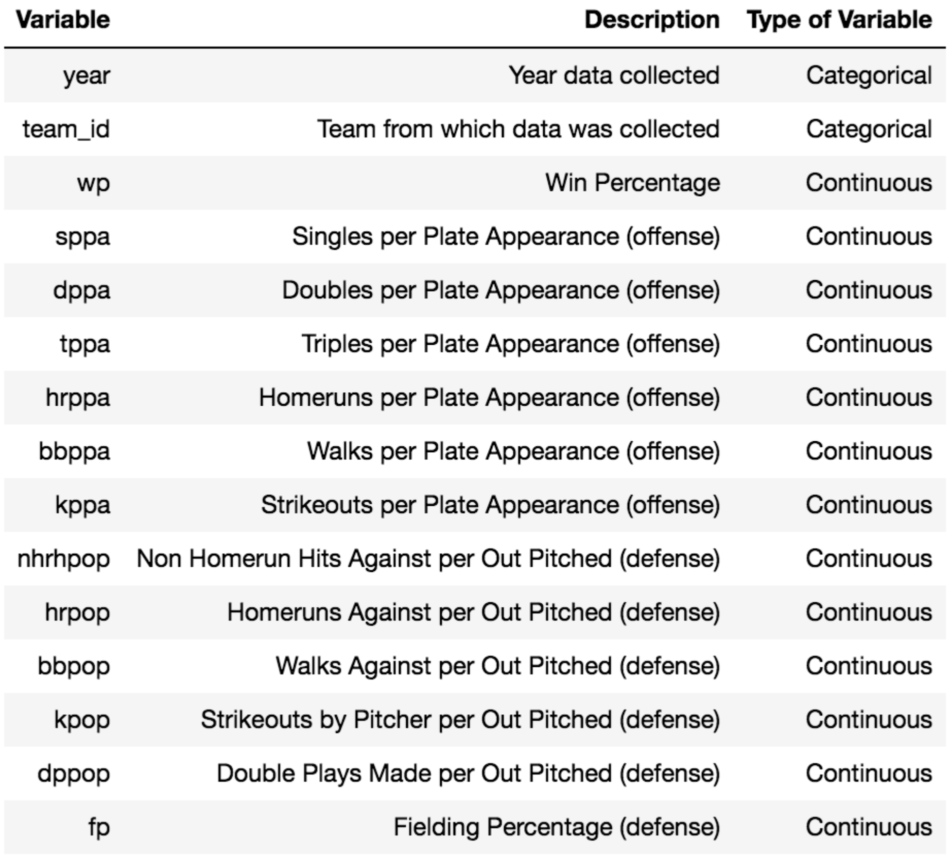
- Added Win Percentage (wp) to the data
  - Win Percentage (wp) = Wins (w) / Games (g)
- James' Pythagorean Theorem on the data

DATA

- Since predicting win percentage, adjusted the offensive & defensive data collected
  - Per Plate Appearance (PA) for offensive data and per Out Pitched (OP) for defensive data
  - For offense, cannot use At Bats (AB) because doesn't take into account walks, sacrifice flies, catcher's interference, etc.
  - Fielding Percentage (fp) is already a percent so it was not adjusted
    - (Putouts + Assists) / Total Chances (Putouts + Assists + Errors)

DATA

- Offense: singles, doubles, triples, home runs, walks, strikeouts
- Defense: non home run hits against, home runs against, walks against, strikeouts by pitcher, double plays, fielding percentage



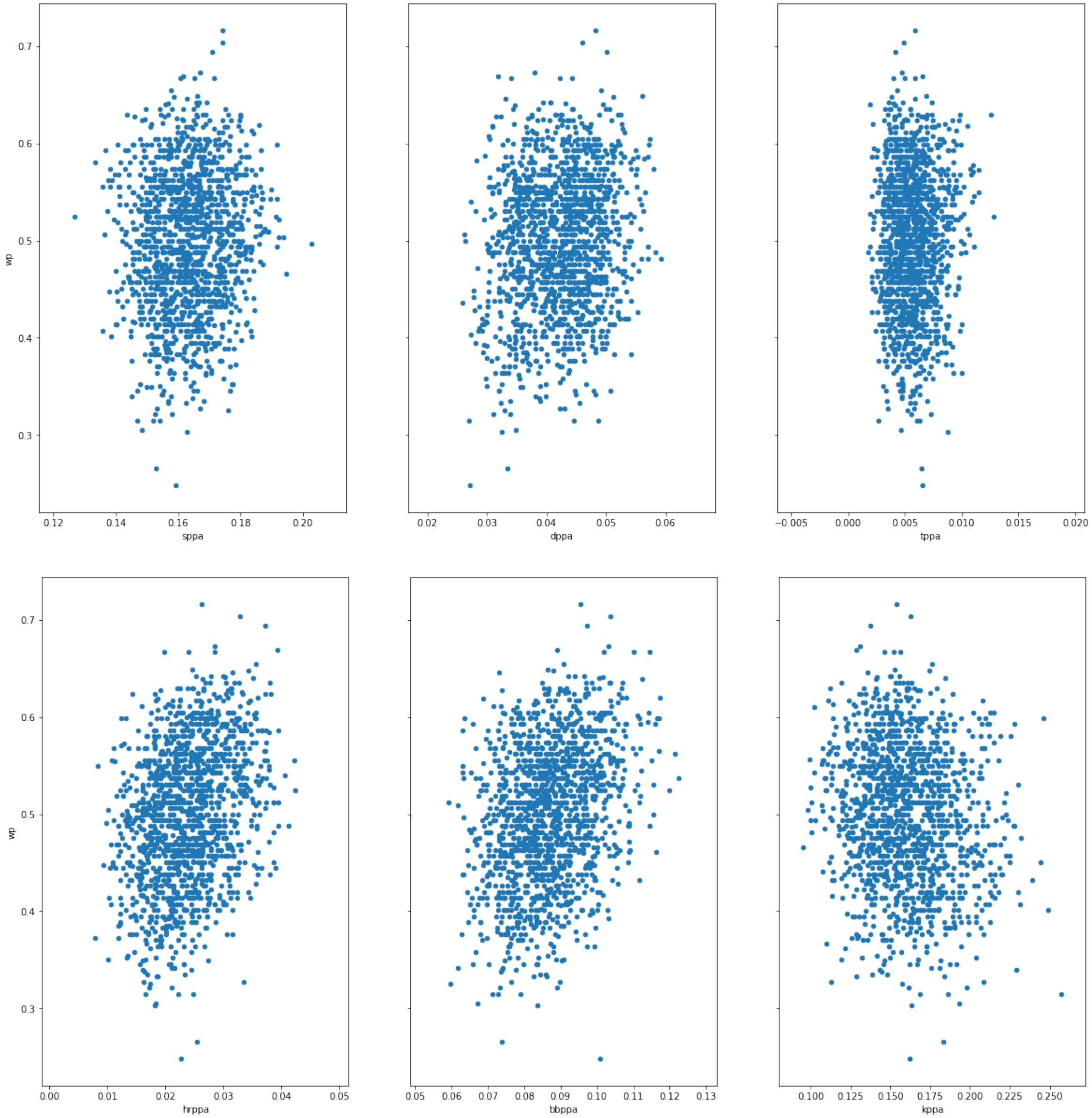
Variable	Description	Type of Variable
year	Year data collected	Categorical
team_id	Team from which data was collected	Categorical
wp	Win Percentage	Continuous
sppa	Singles per Plate Appearance (offense)	Continuous
dppa	Doubles per Plate Appearance (offense)	Continuous
tppa	Triples per Plate Appearance (offense)	Continuous
hrppa	Homeruns per Plate Appearance (offense)	Continuous
bbppa	Walks per Plate Appearance (offense)	Continuous
kppa	Strikeouts per Plate Appearance (offense)	Continuous
nhrhpop	Non Homerun Hits Against per Out Pitched (defense)	Continuous
hrpop	Homeruns Against per Out Pitched (defense)	Continuous
bbpop	Walks Against per Out Pitched (defense)	Continuous
kpop	Strikeouts by Pitcher per Out Pitched (defense)	Continuous
dppop	Double Plays Made per Out Pitched (defense)	Continuous
fp	Fielding Percentage (defense)	Continuous

ANALYSIS

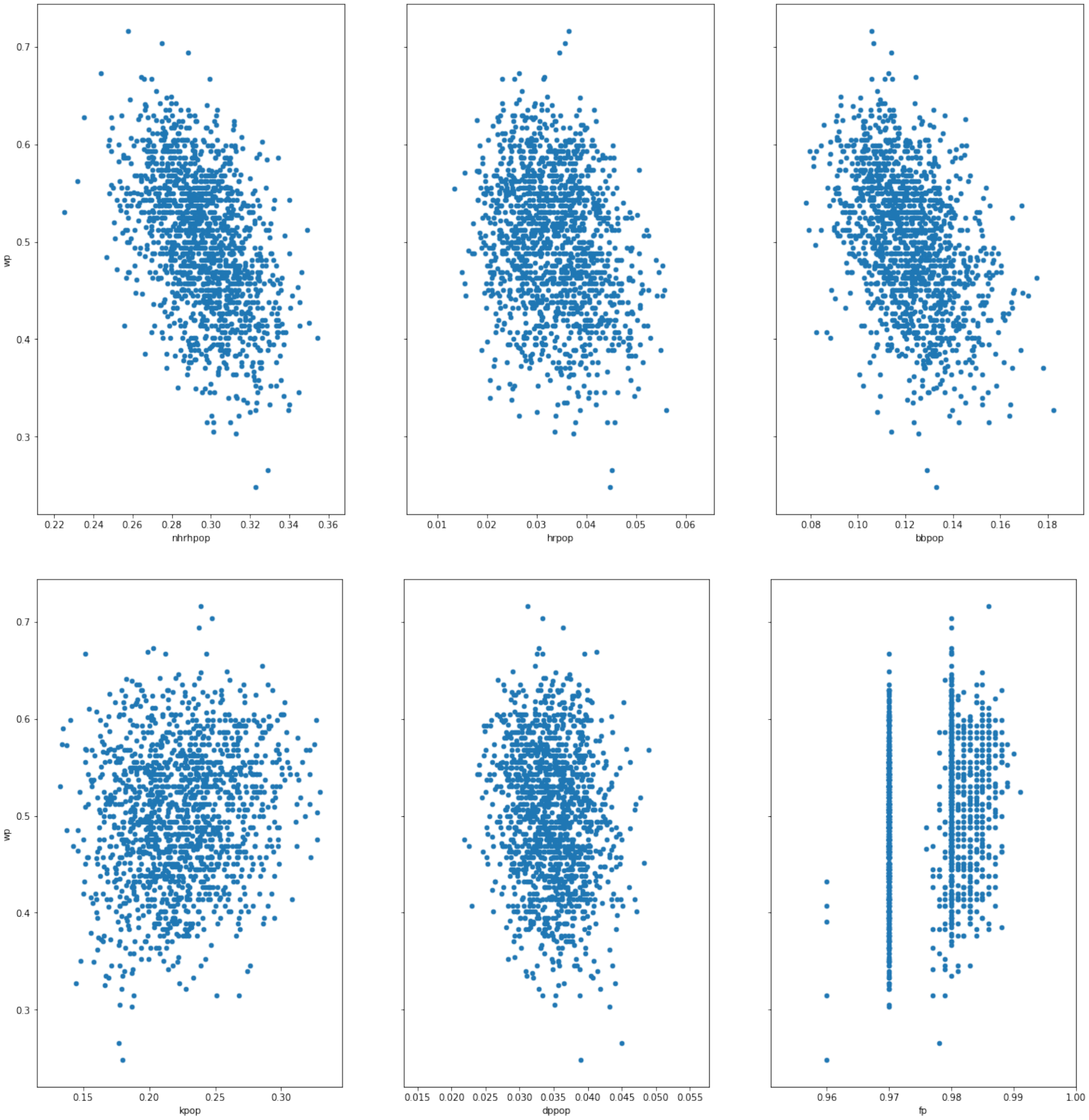
- Using offensive and defensive predictor variables, create a model to explore the association between these variables win percentage and predict a wp outcome
- Null Hypothesis – no association between offensive/defensive predictor variables and win percentage ($p > 0.05$)
- Hypothesis – there is an association between offensive/defensive predictor variables and win percentage ($p < 0.05$)
- Compare to Bill James' "Pythagorean Theorem"

PREDICTING WIN PERCENTAGE FOR MLB TEAMS

ANALYSIS



ANALYSIS

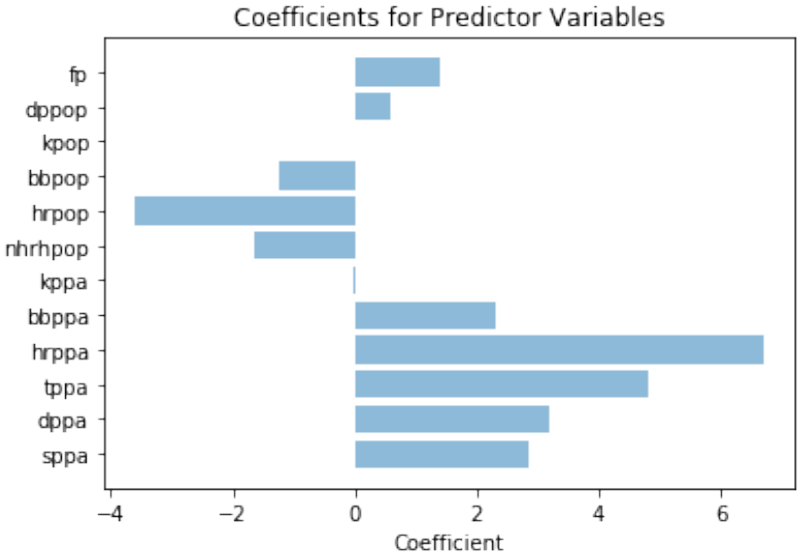

wp
0.7
0.6
0.5
0.4
0.3
0.22
0.24
0.26
0.28
0.30
0.32
0.34
0.36
nhrhpop
0.01
0.02
0.03
0.04
0.05
0.06
hrpop
0.08
0.10
0.12
0.14
0.16
0.18
bbpop
wp
0.7
0.6
0.5
0.4
0.3
0.15
0.20
0.25
0.30
kpop
0.015
0.020
0.025
0.030
0.035
0.040
0.045
0.050
0.055
dppop
0.96
0.97
0.98
0.99
1.00
fp

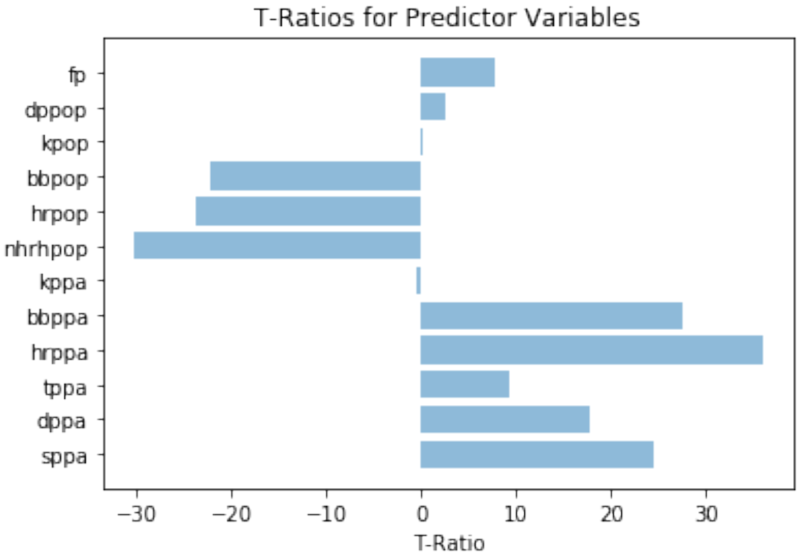
MODELING

- Ordinary Least Squares (OLS) Regression
  - $R^2 = 0.8137$
  - MSE = 0.0009
  - All predictor variables significant ($p < 0.05$) except for Strikeouts per PA (kppa) and Strikeouts by Pitcher per OP (kpop)
  - Model chosen for interpretability

MODELING

- OLS Regression – $R^2 = 0.8137$, MSE = 0.0009



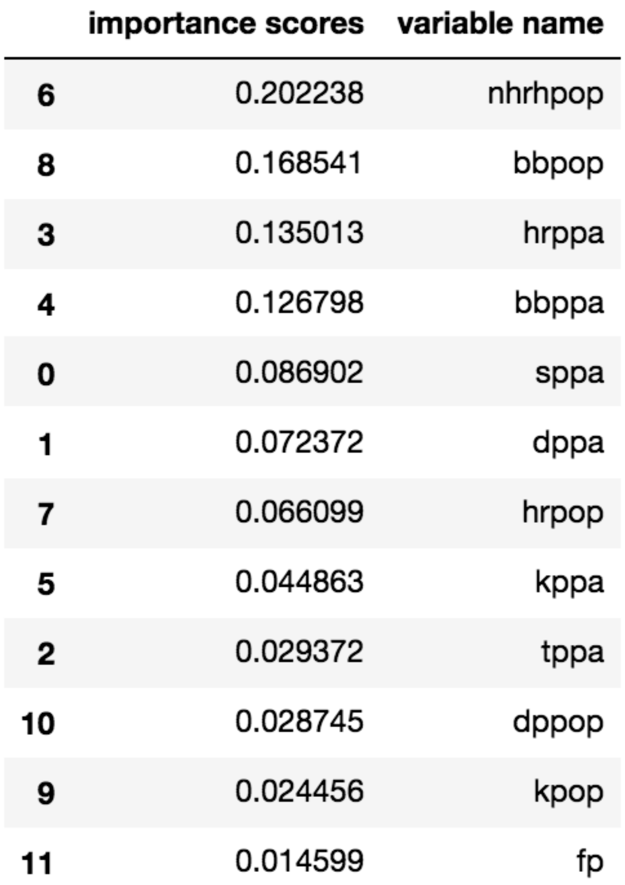


- Biggest Takeaway – home runs & walks really matter while strikeouts don't! For both offensive production & defensive prevention of these happening!

- Only the offensive stats can be compared to each other; only the defensive stats can be compared to each other

MODELING

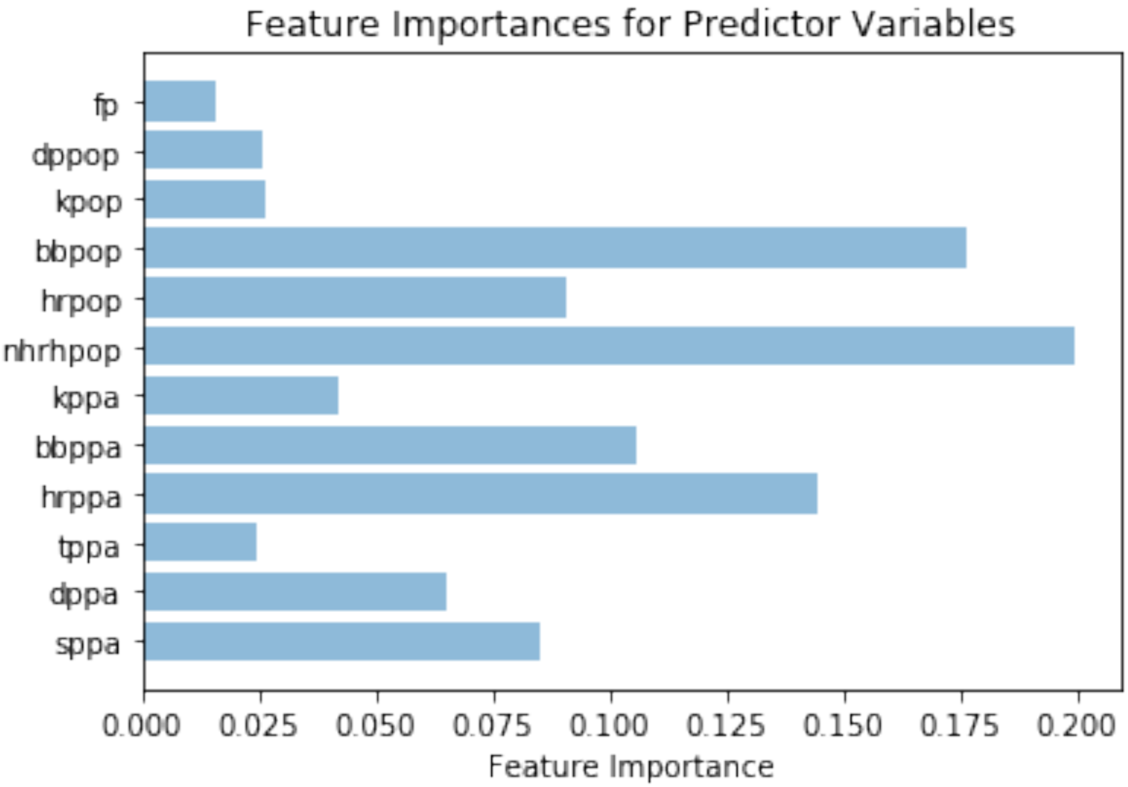
- Random Forest Regression
  - R^2 = 0.8798 (6-Fold Cross Validation)
  - MSE = 0.0019
  - Model chosen for interpretability



	importance scores	variable name
6	0.202238	nhrhpop
8	0.168541	bbpop
3	0.135013	hrppa
4	0.126798	bbppa
0	0.086902	sppa
1	0.072372	dppa
7	0.066099	hrpop
5	0.044863	kppa
2	0.029372	tppa
10	0.028745	dppop
9	0.024456	kpop
11	0.014599	fp

MODELING

- Random Forest Regression – $R^2 = 0.8798$ (6-k CV), MSE = 0.0019



MODELING

- Random Forest Regression – $R^2 = 0.8798$ (6-k CV), MSE = 0.0019
- Again, walks & home runs are key! Offense AND Defense matter!
- Feature importances can all be compared to each other, despite different denominators
- 2 most important features are defensive: Non HR hits per OP (nhrhpop) & Walks per OP (bbpop)
    - Note that Non HR hits per OP is not broken down into Singles, Doubles, or Triples
- Next 2 most important features are offensive: HR per PA (hrppa) & Walks per PA (bbppa)
- Strikeouts per PA or per OP are of less importance

CONCLUSION

- Chose models for interpretability over accuracy
- Home runs and walks are important (offense & defense)
- Strikeouts not as much!
- From random forest regression model, defense is of more importance than offense for predicting win percentage, but certainly both contribute
- Models based on the entire data set; next time set aside data to test models on

CONCLUSION

- Random forest regression model gives important insights as to adjustments a team can make in the middle of the season if the organization is looking for a particular winning percentage to make the postseason
    - Provides the **HOW** that James' Pythagorean Theorem doesn't
- Trade deadline – organization can look into acquiring a pitcher that doesn't give up a lot of hits or walks, or a low WHIP (walks and hits per inning pitched)
- A player that hits a lot of home runs, draws a bunch of walks, and strikes out many times (all per PA) … also worth signing at the trade deadline!
- Though it may intuitively seem obvious, now there are numbers to show that these are the four most important features :)

FUTURE STEPS

- Test the models on data from the 2016 and 2017 seasons.
- Incorporate more data
  - Ballpark factor, payroll, etc.
  - Look into offensive & defensive trends over time ("Steroid Era")
- Compare random forest regression model to Bill James' Pythagorean Theorem

SOURCES

- http://seanlahman.com/baseball-archive/statistics
- https://www.kaggle.com/seanlahman/the-history-of-baseball
- http://sabr.org/research/new-formula-predict-teams-winning-percentage
- http://www.history.com/news/ask-history/what-is-baseballs-modern-era
- https://www.villanovau.com/resources/bls/history-free-agency-pro-sports/#.WkRX9VQ-fOQ
- Introduction to Statistical Learning – http://www-bcf.usc.edu/~gareth/ISL/
- http://www.saedsayad.com/decision_tree_reg.htm
- http://baseballanalysts.com/archives/2010/02/there_are_two_t.php

PREDICTING WIN PERCENTAGE FOR MLB TEAMS

THOUGHTS?