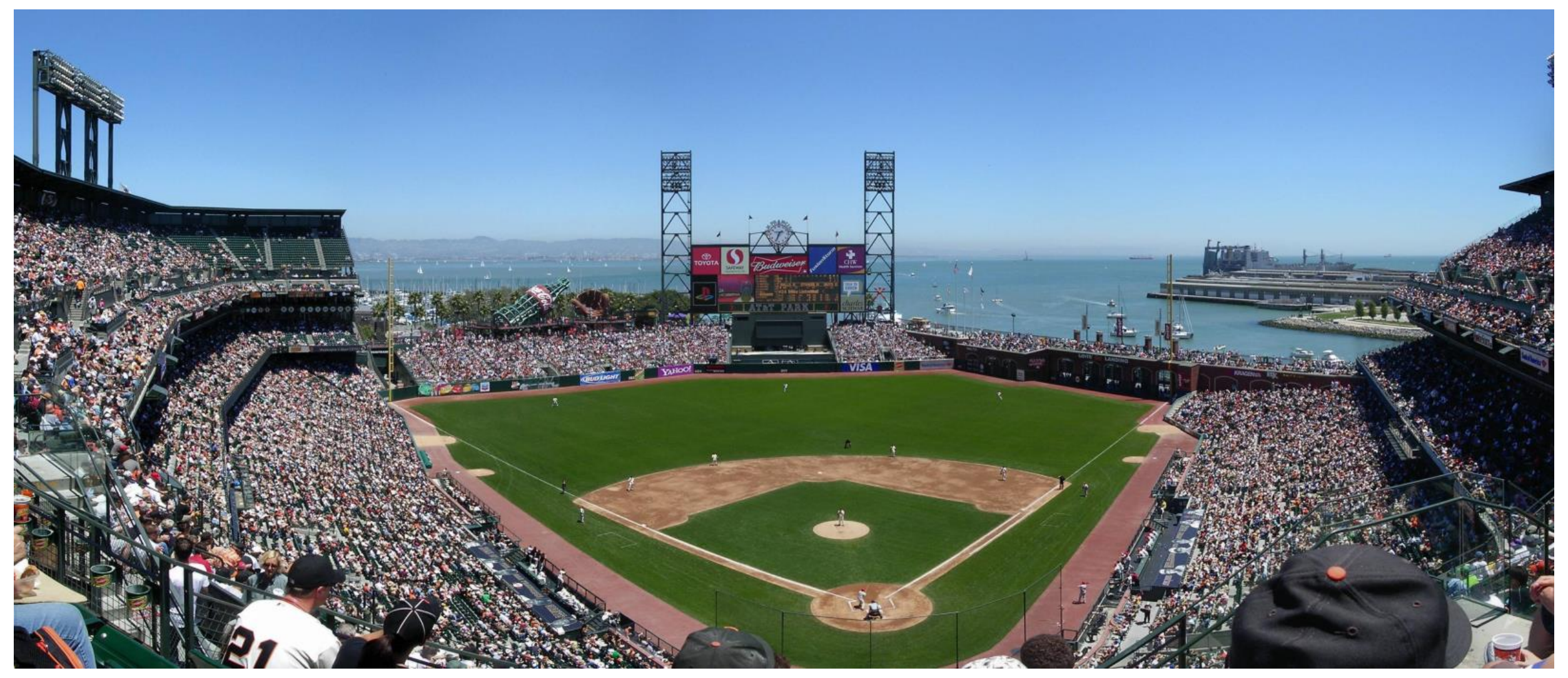
Predicting Baseball Postseason Results from Regular Season Data

Ruiqi Chen[1], Alexander Hobbs[2], and Walter Maier[3]

Department of Aeronautics and Astronautics

[1]rchensix@stanford.edu, [2]ashobbs@stanford.edu, [3]wmaier@stanford.edu

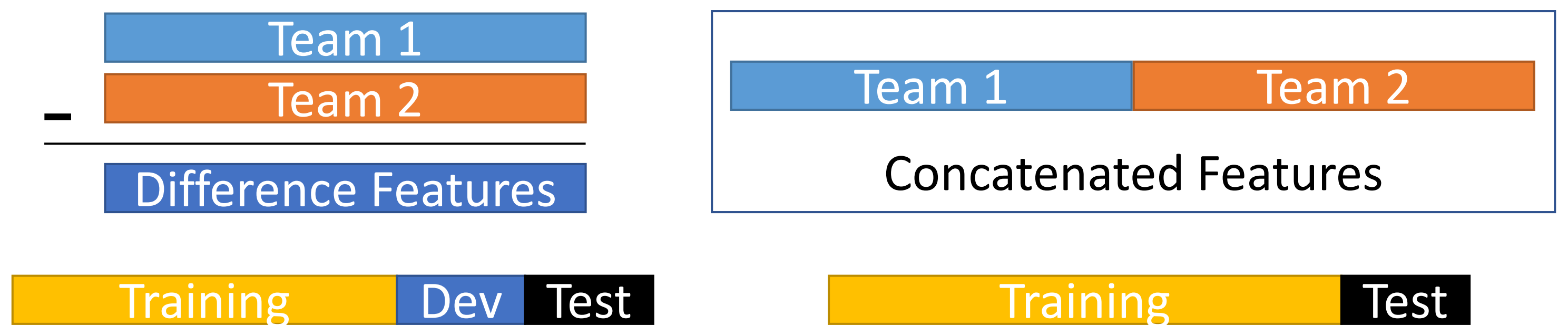
Motivation



Accurately predicting postseason results of sporting events is a multi-million dollar industry [1]. The purpose of this project is to use supervised learning techniques to learn and predict the outcome of baseball playoff series, given regular season player and team statistics.

Data

We use a comprehensive dataset compiled by Sean Lahman [2], which includes complete batting and pitching statistics for Major League Baseball from 1871 to 2016. Both regular season and postseason statistics are included. Results of each playoff series (win or loss) serve as ground truth for training.



60/20/20 dataset used for higher order features model

80/20 dataset used for all other models

Features

Groups of team, batting, and pitching stats were used for a total of 24 features.

Team

Batter park factor (BPF)
Double plays (DP)
Errors (E)
Fielding percentage (FP)
Pitcher park factor (PPF)

Running

Caught stealing (CS)
Runs scored (R)
Stolen bases (SB)

Batting

At bat (AB)
Double (2B)
Hits by batter (H)
Triple (3B)
Home run (HR)
Strikeout (SO)

Pitching

Base on balls (BB)
Earned run (ER)
Earned run average (ERA)
Hits allowed (HA)
Home runs allowed (HRA)
Run average (RA)
Save (SV)
Shutout (SHO)
Strikeout (SOA)
Walks allowed (BBA)

Models

Logistic Classification

$$h_\theta(x^{(i)}) = \frac{1}{1+\exp(-\theta^T x^{(i)})}$$

$$l(\theta) = \sum_{i=1}^{m} y^{(i)} \log\left(h_\theta(x^{(i)})\right) + (1-y^{(i)}) \log\left(1-h(x^{(i)})\right)$$

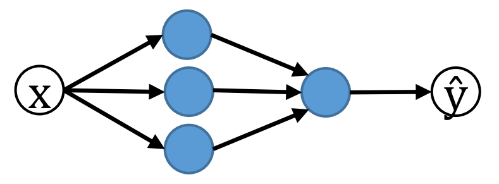
$$\theta \leftarrow \theta - \left(\nabla^2 l(\theta)\right)^{-1} \nabla_\theta l(\theta)$$

Binary logistic classification was implemented using Newton's method. Both subtracted and concatenated features were used. Different ranges of years for training data were explored.

Higher Order Features

Certain groups of features, such as running, batting, and pitching stats, were squared to see effect on accuracy.

Two Layer Neural Network



$$Z = WX + B$$

$$a = \frac{1}{1+\exp(Z)}$$

$$\mathcal{L}(\hat{y}, y) = -(y \log(\hat{y} + (1-y)\log(1-\hat{y}))$$

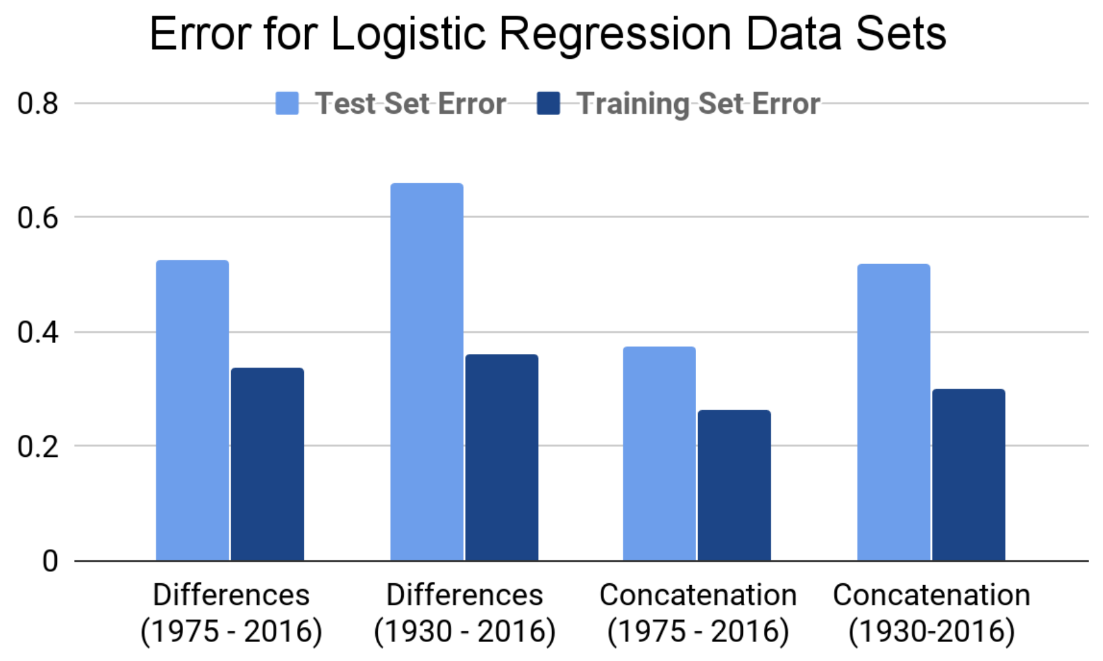
Sigmoid activation neurons were used in a two layer neural network with concatenated features. Experimental parameters include the number of hidden layer neurons, learning rate, and amount of regularization.

Feature Rejection

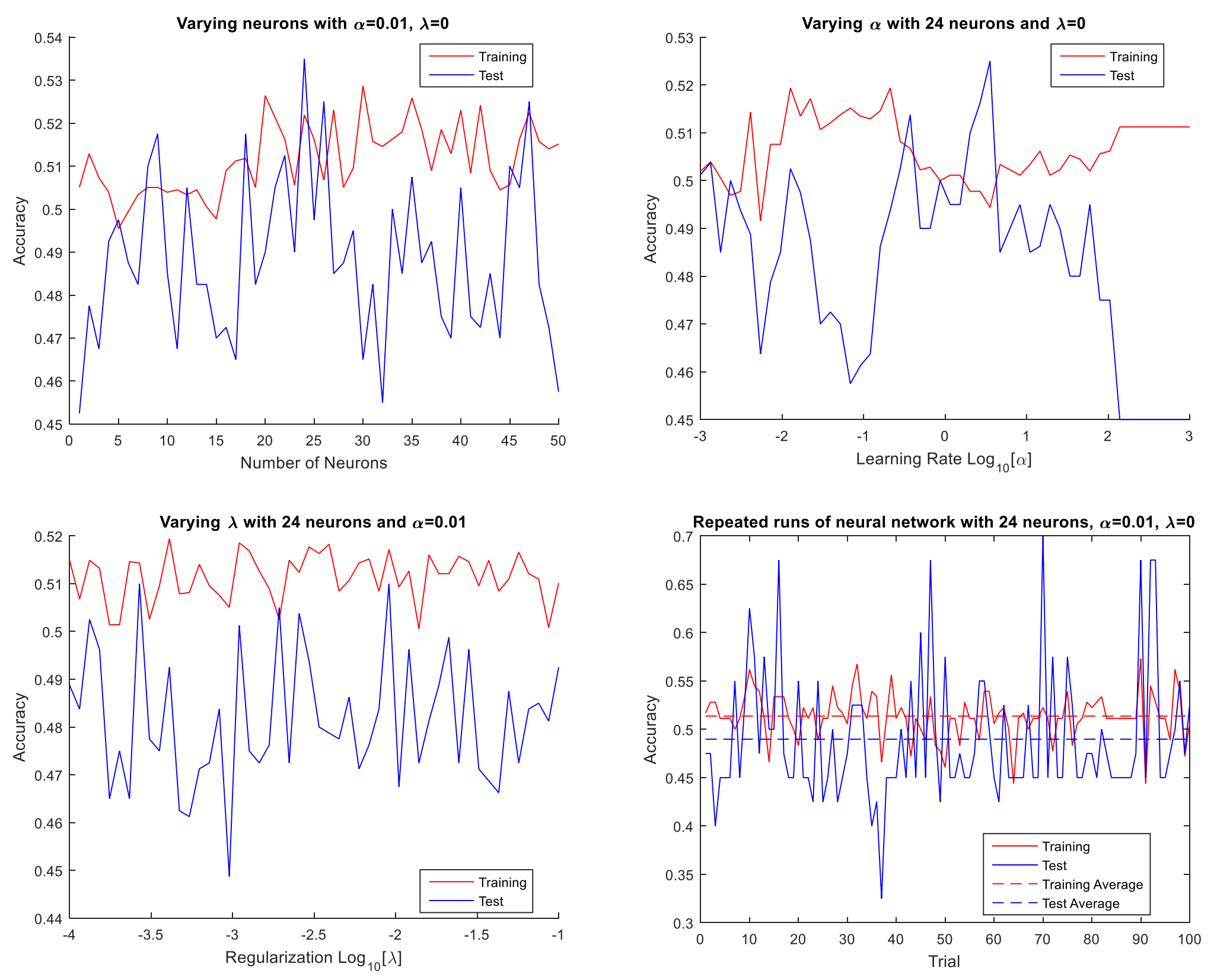
A single feature was removed at a time to see which features caused largest change in accuracy.

Results

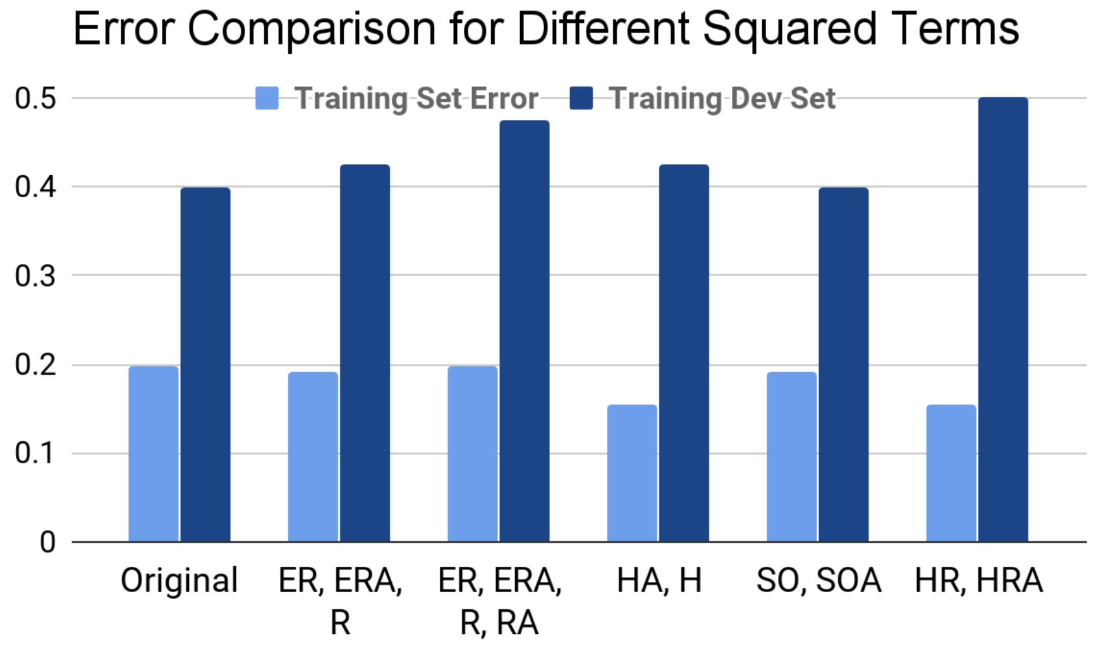
Logistic Classification



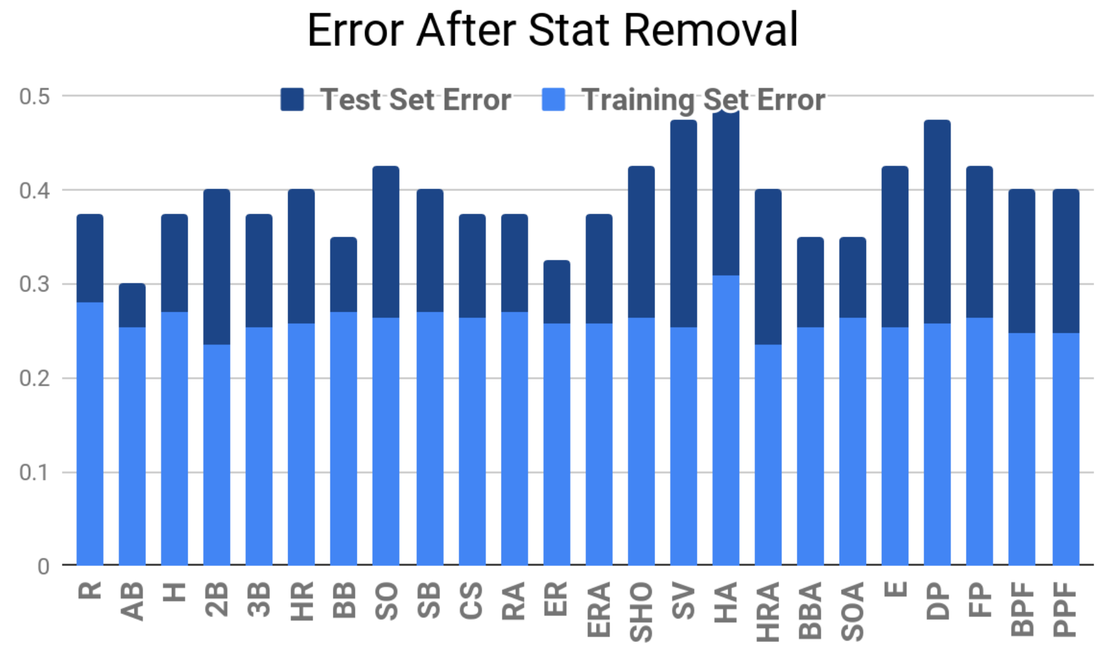
Two Layer Neural Network



Higher Order Features



Feature Rejection



Discussion

The logistic model with concatenated data from 1975-2016 performed the best with 26.4% and 37.5% training and test error, respectively. These results are in the ballpark of related models [3][4]. Concatenated features generally performed better than differential features. Adding additional years resulted in higher error, which can be attributed to rule changes and irrelevance of older statistics [5].

Feature rejection revealed the most relevant statistics in predicting series outcomes are hits allowed (HA) and saves (SV), which makes sense as both statistics greatly affect the number of points a defending team gives up.

Higher order features were able to reduce the training error but generally increased training-dev error due to overfitting.

The neural network generally performed poorly even with carefully selected network parameters. We believe this is due to the small amount of data available.

Future

Currently, the results of each playoff series is assumed to be independent. Future models could look at modeling the playoffs as a Bayesian network, where the nodes represent the playoff bracket.

Additional statistics related to the offseason (drafts, trades, player salaries, etc.) as well as the effect of home field advantage may be explored.

A partially observable Markov decision process (POMDP) could be used to model and find an optimal betting policy.

Finally, due to the general lack of data (due to years before 1975 tending to not be representative of modern day baseball), cross-validation may be a worthwhile endeavor.

References

[1] N. Rayman, "No One Won Warren Buffett's $1 Billion Bracket Challenge," *Time Magazine*

[2] S. Lahman, "Lahman Baseball Database," *SeanLahman.com*

[3] M. Painter, S. Hemmati, and B. Beigi, "Beating the Odds: Learning to Bet on Soccer Matches using Historical Data," *CS229 2016 Projects*

[4] K. Bishop, "Data-Driven Insights into Football Match Results," *CS229 2016 Projects*

[5] "Baseball Rule Change Timeline," *Baseball Almanac*