Baseball Hall of Fame Futures Betting

By Ken Hoffman & Eryk Wdowiak

ROAD MAP

STEP 1

- Gathered and cleaned data for 2260 baseball players that played more than 10 seasons in the MLB from the Lahman Baseball Database

STEP 2

- EDA in order to fully understand the dataset
- Show how the target variable correlates to the features in the dataset as well as each other

STEP 3

- Determined which classification model best fits the dataset
- Used : feature engineering , feature selection & hyperparameter tuning

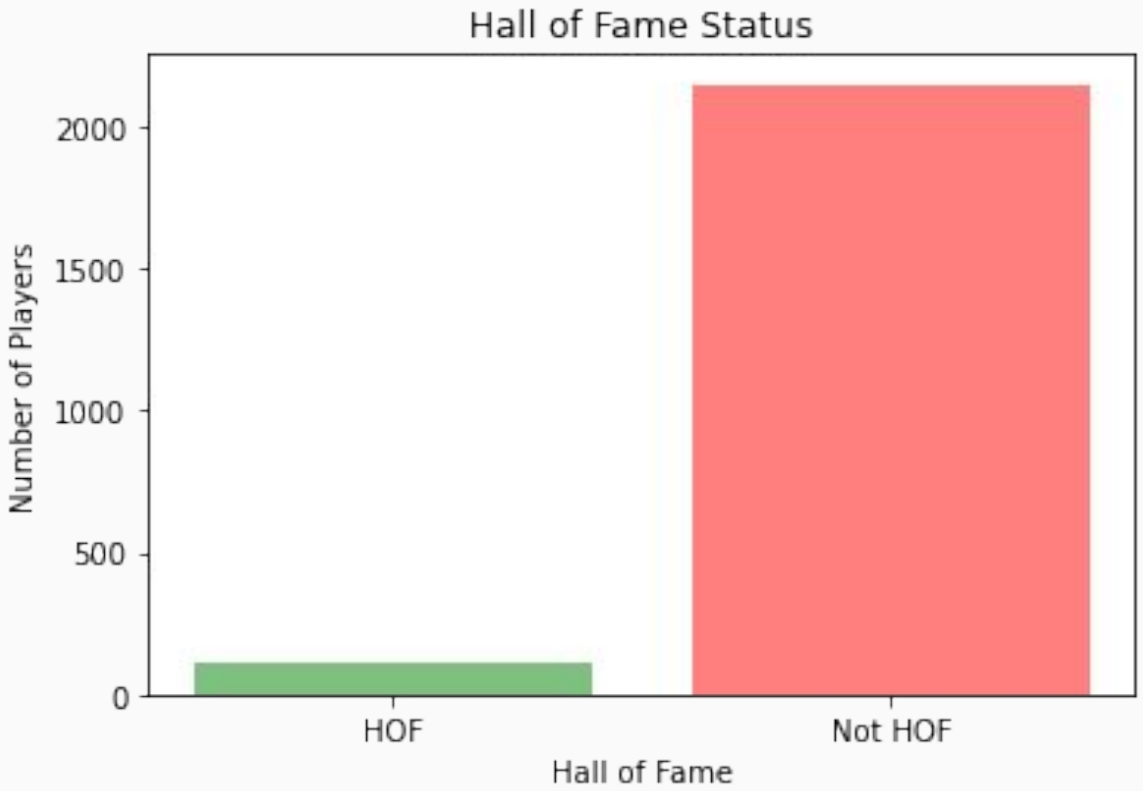
STEP 4

Formulated final analysis & interpret results to apply to real world !

Background and Objective:

- Our goal is to create a model that can accurately predict whether a baseball player will make it into the Hall of Fame
- Dataset contains 2,260 baseball players that have played at least 10 years in the MLB
- Focused on regular season statistics in order to predict Hall of Fame probability

Distribution of Hall of Famers



Predicting Pitchers in the Hall of Fame

Machine Learning Model	F1 Score
Logistic Regression	.7333
Decision Tree	.8571
Random Forest	.9091

- Most important features:
  - Games Played
  - Wins
  - Strikeouts
  - Earned Runs
  - Innings Pitched
  - Years since Retiring

- Hall of Fame Distribution
  - HOF - 35
  - Not HOF - 865

Predicting Catchers in the Hall of Fame

Machine Learning Model	F1 Score
Logistic Regression	.8889
Decision Tree	1.000
Random Forest	1.000

- Most important features:
  - Home Runs
  - Putouts
  - Fielding - Caught Stealing
  - Years since Retiring

- Hall of Fame Distribution
  - HOF - 9
  - Not HOF - 228

Predicting Infielders in the Hall of Fame

Machine Learning Model	F1 Score
Logistic Regression	.8462
Decision Tree	.9565
Random Forest	.9565

- Most important features:
  - Games Played
  - Runs
  - RBIs
  - Double Plays
  - Years since Retiring

- Hall of Fame Distribution
  - HOF - 39
  - Not HOF - 595

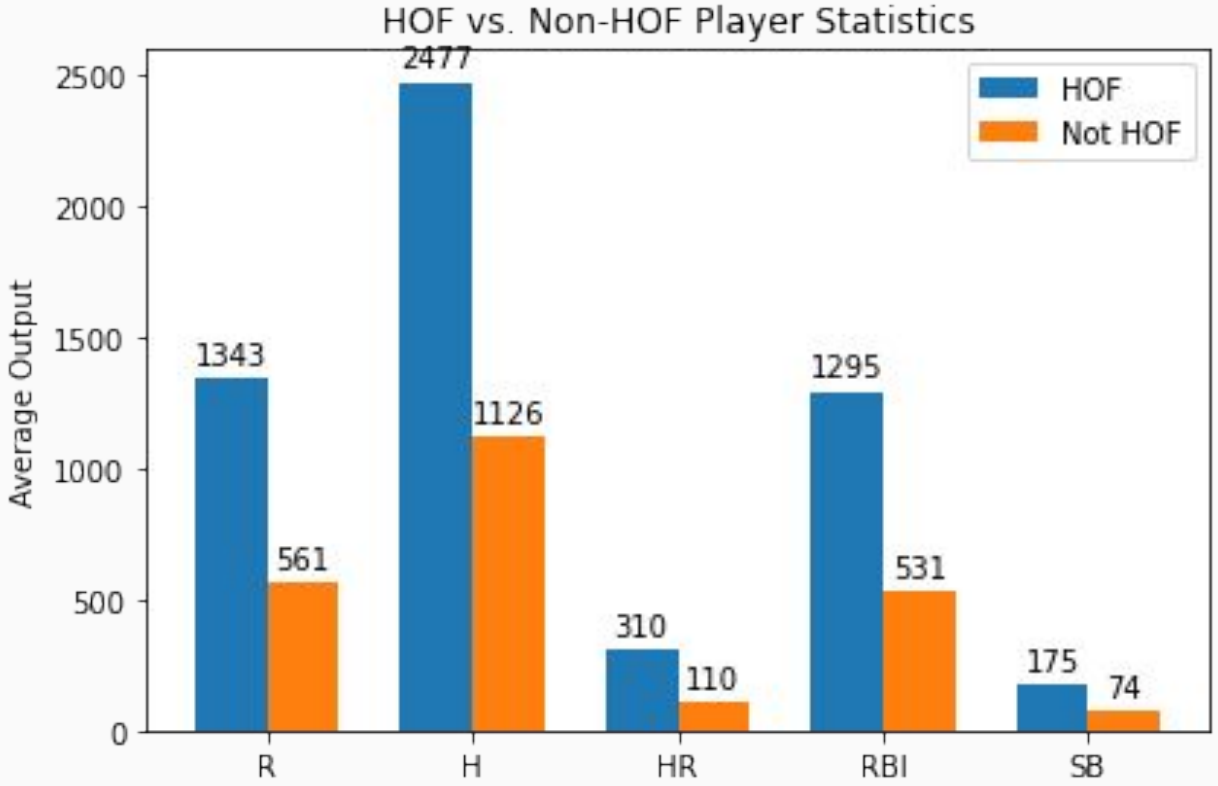
Predicting Outfielders in the Hall of Fame

Machine Learning Model	F1 Score
Logistic Regression	.7586
Decision Tree	.7500
Random Forest	.8571

- Most important features:
  - Runs
  - RBIs
  - Years since Retiring

- Hall of Fame Distribution
  - HOF - 28
  - Not HOF - 461

Comparing HOF to non-HOF batters



Conclusions:

- Certain statistics hold more importance for some positions than they do for others
- Years since retiring is an important measure for every position in determining whether a player is a Hall of Famer
- Fielding statistics hold more weight for Catchers than for other positions

For Further Improvement

- Look at how the following affect a baseball player's chances of making the Hall of Fame:
  - Awards
  - World Series wins
- Account for steroid users and players caught gambling in order to improve model predictions