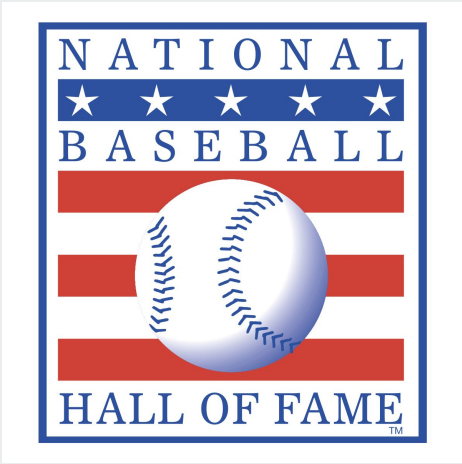
Classifying MLB Hall of Famers

Yarden Sasson & Hunter Harralson



Background

- The National Baseball Hall of Fame is a museum in Cooperstown, NY
- It also represents the pantheon of players elected - the best of the best
  - As of 2018, only 223 Players have been selected
- Players are inducted by Baseball Writers Association of America (BBWAA)
- Any players who played 10+ years can be elected by writers in BBWAA (who each get 10 votes)
- Player must be on 75% of ballots to be elected
  - If less than 5%, no longer eligible
  - Eligible for 10 ballots
  - Example: Barry Bonds & Clemens have been on 6 ballots (~59%)
- Only one unanimous Hall of Famer - Mariano Rivera

Similar Literature

People have made attempts to use neural networks in baseball to:

- Predict player performance
  - Using a recurrent neural network, it reads the data and changes the weights of certain statistics.
  - Forms a matrix based on this data.
  - The output matrices helps the neural network make a prediction, not a sure selection.
  - Trained the network based on statistics of each player per game.
- Predict Pitcher's ERA (Earned Run Average)
  - Using Deep Learning, Linear regression, and gradient descent, they took pitching data and were able to find the ERA for each pitcher throughout the season.

More Similar Literature

- Using Long Term Short Term Memory to predict the next pitch thrown in a game
  - Takes information about the past pitch, current pitcher, and current batter and runs it through a dense neural network.
  - The sigmoid function resulting from the output of the neural network will help it determine the probability of the next pitch.
- Using Reinforcement Learning to decide whether or not to catch a baseball
  - Used the Multilayer Adaptive Heuristic Critic method, to produce a network that will make a probabilistic decision as an event occurs depending on the current risk involved.
  - At the beginning the desired output function is not good enough. As time goes on the system learned to move back instead of moving towards the ball to successfully catch it.

Research Question and Hypothesis

- Research Question:
  - Is it possible to determine Hall of Fame players purely based on their career batting statistics and awards?
- Hypothesis:
  - Accurately classify at least 70% of all batters that the system gets tested on

The Database

- Lahman database
- MLB data updated through 2018 season
- 6 databases used by us:
  - People (name, playerID)
  - Regular season stats
  - Playoff stats
  - Hall of Fame voting - all players voted on
  - Awards
  - All Star Games
- Ended up with 672 Batters (AB > 1000) voted in or out of the HoF

Feature Selection

- 672 batters who were voted in or out of the Hall of Fame
- For each player, we had playerID, First Name, and Last Name
- We then had 33 features based on regular season stats, postseason stats, awards, # of all-star games, # years played
- We chose these features because we believed these were rich enough to distinguish HoFers from losers

Regular Season

[G] [AB] [R] [H] [2B] [3B] [HR] [RBI] [SB] [BB] [IBB]

Postseason

[G] [AB] [R] [H] [2B] [3B] [HR] [RBI] [SB] [BB] [IBB]

Awards

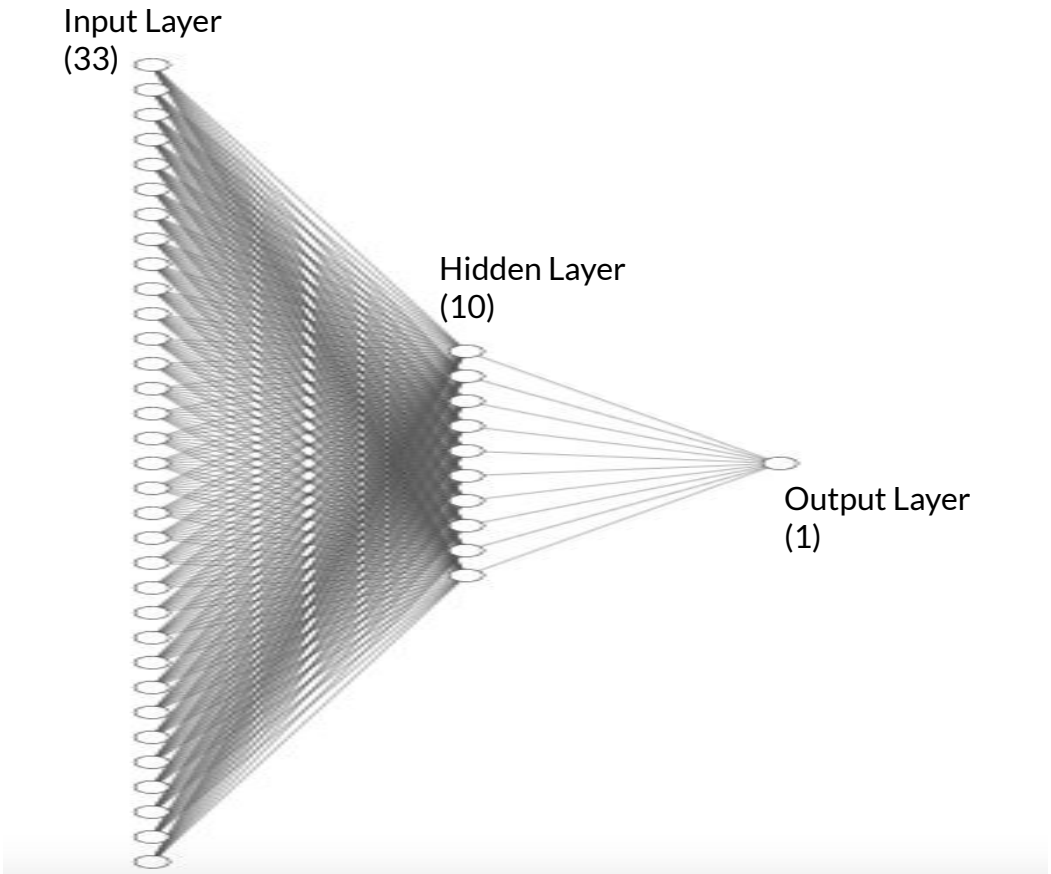
[MVP] [Gold Glove] [ALCS MVP] [NLCS MVP] [WS MVP] [Silver Slugger] [ASG MVP] [Triple Crown] [All-Stars]

Our Approach

- Attempted Models:
  - Linear Associator
    - We found that Hall of Fame is not necessarily linear because there are many variations of statistics and awards when it comes to Hall of Fame players.
- Model We Decided On
  - Multilayer Perceptron that Utilizes Backpropagation

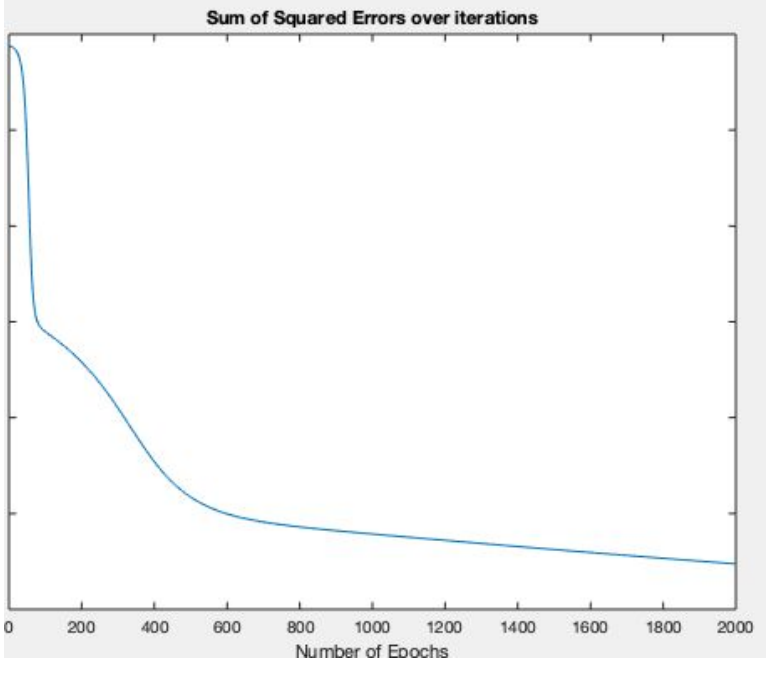
Our Network

- Multilayer Perceptron that utilizes backpropagation
- Input Layer: 33 Nodes
- Hidden Layer: 10 Nodes
    - We played around with this value & this produced the lowest SSE the fastest
- Output Layer: 1 node
- **Binary Classification** - Output gives us a determination of HOF or not



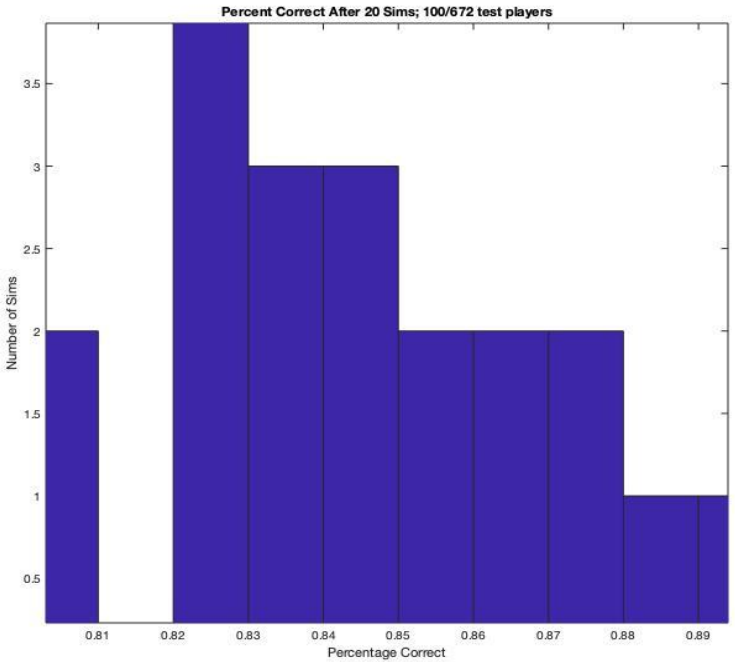
Training the Neural Network

- First run: we trained the model on all players and tested these players
- Testing our training data: ~88 percent accuracy on classification
- Sigmoid activation function
- Begins to converge around 600 epochs



Testing Our Model

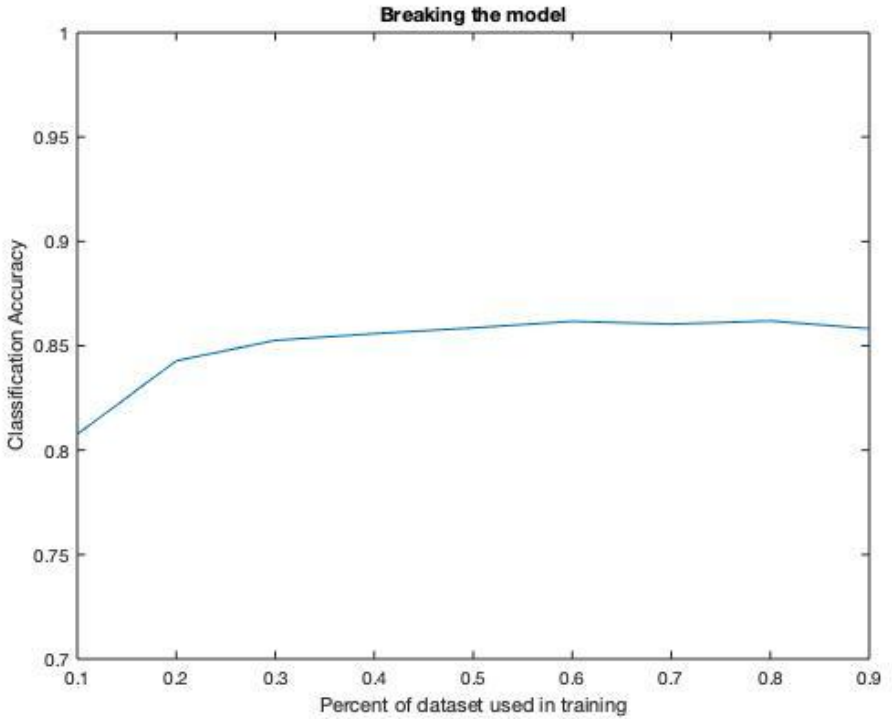
- The first test we did was with 500 players in the training data; 172 in the test
  - In 20 simulations, the accuracy was around 85%
  - Sometimes we got up to 90%
  - distribution of percentages →
- Tended to misclassify players who ARE hall of famers as NOT being hall of famers
  - Actually a good thing. The HOF is very selective and false negatives are better than false positives
  - Also indicative of voting threshold (75%)



Breaking the model

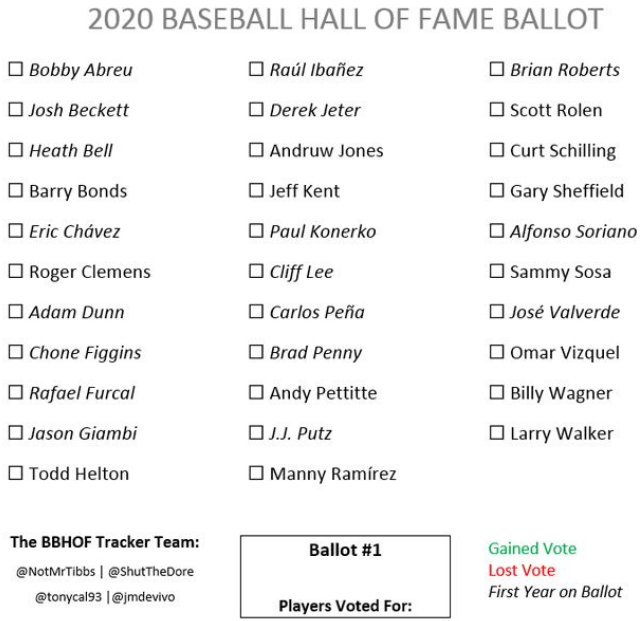
- We incrementally removed players from the training data and placed them in the test data

- Observation: the model is robust, but also plateaus early

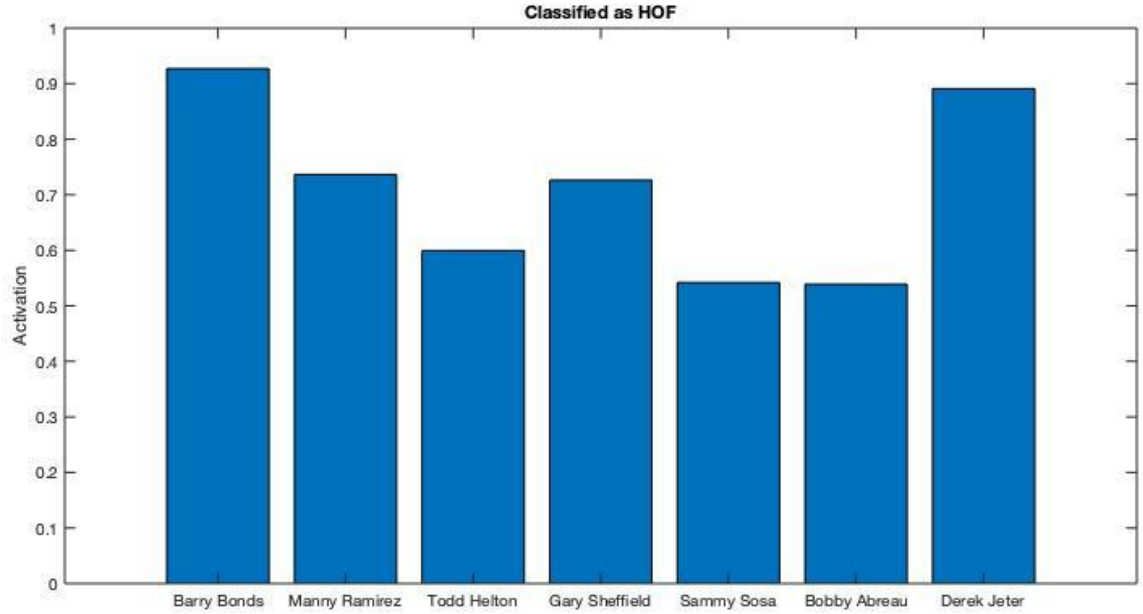


What about this year's Hall of Fame class?

- We tested the 22 batters on the ballot this year.
  - We cannot verify if our model is correct for these new players since the Hall of Fame class has not yet been announced.



Results



- 7/22 players classified as HoF
- Highest Activation: Barry Bonds and Derek Jeter
- Surprised Sosa wasn't higher
- 3 of 4 highest were on Mitchell Report
- Helton played at Coors Field - could be interesting to account for ballpark in model

Future Improvements and Applications

- To train the network to incorporate pitchers and their data.
- Include data from the Mitchell report to help better classify Hall of Famers to the actual standards.
- Using the neural network to help establish a criteria of sorts for Hall of Fame players.
- Determining the contribution that a single season could have to a player's Hall of Fame nomination.

Works Cited

- https://tht.fangraphs.com/using-recurrent-neural-networks-to-predict-player-performance/
- https://github.com/jacobdanovitch/Deep-Neural-Networks-for-Baseball/blob/master/PitcherData.xlsx
- https://cs230.stanford.edu/projects_spring_2018/reports/8290890.pdf
- https://pdfs.semanticscholar.org/fbaf/d78bb02a393bce74ae9df7f4f103d63f5ccf.pdf?_ga=2.34034728.86528715.1575451200-495942217.1575451200

Thank You!

Any Questions?