Classification Project

Colin Gallagher

Introduction

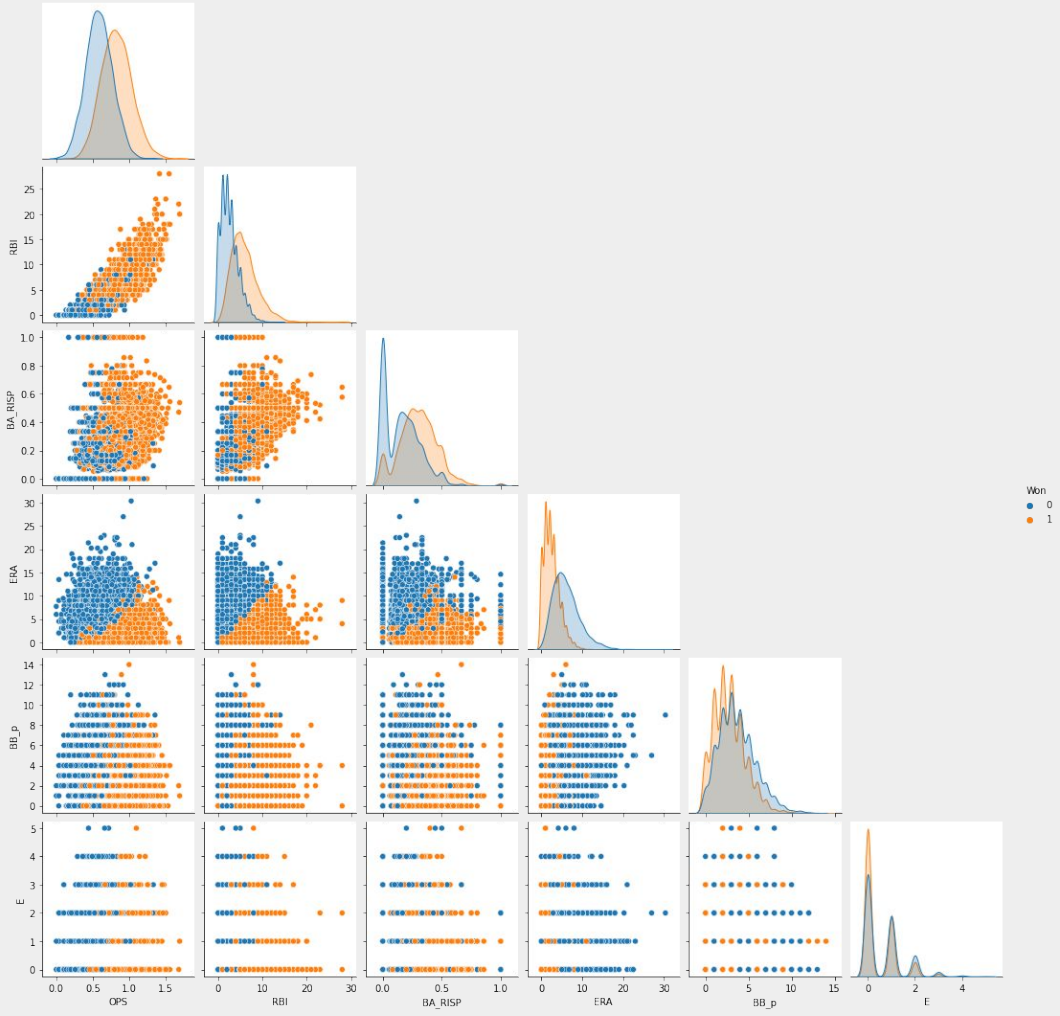
- The goal of this project was to create a classification model
- Predicts a baseball team either getting a win or a loss based on the team performance statistics during the game
- Important problem being solved: **What statistics are most important?**
- Useful for players, managers, owners, fans, gamblers, and sports analysts

Methodology

- Scraped statistics from the baseball-reference.com website using BeautifulSoup and Selenium
- 3 seasons/years (2020-2022) games' box score webpages
- Each games' page had 4 main tables:
  - Batting statistics (one for each team)
  - Pitching statistics (one for each team)
- Each teams' total row from each table was combined to form one row
- Every page would then produce 2 rows of data (one for each team)
- A total of 5,785 games were scraped, making 11,570 total rows of data

Methodology Continued

- Data saved to SQL Database
  - 3 tables (2020, 2021, 2022)
  - 11,570 rows of data total
  - Columns/features/stats: **Won (1/0)**, hits, homeruns, RBIs, ERA, walks, errors, etc.
- Model Creation
  - Seaborn Pair Plot for initial feature selection
  - KNN, **Logistic Regression**, Decision Tree/Random Forest
  - Lots of trial and error with feature selection
  - 3 final candidate feature pools: batting only, pitching only, and combination
    - Two stats dominated: RE24 batting and pitching
    - Broke that into: OPS, RBI, BA (RISP), ERA, Walks, and Errors


RBI
BA_RISP
ERA
BB_p
E
OPS
RBI
BA_RISP
ERA
BB_p
E
Won
0
1

Results (KNN Batting/Pitching)

Features: On Base Plus Slugging, Runs Batted In, Batting Average w/RISP, Earned Run Average, Walks Given Up, Errors

K	21
Accuracy	**95.4%**
Precision	95.3%
Recall	95.6%

Results (Logistic Regression Batting/Pitching)

Features: On Base Plus Slugging, Runs Batted In, Batting Average w/RISP, Earned Run Average, Walks Given Up, Errors

Accuracy	**96.7%**
Precision	96.2%
Recall	97.4%

Results (Logistic Regression Batting/Pitching) cont.

Coefficient Values

Intercept	-0.54
OPS	1.25
RBI	**5.31**
BA (RISP)	0.62
ERA	**-6.98**
BB (pitching)	-0.35
Errors	-1.00

Results (Decision Tree/**Random Forest**)

- Since not as interpretable, was not super interested in picking this one

Accuracy	**96.8%**
Precision	96.2%
Recall	97.2%

OPS	0.15
RBI	**0.26**
BA (RISP)	0.07
ERA	**0.46**
BB (pitching)	0.03
Errors	0.02

Logistic Regression in Practice (Good Team)

Los Angeles Dodgers

Standard Scaled Stats:

[[0.27255026 0.26909954 0.14592825 -0.43859867 -0.32732059 -0.03908101]]

Probability= 98.9%

Logistic Regression in Practice (Bad Team)

Colorado Rockies

Standard Scaled Stats:

[[-0.01189104 -0.0304485 0.03404805 0.28919514 0.07342408 0.10871292]]

Probability= 5.5%

Logistic Regression in Practice (Average Team)

Chicago White Sox

Standard Scaled Stats:

[[-0.08877306 -0.06556793 0.09407451 -0.08336666 0.04546515 0.12513447]]

Probability= 37.8%

Conclusions

- A logistic regression classification model was made to determine which baseball stats are most important when it comes to winning games
- OPS, RBI, BA (RISP), ERA, Walks, and Errors were best features to use when classifying a team’s performance as a win or loss
- RBI and ERA being the best of the best
    - Tough to avoid this “obvious” outcome

Future Work

- Playoff performance only
- Playing field variable
- American League vs National League
- Incorporate weather data
- Was there a better way to handle the shortened season?
- Handling of outliers
- More features scraped

Questions?