A Predictive Model of Whether a Major League Baseball Player Will be Inducted into the Hall of Fame

Aaron Springer

Abstract

This study seeks to predict if a Major League Baseball player will be inducted into the National Baseball Hall of Fame based upon the player's career accomplishments. Using 187 players that are either currently in the Hall of Fame or have been voted to not be in the Hall of Fame, we use 35 different career accomplishments and statistics to solve the classification problem of whether a player is a Hall of Famer or not. Ten different models from four different model categories were tested, and the best model from each category was used to create a final ensemble model. This final ensemble model proved to effectively predict the Hall of Fame fate for the players placed in the testing set and can therefore be used to evaluate the quality of a player's Hall of Fame candidacy that is still on the ballot or will be in the future.

Background and Significance

The National Baseball Hall of Fame includes 333 stellar individuals that have greatly contributed to the history and growth of baseball as a player, manager, umpire, or pioneer/executive ("Hall of Famers"). 22,238 players have participated in the history of Major League Baseball since its inception in 1876, but only 263 players have been inducted into the Hall of Fame ("Major League Baseball & Major League"). Clearly, being a Hall of Famer is an incredible honor that puts a player in the top 1% of players all-time.

Players are generally inducted into the Hall of Fame via a ballot voting process whereby they must receive a substantial majority of votes from a large pool of experienced baseball experts. Starting 5 years after their retirement, potentially prominent players are placed on the ballot and to be inducted must receive at least 75% of the votes. Players have 10 years of being on the ballot to reach this 75% threshold to be inducted and are removed from the ballot once the 10 years are up or if they ever receive less than 5% of the votes. The voters are members of the Baseball Writers' Association of America, or the BBWAA, a professional organization of experienced baseball journalists ("BBWAA Election Rules"). In addition to the ballot voting process, previously denied players can still be inducted via the Era Committees, which consist of a smaller group of 16 baseball experts that review the Hall of Fame qualifications of players that were potentially egregiously rejected by the ballot voting process ("Era Committees").

The goal of this study is to use the career accomplishments of a baseball player to determine if they will be inducted into the National Baseball Hall of Fame. The resulting model can be used to predict the Hall of Fame fate of players currently still on the ballot, as well as players that will be eligible for the ballot in the future. This will allow voters to be more informed on the quality of a player's Hall of Fame candidacy prior to submitting their ballots. The results can also be used to determine which factors are most important in determining if a player is a Hall of Famer, and these factors can further be used to evaluate the quality of current players.

Methods

Data Collection

The dataset used consisted of 187 total players, 70 in the Hall of Fame and 117 not in the Hall of Fame. The sport of baseball has 9 positions, 8 of which are batters/fielders, and 1 of which is the pitcher. Whereas the other 8 positions are primarily judged based on their hitting/offensive and fielding/defensive skills, pitchers are generally judged only on their fielding and unique pitching skills, and therefore no pitchers were used in the dataset.

All Hall of Fame players that were not pitchers and retired after 1957 were used. Earlier Hall of Fame players were not used because of the low run-scoring environment that existed in early baseball (Gordon). Additionally, the sport of baseball was racially segregated from 1920 to 1947, and furthermore many of baseball's awards did not exist prior to 1957 ("Negro Leagues", "Major League Baseball Awards"). By only using players that played after 1957, we avoid any type of historical skew and ensure that the players in the dataset all had access to most awards and that they all played in a similar racially integrated run-scoring environment.

The best non-Hall of Famers were sought out and used so that the model could adequately distinguish good from great, and the same rules of no pitchers and no players that played before 1957 applied. The positional JAWS pages on *Baseball Reference* and the positional player rankings on *Baseball Egg* were used to determine which non-Hall of Fame players to include ("Hall of Fame Monitor", "All-Time MLB"). Approximately the same number of non-Hall of Famers as Hall of Famers were used at each position.

Initially, 99 predictors were considered from 6 different categories; how many times a player led the league in a certain offensive category, how many times a player won a certain award, a player's career offensive and defensive statistics, and a player's 162-game average offensive and defensive statistics. The offensive and defensive statistics for every player were found on the player's page on *Baseball Reference*, specifically the 'Standard Batting' and

'Standard Fielding' tables. These pages also gave the info for which awards a player won and which seasons they led the league in a certain offensive category ("Baseball Encyclopedia").

Variable Creation

Viewing the relationships between each predictor and the response showed that some predictors were not good indicators of whether a player would end up in the Hall of Fame. Scatterplots, boxplots, and summary statistics for each predictor, by Hall of Fame status, were created and examined, and those predictors that did not appear to be good Hall of Fame indicators were removed from the dataset. Illogical predictors, such as Hall of Famers having more of a bad event or having less of a good event, were also removed.

Predictors that were highly correlated were removed to avoid multicollinearity. The initial matrix correlation plot with all 99 predictors showed that several highly correlated predictors existed. To fix this, all predictor pairs with correlations of 0.75 or greater were found. The predictor that appeared to be the least indicative of Hall of Fame status was removed, which allowed us to keep the best indicators of Hall of Fame status while still avoiding the use of highly correlated predictors. Figure 2 in the appendix shows the final matrix correlation plot.

One last predictor was also removed because it varied starkly by position. Since our model does not predict players that play different positions in different ways, it is important that no predictors are used that would put certain positions at a severe disadvantage.

The removal of all these predictors brought the final number of predictors used down to 35, 10 of which were for leading the league in a certain offensive category, 5 for winning a certain award, 10 for certain career offensive statistics, 3 for certain career defensive statistics, 7 for certain 162-game average offensive statistics, and none for 162-game average defensive statistics. Table 2 in the appendix describes all variables used in the final dataset.

Analytic Methods

A total of 10 different models from 4 different model categories were tested on the final dataset. The first model category was logistic regression, where a standard and a penalized version were used. The second category was discriminate analysis and consisted of a Linear Discriminate Analysis (LDA), Partial Least Squares Discriminate Analysis (PLSDA), and a Flexible Discriminate Analysis (FDA) model. The third category was a Support Vector Machine (SVM), and a linear, radial, and polynomial kernel were tested. The fourth and final category were neural network models, and a standard and model-averaged version were tested. Each of the 10 different models used the exact same training set and testing set, with 75% of the players placed in the training set and the rest placed in the testing set. The penalized logistic had the alpha and lambda parameters tuned, whereas the number of components was tuned for PLSDA and the degree and number of prunes were tuned for FDA. The neural network models had the size and decay parameters tuned, and all SVM models had the cost parameter tuned. The radial kernel also had the sigma parameter tuned and the polynomial kernel also had the degree and scale parameters tuned. Each of these models were tuned using 10-fold cross validation.

The best model in each category, as measured by the Area Under the Curve (AUC), was used to create a final ensemble model. Table 3 in the appendix shows the AUC for each of the 10 models tested, as well as the 4 models that were used to create the ensemble model. Since the PLSDA and FDA models had the same AUC, the FDA model was used because it had the higher Kappa value when the desired decision rule of 0.5 was used to predict the testing set.

As Table 3 shows, the 4 models used to create the ensemble model were the penalized logistic regression model, the FDA model, the SVM with a radial kernel, and the model-averaged neural network. Graphs of the tuning process for each model can be seen in Figures 3 through 6 in the appendix. In terms of variable importance, the predictors of *AS*, *Singles*, *R*, *RBI*, *MVP*, *SB.Dif*, *RFGDif*, and *Inn* were in the top 5 for at least one of the 4 models.

The final ensemble model was created by taking the soft predicted probabilities of being in the Hall of Fame from each of the 4 models and computing a simple average. From there,

players whose average was greater than 0.5 were assigned to the Hall of Fame class, and players whose average was less than 0.5 were assigned to the non-Hall of Fame class.

Results

The final ensemble model proved superior to each of the 4 models that were used to create it, as it had the highest AUC, Accuracy, Kappa, Sensitivity, and Specificity. Table 4 in the appendix shows a summary of the Accuracy, Kappa, Sensitivity, and Specificity values of the final ensemble model compared to each of the other 4 models. All these models used the same decision rule of 0.5 to determine the player's class. The ROC curve of the final ensemble model compared to each of the other 4 can be seen below in Figure 1, and the AUC of the final ensemble model compared to the other 4 models can be seen below in Table 1.

Figure 1: Final ensemble model's ROC curve compared to the ROC curves of each of the other 4 models that were used to create the final ensemble model

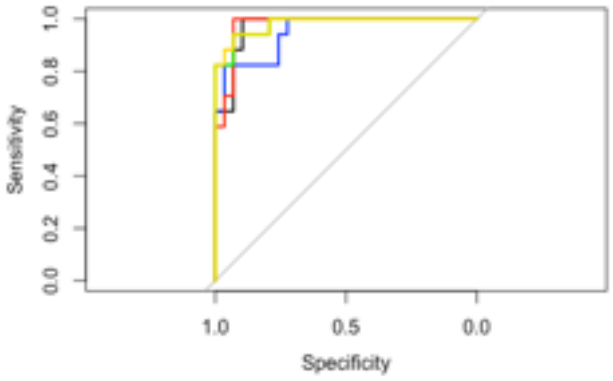

Table 1: AUC for the final ensemble model compared to each of the other 4 models

Model	AUC
Penalized Logistic	0.9716
FDA	0.9493
SVM with Radial Kernel	0.9797
Model-Averaged Neural Network	0.9757
Final Ensemble	0.9817

The final model's confusion matrix can be seen in Table 5 in the appendix. The model correctly predicted 15 of the 17 Hall of Famers and 28 of the 29 non-Hall of Famers in the testing set. The one non-Hall of Famer that was predicted as a Hall of Famer was Dave Parker. The two Hall of Famers that were predicted as non-Hall of Famers were Alan Trammel and Lou Brock.

Discussion and Conclusion

As evidenced by the ROC curve and the confusion matrix, the final ensemble model is an effective predictor of whether a player will be in the Hall of Fame. Only 3 of the 48 players in the testing set were misclassified, and only a single false positive was recorded. Alan Trammel is a fairly acceptable false negative as it took an Era Committee for him to be inducted, and Dave Parker is a fairly acceptable false positive as he was only removed from the ballot after being on it for too many years. Lou Brock, however, is an egregious false negative, as he was inducted with 79.7% of the votes in his first year on the ballot ("Hall of Fame Ballot History").

One limitation of the model is the use of Silver Slugger awards, which players that played before 1980 did not have access to ("Major League Baseball Awards"). By using this award as a predictor, earlier players were less likely to be determined as Hall of Famers. Not using Silver Sluggers as a predictor could solve this issue. Another limitation are the positional impacts on different predictors. While no predictors that obviously favored certain positions were used, some predictors were used that still slightly favored certain positions. To improve this, data from more players could be obtained and separate models could be run for each position.

Overall, this model can effectively be used to predict if players belong in the Hall of Fame. The data for the 35 predictors can be collected to predict the Hall of Fame outcomes for players currently on the ballot and players eligible for the ballot in the future. The results of these predictions can be used to better inform baseball fans and BBWAA voters on which players belong in the Hall of Fame. Furthermore, predictors that are important in determining Hall of Fame status can be used to evaluate players for certain awards, for retention on a team's roster, and for whether a player should be in the starting lineup.

References

"About." *BBWAA*, Baseball Writers' Association of America, 7 Dec. 2021, https://bbwaa.com/about/.

"All-Time MLB Player Rankings." *Baseball Egg - Baseball Rankings and History*, 30 Nov. 2021, https://baseballegg.com/all-time-player-rankings/.

"Baseball Encyclopedia of Major League Players." *Baseball Reference*, Sports Reference, https://www.baseball-reference.com/players/.

"BBWAA Election Rules." *National Baseball Hall of Fame*, https://baseballhall.org/hall-of-famers/rules/bbwaa-rules-for-election.

"Era Committees." *National Baseball Hall of Fame*, https://baseballhall.org/hall-of-famers/rules/eras-committees.

Gordon, David J. "The Rise and Fall of the Deadball Era." *SABR*, Society for American Baseball Research, 27 Nov. 2018, https://sabr.org/journal/article/the-rise-and-fall-of-the-deadball-era/.

"Hall of Fame Ballot History." *Baseball Reference*, Sports Reference, https://www.baseball-reference.com/awards/hall-of-fame-ballot-history.shtml.

"Hall of Fame Monitor Leaders." *Baseball Reference*, Sports Reference, https://www.baseball-reference.com/leaders/hof_monitor.shtml.

"Hall of Famers." *National Baseball Hall of Fame*, https://baseballhall.org/hall-of-famers.

Hastie, Trevor, et al. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed., Springer, 2016.

Kuhn, Max, and Kjell Johnson. *Applied Predictive Modeling*. Springer, 2013.

Liebman, Ronald G. "Schedule Changes Since 1876." *SABR Research Journals Archive*, Society for American Baseball Research, http://research.sabr.org/journals/schedule-changes-since-1876.

"Major League Baseball & Major League Encyclopedia." *Baseball Reference*, Sports Reference, https://www.baseball-reference.com/leagues/index.shtml.

"Major League Baseball Awards." *Baseball Almanac*, https://www.baseball-almanac.com/me_award.shtml.

"MLB All-Star Game History." *Baseball Almanac*, https://www.baseball-almanac.com/asgmenu.shtml.

"Negro Leagues History." *NLBM*, Negro Leagues Baseball Museum, 17 Oct. 2021, https://www.nlbm.com/negro-leagues-history/.

Appendix

Figure 2: Final matrix correlation plot of the final 35 predictors

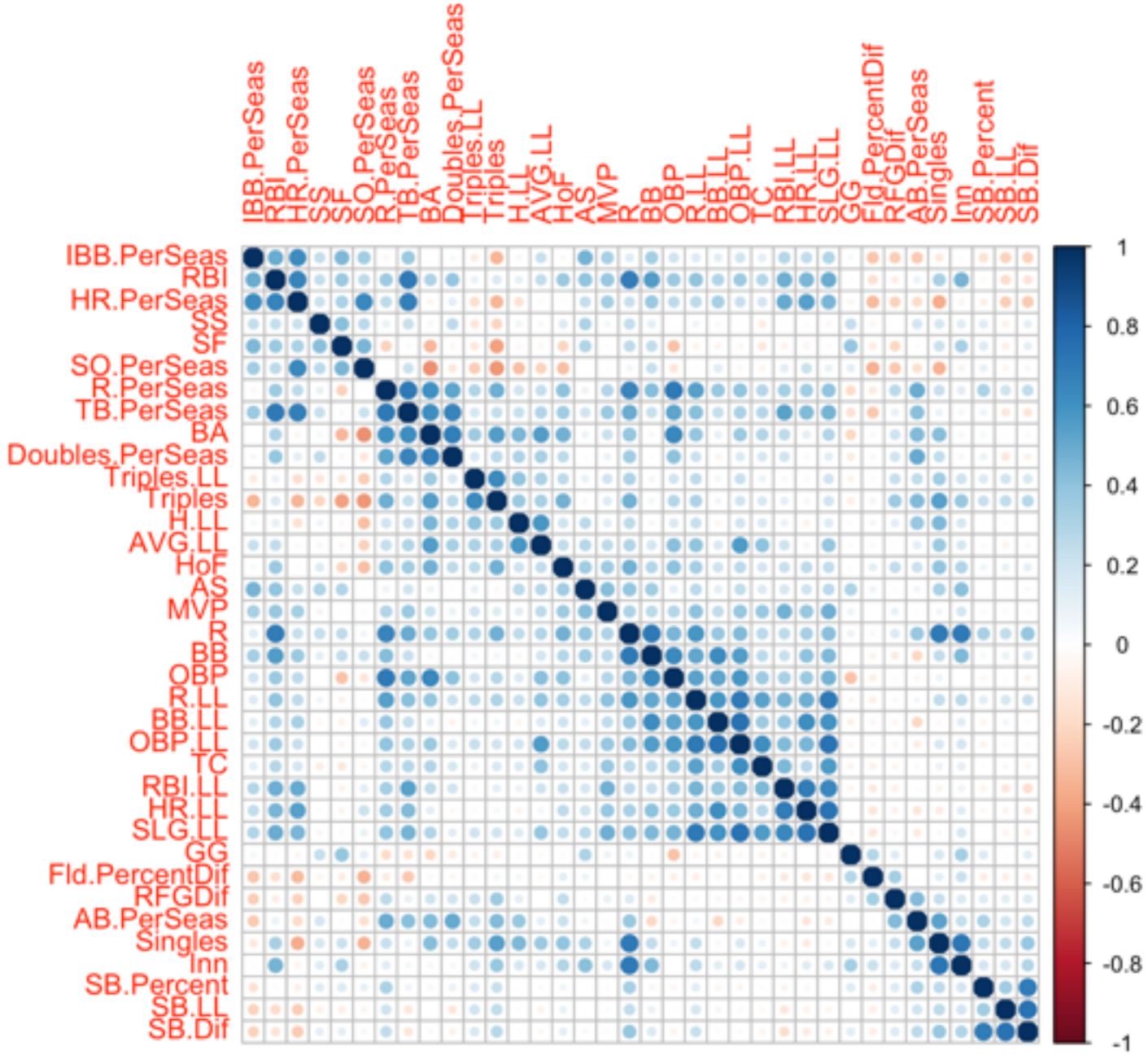

Table 2: Explanation of all variables in the final dataset, including the 35 predictors, the binary Hall of Fame response, and a few descriptive variables

Variable Name	Variable Type	Description
Player	Descriptive – not a predictor	The first and last name of the baseball player
Position	Descriptive – not a predictor	The primary career position of the player, as indicated by the position they played the most games at or achieved the most success at; LF, CF, and RF were grouped into a singular OF position
HoF	Binary Response	Indicator of whether a player is in the Hall of Fame (denoted by a 1) or not in the Hall of Fame (denoted by a 0)
First.Year	Descriptive – not a predictor	The year of the first season the player played in the MLB
Last.Year	Descriptive – not a predictor	The year of the last season the player played in the MLB
R.LL	Discrete	The number of seasons a player led their league in Runs scored
H.LL	Discrete	The number of seasons a player led their league in Hits
Triples.LL	Discrete	The number of seasons a player led their league in Triples
HR.LL	Discrete	The number of seasons a player led their league in Home Runs
RBI.LL	Discrete	The number of seasons a player led their league in RBI (Runs Batted In)
SB.LL	Discrete	The number of seasons a player led their league in Stolen Bases
BB.LL	Discrete	The number of seasons a player led their league in Bases on Balls, commonly referred to as walks
AVG.LL	Discrete	The number of seasons a player led their league in Batting Average; this accomplishment is referred to as a ‘batting title’

OBP.LL	Discrete	The number of seasons a player led their league in On Base Percentage
SLG.LL	Discrete	The number of seasons a player led their league in Slugging Percentage
AS	Discrete	The number of seasons a player participated in the All-Star game; the All-Star game is an exhibition game between the best players in the league at each position
GG	Discrete	The number of Gold Glove awards a player won; the award is given to the best defensive player in the league at each position
SS	Discrete	The number of Silver Slugger awards a player won; the award is given to the best offensive player in the league at each position
MVP	Discrete	The number of Most Valuable Player awards a player won; the award is given to the best overall player in the league
TC	Discrete	The number of Triple Crowns a player achieved; this accomplishment is achieved when a player leads the league in Home Runs, RBI, and Batting Average in the same season
R	Continuous	Runs; the number of times a player scored a run by crossing home plate in their career
Singles	Continuous	The number of singles a player recorded in their career, meaning the number of times they recorded a hit and finished on first base
Triples	Continuous	The number of triples a player recorded in their career, meaning the number of times they recorded a hit and finished on third base
RBI	Continuous	Runs Batted In; the career number of times a player's offensive actions led to

		another player scoring a run by crossing home plate
SB.Percent	Continuous	Stolen Base Percentage; the player's career percent of stolen base attempts that were successful
SB.Dif	Continuous	Stolen Base Difference; how many more times a player was successful than unsuccessful at stealing bases in their career
BB	Continuous	Bases on Balls; commonly referred to as walks, the career number of times a player advanced to first base as a result of the pitcher throwing 4 balls (as opposed to strikes) during the plate appearance
BA	Continuous	Batting Average; a measure of a player's career ability to get on base via recording a hit
OBP	Continuous	On Base Percentage; a measure of a player's career ability to get on base via recording a hit, getting walked, or getting hit by a pitch
SF	Continuous	Sacrifice Flies; the career number of times a player hit the ball into the air and got out, but as a result a teammate that was already on base during the player's at-bat advanced to home plate and scored a run
Inn	Continuous	Innings; the total number of innings a player played in the field during their career
Fld.PercentDif	Continuous	How much higher a player's career fielding percentage was than the league average fielding percentage at their position throughout their career
RFGDif	Continuous	How much higher a player's career range factor per game was than the league average

		at their position throughout their career
AB.PerSeas	Continuous	A player's average number of at-bats per 162-game season
R.PerSeas	Continuous	A player's average number of runs scored per 162-game season
Doubles.PerSeas	Continuous	A player's average number of doubles per 162-game season
HR.PerSeas	Continuous	A player's average number of home runs per 162-game season
SO.PerSeas	Continuous	A player's average number of strikeouts per 162-game season
TB.PerSeas	Continuous	A player's average number of total bases per 162-game season
IBB.PerSeas	Continuous	A player's average number of intentional walks per 162-game season

Table 3: The AUC of each of the 10 models tested

Model	Area Under the Curve (AUC)
Logistic Regression	0.8316
Penalized Logistic Regression	**0.9716**
Linear Discriminate Analysis (LDA)	0.9209
Partial Least Squares Discriminate Analysis (PLSDA)	0.9493
Flexible Discriminate Analysis (FDA)	**0.9493**
Support Vector Machine with Linear Kernel	0.9594
Support Vector Machine with Radial Kernel	**0.9797**
Support Vector Machine with Polynomial Kernel	0.9655
Neural Network	0.9716
Neural Network with Model Averaging	**0.9757**

Figure 3: Tuning graph for the penalized logistic regression model to find the optimal alpha (Mixing Percentage) and lambda (Regularization Parameter) values

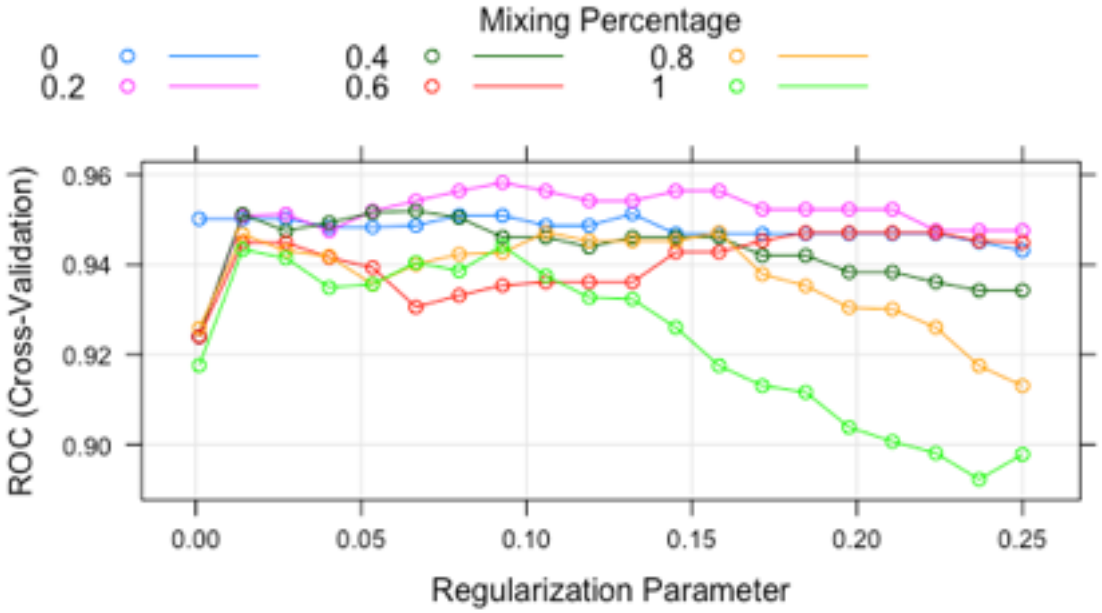

Figure 4: Tuning graph for the flexible discriminate analysis model to find the optimal values for the degree (Product Degree) and number of prunes (#Terms)

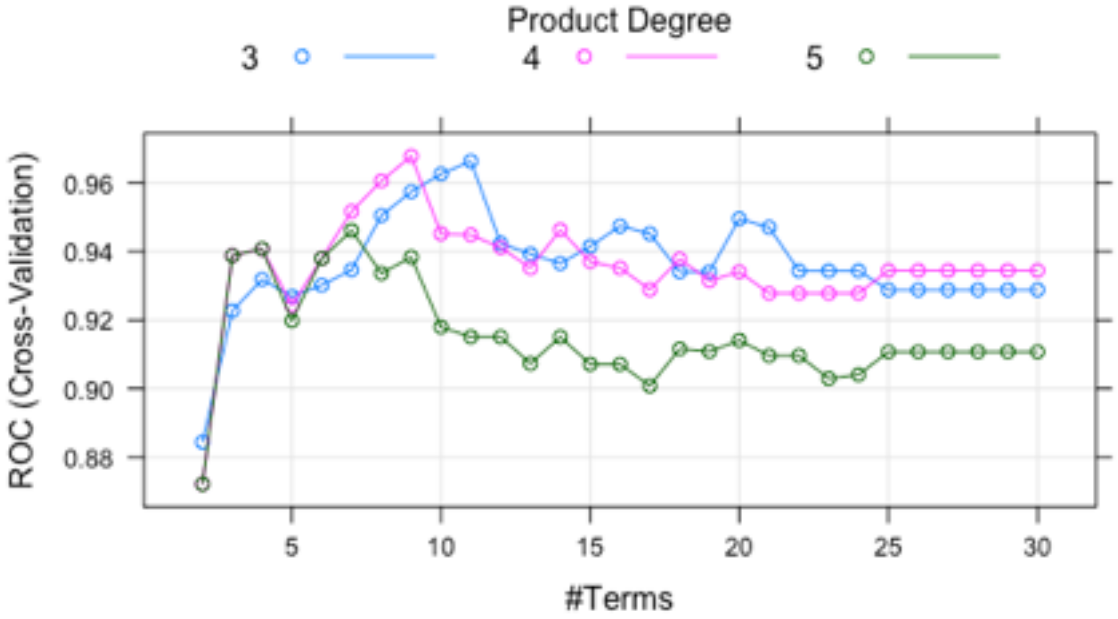

Figure 5: Tuning graph for the support vector machine with radial kernel to find the optimal cost and sigma values

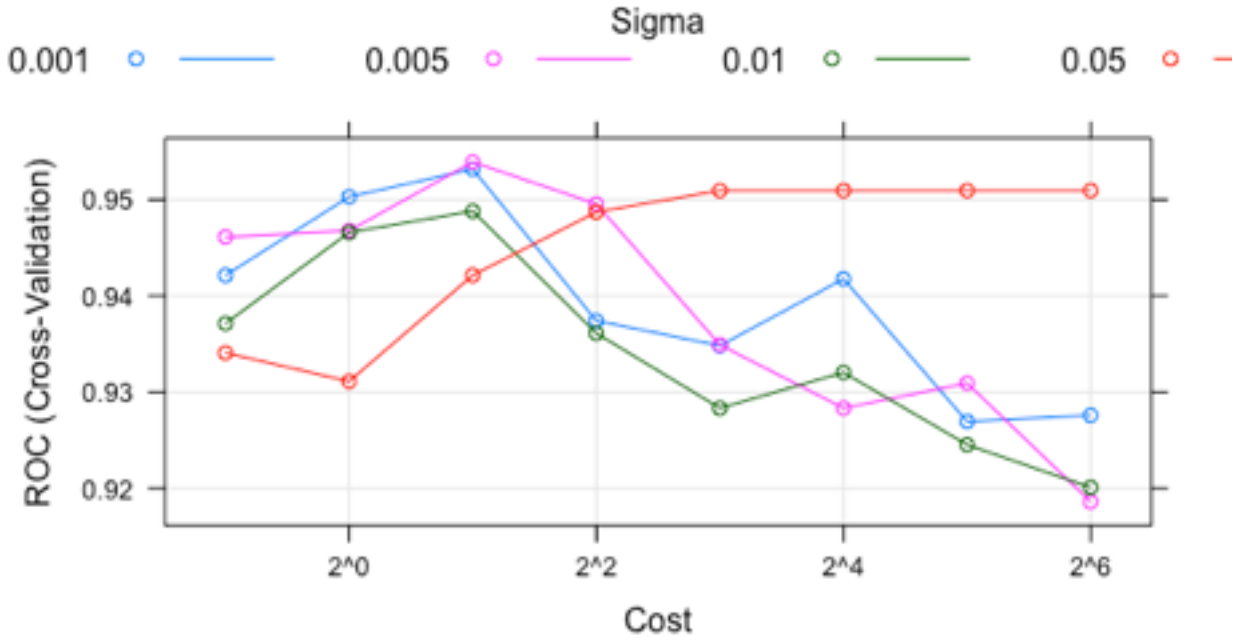

Figure 6: Tuning graph for the model-average neural network to find the optimal size (#Hidden Units) and decay (Weight Decay) values

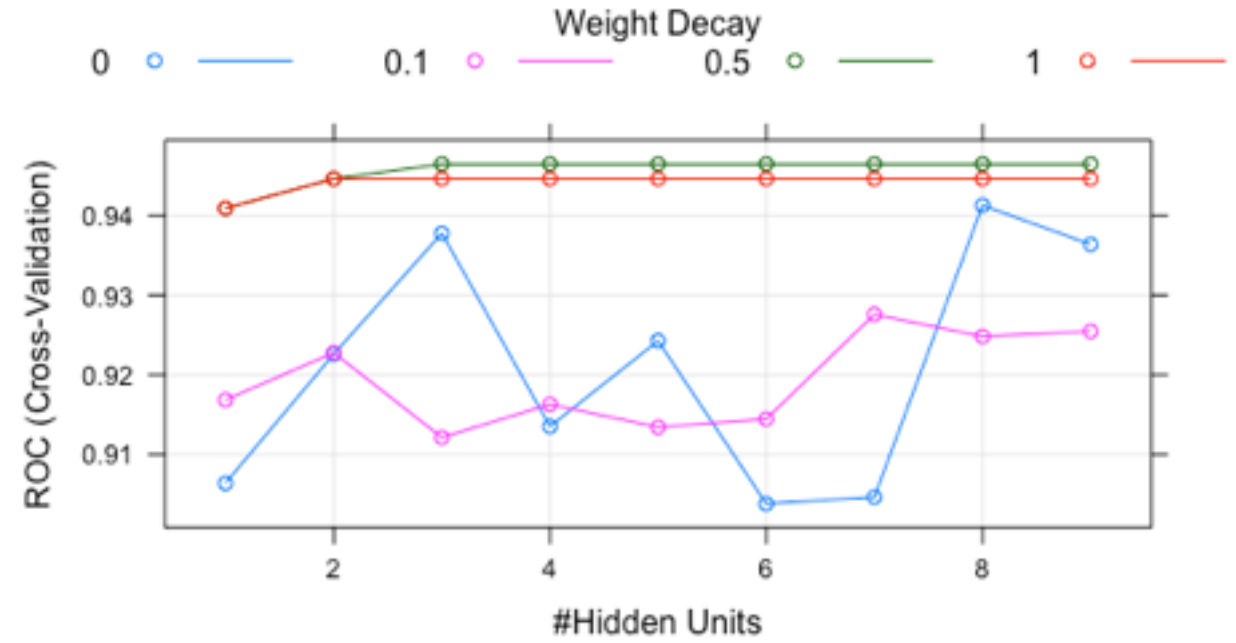

Table 4: A summary of the accuracy, kappa, sensitivity, and specificity values of the final ensemble model and each of the 4 other models used to create it

Model	Accuracy	Kappa	Sensitivity	Specificity
Penalized Logistic	0.8478	0.6523	0.9655	0.6471
FDA	0.8696	0.7201	0.8966	0.8235
SVM with Radial Kernel	0.913	0.8087	0.9655	0.8235
Model-Average Neural Network	0.8696	0.7058	0.9655	0.7059
Final Ensemble	0.9348	0.8583	0.9655	0.8824

Table 5: Final confusion matrix

		Actual	
		Not Hall of Fame	Hall of Fame
Predicted	Not Hall of Fame	28	2
	Hall of Fame	1	15