The Analysis of MLB Playoff Teams

STAT 550 Final Report

Mario A. Leon
Anh Vu
Grant Williams

California State University, Long Beach
December 15, 2022

Abstract

The primary objective for professional sports organizations is to build a winner, as players and front offices often say, "we want to bring a championship to this city". In Major League Baseball, in order to have a chance at a championship, a team needs to qualify for the yearly October postseason, where only 8-10 out of 30 teams qualify. A team can achieve this in various ways, whether that means spending a lot of money, focusing on a specific aspect of a game, dominating a statistic, being a solid home and/or away team, and so on. We will look at Major League Baseball data from 2005 to 2015 along with testing data between 2018 and 2019 in order to analyze specific trends that these playoff teams have. The goal of this study is to comprehend which specific variables likely contribute to a playoff team and how accurate they classify playoff teams and non-playoff teams.

Introduction

Major League Baseball has a very storied history, with the league starting in the late 1800's. Ever since, there have been a handful of teams that have had a lot of success, winning World Series after World Series. The New York Yankees are perhaps the most successful sports franchise in all of North America, achieving a whopping 27 World Series Championships. A lot of people look at the Yankees as the perfect example of what a sports franchise should be. However, what a lot of people do not realize is that the MLB was a lot different back in the day. The "playoffs", per se, was simply a battle between the two best teams in each league and the game was not as developed. There was no bracket and the competition was not like what we see today. There were also less teams back in the day, with the MLB being split into two separate leagues, the American League and National League, increasing to 30 teams in 1998. Nowadays, every MLB team is talented, even the weak teams, and about 10 teams qualify for the postseason.

At the end of the postseason, two teams (the winner of the American League bracket and the winner of the National League bracket) face off in a championship series (known as the World Series) to determine the champion for that year. The time period we will be looking at is 2005 to 2015 but before getting into the analysis, it is important to note that only 8 teams in total (4 in each league) qualified for the postseason from 2005 to 2011 while the league increased the field from 8 to 10 teams in 2012, which makes up a total of 96 playoff teams between 2005 to 2015.

Exploratory Data Analysis

Our data consisted of about forty variables such as Hits, Runs, At-Bats, Strikeouts, and many other typical baseball statistics that one would see. It also consisted of park factors, which shows how good hitting or pitching is in road and away ballparks, attendance, and the payroll for that team in that season, which accumulates the total payroll of the roster. A lot of statistics were combined into baseball metrics such as WHIP, BABIP (batting average on balls in play), and OPS. Really

obvious variables such as wins and losses are deleted as we already know more wins gets a team in the playoffs. If we were looking at just the success of a team, we would be looking at wins or winning percentage to separate the over .500 teams and the under .500 teams but this analysis is only focused on playoff teams. Other useless variables were deleted but we chose to keep AB, at-bats, to determine if playoff teams should be having more at-bats or less due to the fact that certain plate appearances are not counted as at-bats. If we look at certain variables like payroll,

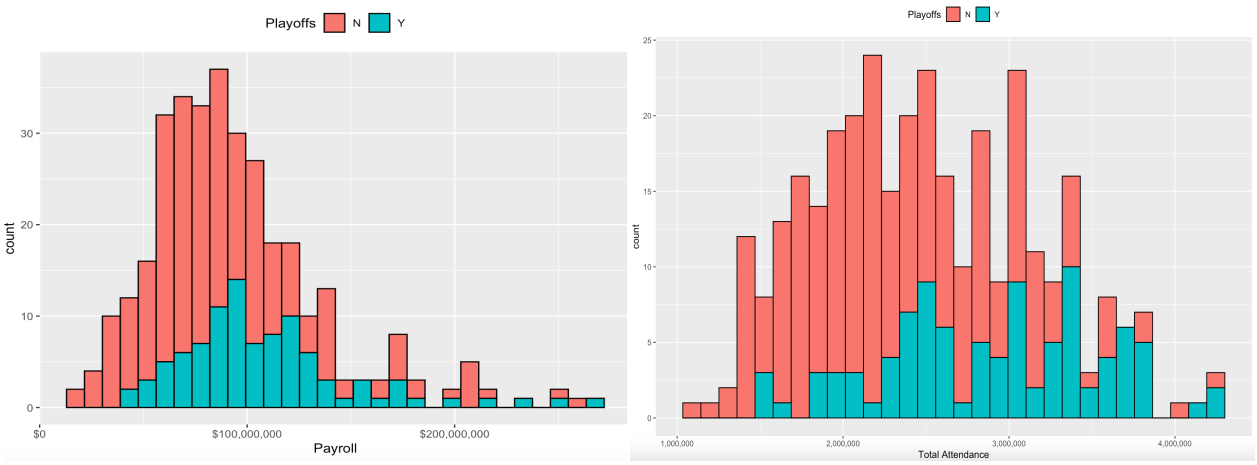


Fig. 1. Histograms showing the relationship of playoff teams vs Payroll and Attendance.

we see that a lot of teams are near the $100,000,000 range and a good chunk of those teams end up making the playoffs. We notice a similar trend with attendance figures and playoff teams, so we would like to see if variables like these are other "not as obvious" variables have some effect on a team making the playoffs or not.

Analysis

In order to analyze different playoff teams, we performed principal component analysis in order to reduce the number of variables we are working with. After narrowing down the dataset to variables that were both relevant and non-obvious contributors to playoff eligibility (for example: number of wins during the regular season could not be included), there were still twenty-eight variables to consider. This reduced dimensionality often helps other models perform better on training data because it removes white noise and irrelevant information that often leads to over-specificity in linear models. Since the data is from 2005 to 2015, we were also interested in testing out future seasons so we chose the 2018 to 2019 seasons so that the split, based on yearID, is close to an 80/20 split.

Because there were many types of data in the original dataset, the scales for each of the columns ranged from decimals to units of over one million (like payroll and attendance). To account for

this, the dataset was completely standardized. In order to complete this procedure, we calculated each column's minimum value, maximum value, and mean. The formula we used for each entry was $\frac{X_{ij}-\mu_i}{max(X_i)-min(X_i)}$, where i represents each column and j represents each row. This standardization assured that any variance associated with each column would not be biased toward variables with a larger scale.

Fig. 2. PCA for MLB 2005-2015 Data

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	5.49313679	1.16719539	0.2034	0.2034
2	4.32594139	1.31063712	0.1602	0.3637
3	3.01530427	1.09084044	0.1117	0.4753
4	1.92446383	0.30778940	0.0713	0.5466
5	1.61667443	0.21166461	0.0599	0.6065
6	1.40500982	0.32882532	0.0520	0.6585
7	1.07618450	0.13111532	0.0399	0.6984
8	0.94506918	0.06442747	0.0350	0.7334
9	0.88064171	0.04889339	0.0326	0.7660
10	0.83174832	0.05779622	0.0308	0.7968
11	0.77395210	0.06704561	0.0287	0.8255
12	0.70690649	0.08562685	0.0262	0.8517
13	0.62127964	0.06720553	0.0230	0.8747
14	0.55407411	0.09721684	0.0205	0.8952
15	0.45685727	0.03187711	0.0169	0.9121
16	0.42498016	0.02737490	0.0157	0.9279
17	0.39760526	0.00153109	0.0147	0.9426
18	0.39607417	0.10666025	0.0147	0.9573
19	0.28941392	0.04627733	0.0107	0.9680
20	0.24313660	0.02076866	0.0090	0.9770
21	0.22236794	0.04066709	0.0082	0.9852
22	0.18170085	0.07280500	0.0067	0.9919
23	0.10889585	0.05535422	0.0040	0.9960
24	0.05354164	0.01429746	0.0020	0.9980
25	0.03924418	0.02942505	0.0015	0.9994
26	0.00981913	0.00384268	0.0004	0.9998
27	0.00597645		0.0002	1.0000

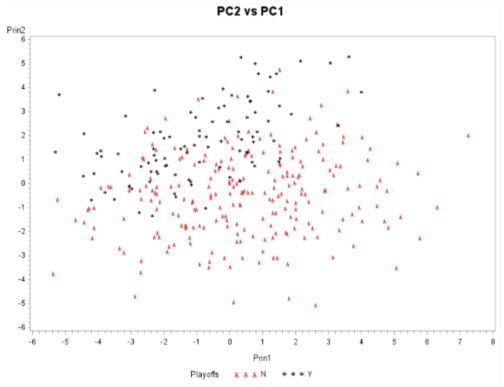
After reducing the dimensions using PCA, Logistic Regression was used to classify the teams as making the playoffs (indicated by a "Y") or not ("N"). Logistic regression particularly benefits from the PCA done earlier because the clarity of the data aids the model in fitting data into distinct classes. This gave the model better predictive power for the 2018 and 2019 seasons, which were used as our testing data. One drawback about using logistic regression with still such a large amount of variables is that it was uncertain whether or not the model would be able to classify the points distinctly, or if there was still too much overlap. For this reason, we also chose to look at linear discriminant analysis to compare the results.

Linear discriminant analysis was similar to logistic regression in that it would be able to classify our data points into binary categories. However, this analysis would also further decrease the dimensionality of our dataset and flatten it into just one or two dimensions. Because information about the classes is known with LDA, a supervised analysis can minimize variance within classes while still optimizing variance between them. We thought this could lead to better predictive power and avoid over-specificity, as the model would be able to handle the same information, but also be able to more clearly separate the data. Now that we know which methods we will be applying, we will see what results we gathered using these methods.

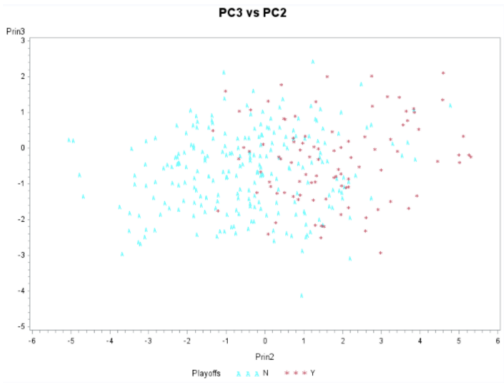
Which variables tells us the most information?

When we applied PCA to the dataset, we see that the the first 18 principal components explain roughly 95%. We chose 95% as out cutoff point to cover all important aspects of the dataset to be safe. The variables featured seem to be a variety of different reasons a team would make the playoffs. We will discuss the interpretations later.

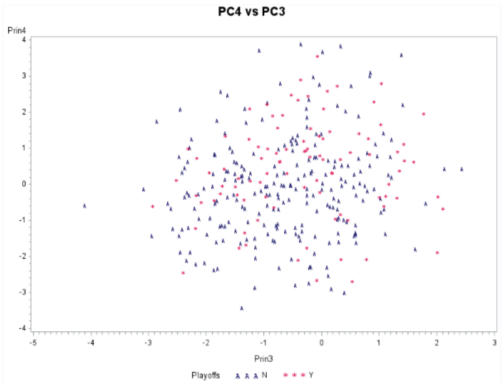
As shown on **App.Fig.1**, we see that the 18 principal components range from $Y_1 = 0.172572 * R + ... - .089270 * AvgAttendance$ to $Y_{18} = 0.0.096982 * R + ... - .241227 * AvgAttendance$. We will test which of these are going to end up significant. As mentioned earlier, 96 teams out of 330 made the playoffs, which is roughly 30%. This is what we will use as the cutoff point for our logistic regression model prediction. The scatterplots of a couple of principal components in **Fig.3** show



(a) PC2 vs PC1



(b) PC3 vs PC2



(c) PC4 vs PC3

Fig. 3. Scatterplots of several principal components.

approximate normality as they roughly form an elliptical shape. Nevertheless, in each of the plots, there is a lot of mixing between each population, implying that Linear Discriminant Analysis may not perform as well as Logistic Regression, since Logistic Regression does not require normality assumption. We will find out if that is indeed the case.

Analysis of the Principal Components

The logistic regression fitted model based on this output is:

$$log(\frac{P(YesPlayoffs)}{P(NoPlayoffs)}) = -3.1142 - 0.8720Prin1 + 1.6308Prin2 + ... - 0.5312Prin15.$$

When we test the principal components, all the principal components in the SAS output in **Fig. 3** are all significant at an $\alpha = 0.1$ significance level. The interpretation of each component of the PCs is that for every 1 unit increase, given the other components remain constant, the log odds that a team makes the postseason increases by $b_(i)$, where i=1,2,3,4,5,6,8,10,11,13,15, all the significant

PCs. If $b_i > 0$, then log odds increase and vice versa. Principal Component 1 (PC1), PC5, PC11, and PC15 are all inversely related to the odds of a team making the postseason due to their negative coefficients.

We can see the relationships a bit more clearly with the correlation matrix for the training set shown on **App.Fig.4**. When we focus on the PCs that are inversely related to the odds of making the playoffs, which are PC1, PC5, PC11, and PC15, we see some interesting characteristics. For the PCs that are inversely related to the odds of success, PC1 has certain offensive statistics such as R and OPS being positively correlated, so they will have an inversely related to PC1. Statistics such as Shutouts, Saves, Home Runs Allowed, and Strikeouts Allowed are positively related to the odds of success so it seems like this PC is more focused on having a team to have shutdown pitching, as also indicated by the strong ERA and strong WHIP. We want a playoff team to have an ERA, HRA, SOA and WHIP as small as possible.

Fig. 4. Significant Principal Components at an $\alpha = 0.1$ significance level.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr> ChiSq
Intercept	1	-3.1142	0.4154	56.2144	<.0001
Prin1	1	-0.8720	0.1348	41.8292	<.0001
Prin2	1	1.6308	0.2126	58.8458	<.0001
Prin4	1	0.7339	0.1671	19.2827	<.0001
Prin5	1	-0.8987	0.2034	19.5322	<.0001
Prin6	1	0.4191	0.1723	5.9154	0.0150
Prin8	1	0.5634	0.2055	7.5159	0.0061
Prin10	1	0.5213	0.2077	6.2986	0.0121
Prin11	1	-0.4242	0.2336	3.2983	0.0694
Prin13	1	0.5553	0.2414	5.2895	0.0215
Prin15	1	-0.5312	0.2680	3.9279	0.0475

PC5 is then focused more on Park Factors, which possibly tells us it does not matter if a team has success in certain ballparks due to the inverse relation. However, it's interesting to note that a higher payroll here diminishes a team's odds of making the playoffs. PC11 dos not really tell us much but one thing to take from this PC is that a team should not rely so much on having sacrifice flies. A playoff team should not be getting out so much as they should be getting base hits more often to drive in runs. PC15 shows a contrast between shutouts, doubles, home runs, strikeouts, home runs allowed, saves, double plays, WHIP, Batter Age, Pitcher Age, Hit By Pitch, Complete Games, BABIP, and Average Attendance. Pretty interesting how each of the PCs focus on different strengths and weaknesses so far.

That is also evident when we look at the positive PCs, such as PC2, PC4, PC6, PC8, PC10, and PC13. PC2 shows R, AB, doubles, Runs allowed, ERA, OPS, Pitcher Age, and Average Attendance being strongly correlated, so this PC is focusing on offensive statistics. PC2 even adds emphasis to a higher payroll to succeed, which is interesting given what PC5 told us. The more runs and doubles a team gets and the higher the OPS, the higher the chances of making the playoffs. PC4 shows that older teams will have a less likely chance of making the playoffs but with a positive emphasis on park factors. The rest of these have different variables with different that are being focused on. The point is that a playoff team may have different strengths and weaknesses.

Classification of Playoff Teams

Earlier, we explained why we set a cutoff point of 30% for a team's probability of making the postseason. We will use the following model to calculate the probability of making the postseason:

$$P(Yes-Playoffs) = \frac{\exp -3.1142-0.8720*Prin1+1.6308*Prin2*0.7339*Prin4-0.8987*Prin5+...-0.5312*Prin15}{1+\exp -3.1142-0.8720*Prin1+1.6308*Prin2*0.7339*Prin4-0.8987*Prin5+...-0.5312*Prin15}$$

If $P(Yes - Playoffs)$ is greater than 30%, it will be classified as Yes. Otherwise, it's a No. Note that the dataset is unbalanced since we have a lot more "N" responses than "Y" responses. The ROC curve for our model in **Fig.6** shows that our model is great at predicting which teams make the postseason from the training set since the curve is nearing 1 as the x-axis gets bigger.

Frequency
Percent
Row Pct
Col Pct

Table of predPlayoffs by Playoffs

predPlayoffs	Playoffs N	Playoffs Y	Total
No	197 59.70 94.71 84.19	11 3.33 5.29 11.46	208 63.03
Yes	37 11.21 30.33 15.81	85 25.76 69.67 88.54	122 36.97
Total	234 70.91	96 29.09	330 100.00

(a) Confusion Matrix for training data (2005-2015).

predPlayoffs	Playoffs N	Playoffs Y	Total
No	39 65.00 84.78 97.50	7 11.67 15.22 35.00	46 76.67
Yes	1 1.67 7.14 2.50	13 21.67 92.86 65.00	14 23.33
Total	40 66.67	20 33.33	60 100.00

(b) Confusion Matrix for testing data (2018-2019).

Fig. 5. Confusion Matrix that shows how well our model classifies playoff teams given the variables provided.

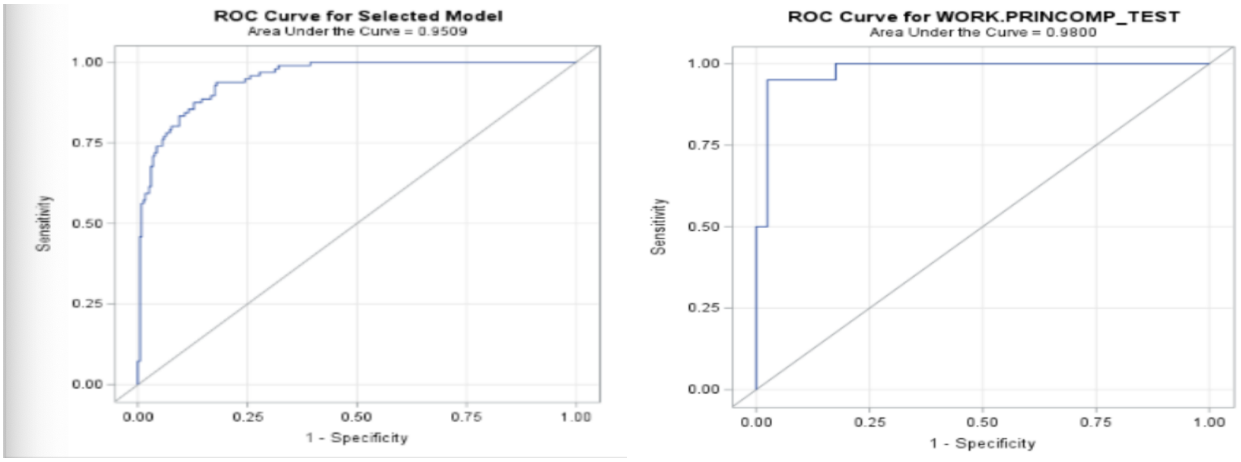
From the confusion matrix in **Fig.5**, we can see that there are 85 observations that were correctly predicted on making the playoffs. However, 11 observations were incorrectly classified as not making the playoffs even though they actually did, so the 85 observations can be called True Positive while 11 observations are called False Negative. These two terms can be used to compute the power of detecting the teams that made the playoffs, which we will call Sensitivity. The formula for it is True Positive divided by the sum of True Positive and False Negative so this calculation gives us a sensitivity of 88.54%.

Another useful feature is Specificity, which is the power to detect teams that did not make the playoffs. This feature can be computed by dividing True Negative by the sum of True Negative and False Positive. True Negative are the number of observations that were correctly classified

as not making the playoffs while False Positive represents the observations incorrectly classified as making the playoffs while they actually did not. The Specificity for this confusion matrix is 84.19%. In overall, the logistic model performed well on the training dataset since its Accuracy is 85.45%. The accuracy is computed by dividing the sum of True Positive and True Negative by the sum of True Positive, True Negative, False Positive, and False Negative. Then, once we obtained the accuracy, the training error rate can be computed by deducting the Accuracy from 100%, which is 14.55%.

When we applied the model to the test data set, the receiver operating characteristic (ROC) curve (**Fig.6**) is close to one as the curve increases which shows that this model is also great at predicting playoff teams. Sensitivity is roughly 65%, which is not as good as the one from training dataset. However, the Specificity came out to be higher, which is 97.50%. The Y-Pred for this validation matrix is also higher, 92.86% while N-Pred is lower than that of the training confusion matrix. Overall, the obtained accuracy and error rate from this confusion matrix is higher and lower respectively than those from the training one.

There were 7 wrong predictions made by the model shown in **App.Fig.6a**. The most interesting one are the Washington Nationals in the 2019 season: they were predicted to not make the playoffs but they actually did and won the World Series in that season. It is likely they were incorrectly classified because they started 19-31 that season and had many struggling points during the season. The Tampa Bay Rays are also an interesting case where they went against the model prediction in both seasons. They made the playoffs while being predicted not in 2019 and vice versa in 2018 likely due to them being a visiting Wild Card team in 2019. The Rockies (2018), Braves (2018-2019), and A's (2018) are the other 4 cases that the model failed to predict for making the playoffs. It may be worth noting that the Rockies and A's were both also the visiting Wild Card teams in 2018 so they barely made the playoffs.



(a) ROC Curve for training data (2005-2015).

(b) ROC Curve for testing data (2018-2019).

Fig. 6. ROC Curves that shows how good the performance of our classification model is.

LDA Classification Rule

Now that we know how the logistic classification is working out, we can apply Linear Discriminant Analysis with our prior probabilities being 70% for "N" and 30% for "Y". We attempted to create the classification rule from the output for the Linear Discriminant Function for Playoffs in **Fig.7**. The left column represents coefficients for each of the 18 principal components associated with N (not making the playoffs) populations. The right column is for Y (making the playoffs) populations. The first row is the constant associated with each population. We created the rule by subtracting coefficients of Y population from coefficients of N population and name this formula $\hat{y}$. After some math, our classification rule is as follows:

$$\hat{y} = 0.57034Prin1 - 1.19526Prin2 + 0.01955Prin3 - 0.52379Prin4 + 0.61553Prin5 - 0.24923Prin6 + 0.08067Prin7 - 0.50905Prin8 + 0.06073Prin9 - 0.38654Prin10 + 0.41096Prin11 + 0.09565Prin12 - 0.51714Prin13 + 0.05186Prin14 + 0.6026Prin15 - 0.04472Prin16 - 0.48486Prin17 - 0.23438Prin18 - 3.51802$$

Fig. 7. Classification Rule of the Linear Discriminant Function for Playoffs.

Linear Discriminant Function for Playoffs		
Variable	N	Y
Constant	-0.73394	-2.78408
Prin1	0.15774	-0.41260
Prin2	-0.14517	1.05009
Prin3	-0.73255	-0.75210
Prin4	0.01138	0.53517
Prin5	0.23559	-0.37994
Prin6	0.09237	0.34160
Prin7	-0.05934	-0.14001
Prin8	0.04986	0.55891
Prin9	-0.31485	-0.37558
Prin10	0.01970	0.40624
Prin11	-0.13647	-0.54743
Prin12	-0.08875	-0.18444
Prin13	-0.37291	0.14423
Prin14	-0.29200	-0.34386
Prin15	0.13282	-0.46978
Prin16	-0.08486	-0.04014
Prin17	-0.08657	0.39829
Prin18	0.23555	0.46993

If a vector, x, with specific values for each of the principal component get substituted in this equation and obtain a value larger than 0, we classify this x as not making the playoffs. If the value is less than 0, we classify this x as making the playoffs.

For example, we could use the rule above to determine if some arbitrary team with specific statistics, say a specific ERA, WHIP, OPS, etc applied to the PCs, will make the playoffs or not. When we look at the confusion matrix when we applied this classification rule on the training dataset (**App.Fig.5**), we get a Sensitivity of 83.33%, Specificity of 90%, and an accuracy of 88.18%. Then, we cross validated the model within the training data to check if the previous confusion matrix holds. We obtain a Sensitivity of 79.16%, Specificity of 88.03%, and an accuracy of 85.45%, which is only slightly lower than before so it is not a drastic change. Applying LDA on the testing dataset gives us a really low sensitiviy of 55%, accuracy of 83.33%, but a really high specificity of 97.5% so this model does a great job predicting non-playoff teams.

The table in **App.Fig.6b** contains classification result for LDA on the validation set. We can see that, similar to the classification output from Logistic Regression, the LDA model also misclasified Washington (2019), Tampa Bay and Atlanta in 2018 and 2019, Oakland (2018), and Colorado (2018). It's possible that statistical analysis did not apply to these teams, especially Washington in 2019, since we know they won the World Series in 2019 despite an awful start to the season. The LDA model also

misclassified Milwaukee (2019), Cleveland (2018), and Oakland (2019). These misclassifications might also be due to the model's poor performance on Sensitivity as Logistic Regression classified these teams correctly.

Conclusion

In conclusion, both the LDA model and the logistic regression model turned out to be relatively accurate predictors for playoff eligibility. The discriminant analysis model had better accuracy and lower error rates when applied to the training data in comparison with logistic regression, but we still think the latter is the better model. For one, the model outperformed LDA in sensitivity when applied to both the testing and training data. Since our main objective is to correctly classify teams that made playoffs given a number of different variables, we mainly focused on sensitivity, which is the power to detect teams that made playoffs. This makes logistic regression the better choice as it implies that even though the LDA model was better at summarizing and classifying the "No" teams, the logistic model was better at prediction. This could be related to the priors we had chosen for the LDA, as well as the fact that the normality plots showed mixing between populations.

The second reason we prefer logistic regression is that based on the confusion matrix, its errors are less systematic. When applied to the test set, the LDA model classified a striking number of "No" values as "Yes." Overall, it seems like a playoff team does not necessarily have to be one specific style of team to qualify. It ultimately depends on what the needs of a team are and where their strengths and weaknesses currently stand. Hitting helps but having too much hitting with bad pitching is not favorable. The season is also really long and it is likely a team was performing poorly for half the season but made a winning run for the other half to make the playoffs, hence the likely incorrect classification that this specific team missed the playoffs.

In the future, we could improve the accuracy of the models by perhaps choosing the twelve 'most likely' teams to make the playoffs per year. This ensures that the correct number of teams get chosen each time, leading to less error in the model overall. This is especially the case as we have seen in real life that certain teams reach the 90 win mark but still fail to make the playoffs while an 80 win team qualifies. We could also attempt to use less PCs (perhaps only 5-8) and see if this has any measurable effect on the outcome of the two models. We did not do that in this case as we wanted to account for at least 95% of the model. It may be possible that logistic works better with more PC variables, but LDA outperforms with less. We could also compare these models to ones which use regular variables instead of PCs in order to assess whether or not they actually improved model accuracy. This topic has several different approaches but for now, we are happy with the results we got given the variables we worked with.

Appendix

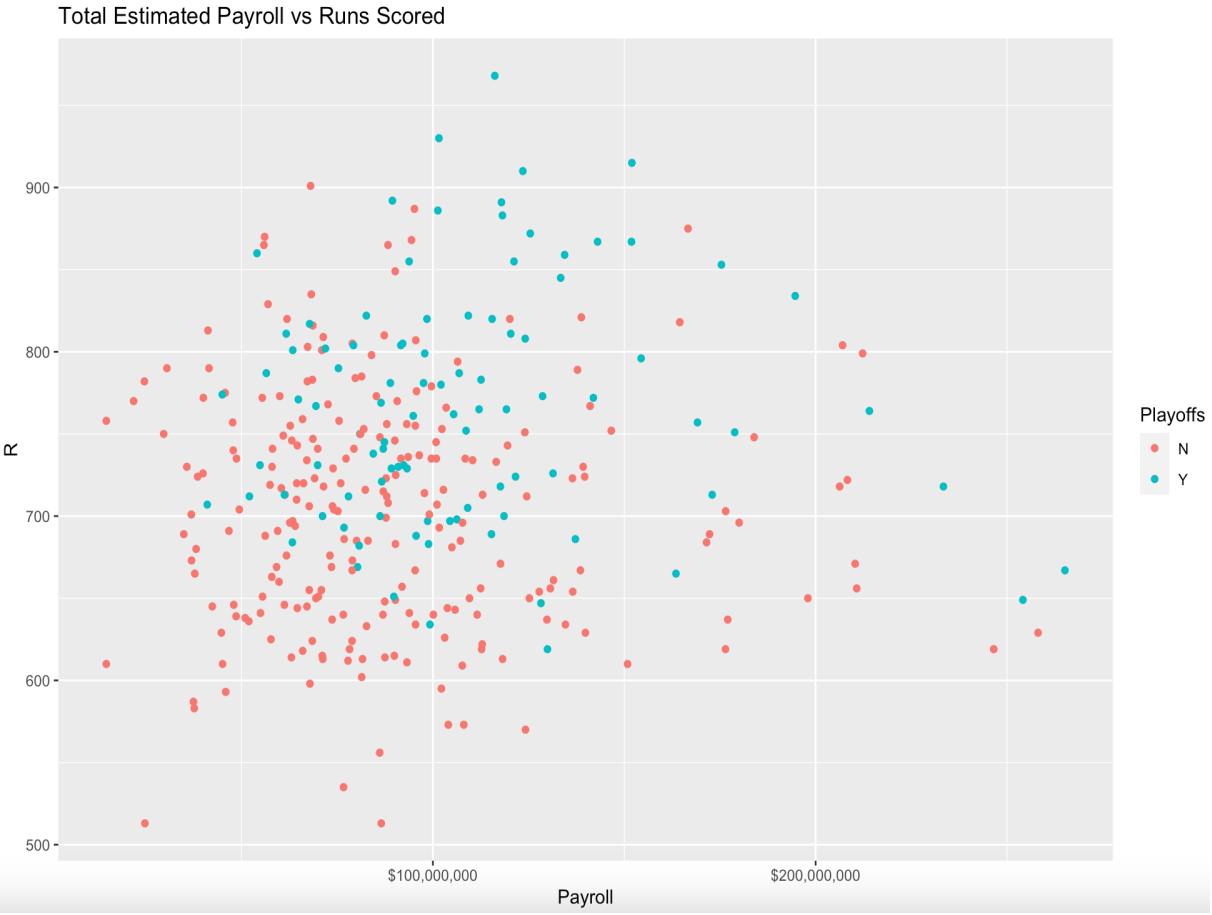
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin16	Prin17	Prin18
R	0.172572	0.355039	0.184082	-.012761	-.222571	0.057692	-.147415	0.105414	0.096982
AB	0.206815	0.249780	0.057246	-.005993	-.124282	-.179015	-.197410	-.476923	-.192630
X2B	0.202996	0.273482	-.012741	0.031682	-.222467	-.101455	0.219681	0.343560	-.307142
X3B	0.074091	0.041537	-.119091	0.270926	0.144900	-.407177	-.203833	0.194425	-.164746
HR	0.075280	0.164538	0.414910	-.086074	-.166569	0.294569	-.193675	0.145892	0.013979
SO	-.062428	-.207909	0.390947	0.140990	-.053570	-.102702	0.164573	0.201187	-.125968
HBP	0.006069	0.039512	0.193334	-.060121	-.314491	0.216209	-.030267	-.068876	0.024919
SF	0.072862	0.233130	-.132823	0.071719	-.208797	-.142615	-.151635	-.082734	-.054370
RA	0.380020	-.135342	0.056746	-.177905	0.056429	-.032911	0.119184	-.003378	0.125955
ERA	0.377841	-.130384	0.056370	-.176485	0.076855	-.014281	0.152802	0.028017	0.185086
CG	-.063563	0.129937	-.294454	-.008398	-.000206	0.334728	0.389782	-.065131	-.164949
SHO	-.289582	0.127801	-.051761	0.101123	-.051776	0.013954	-.050405	-.060732	0.556997
SV	-.221825	0.156617	0.035279	0.141564	-.158887	0.085232	-.541239	-.082647	-.163400
HRA	0.215971	-.096672	0.334169	-.192704	0.010492	0.080180	0.198288	-.161289	-.057467
SOA	-.222986	0.001203	0.404393	0.075386	-.001480	-.161568	-.084577	0.036896	-.132153
DP	0.180928	0.000076	-.253202	0.049408	-.016019	0.207626	-.187850	0.060355	-.295160
FP	-.141214	0.144852	0.028449	0.040471	0.070084	0.197161	0.099586	0.156320	0.191425
BPF	0.154037	0.165165	0.152466	0.400661	0.401626	0.232990	-.020778	-.101875	-.017887
PPF	0.186177	0.103873	0.137940	0.397039	0.420134	0.229774	-.005973	-.110046	-.031000
SBperc	-.006435	0.112046	0.049789	-.014909	0.136573	-.076789	0.093398	0.057489	0.058308
OPS	0.174478	0.356021	-.112656	0.044421	-.154671	0.136336	-.065803	0.180999	0.182557
BABIP	0.179221	0.170063	0.001662	0.278031	-.114683	-.431519	-.385441	-.052889	0.219334
WHIP	0.368828	-.120452	-.089506	-.127037	0.085023	-.030723	-.115992	0.136737	0.212187
BatAge	-.062109	0.229347	-.128986	-.380343	0.288654	0.040077	-.006197	0.234908	-.216030
PAge	-.048276	0.269297	0.037282	-.315374	0.155136	-.113124	-.073793	-.523485	-.029311
EstPayroll	-.146042	0.193756	0.217058	-.251061	0.281532	-.214282	-.056628	0.153948	-.129260
AvgAttendance	-.089270	0.319327	0.027099	-.168412	0.248384	-.098972	-.010847	0.168136	0.241227

App.Fig.1 The 18 Principal Components showing only the first few and last few.

App.Fig.2 A glimpse of our dataset.

yearID	lgID	franchID	divID	Ghome	R	AB	H	X2B	X3B	HR	SO	HBP	SF	RA	ER	ERA	CG	SHO	SV	HRA
2005	NL	ARI	W	81	696	5550	1419	291	27	191	1094	55	45	856	783	4.84	6	10	45	193
2005	NL	ATL	E	81	769	5486	1453	308	37	184	1084	45	46	674	639	3.98	8	12	38	145
2005	AL	BAL	E	81	729	5551	1492	296	27	189	902	54	42	800	724	4.56	2	9	38	180
2005	AL	BOS	E	81	910	5626	1579	339	21	199	1044	47	63	805	752	4.74	6	8	38	164
2005	AL	CHW	C	81	741	5529	1450	253	23	200	1002	79	49	645	592	3.61	9	10	54	167
2005	NL	CHC	C	81	703	5584	1506	323	23	194	920	50	37	714	671	4.19	8	10	39	186
2005	NL	CIN	C	82	820	5565	1453	335	15	222	1303	62	39	889	820	5.15	2	1	31	219
2005	AL	CLE	C	81	790	5609	1522	337	30	207	1093	54	50	642	582	3.61	6	10	51	157
2005	NL	COL	W	81	740	5542	1477	280	34	150	1103	64	34	862	808	5.13	4	4	37	175
2005	AL	DET	C	81	723	5602	1521	283	45	168	1038	53	52	787	719	4.51	7	2	37	193
2005	NL	FLA	E	81	717	5502	1499	306	32	128	918	67	50	732	666	4.16	14	15	42	116
2005	NL	HOU	C	81	693	5462	1400	281	32	161	1037	72	42	609	563	3.51	6	11	45	155
2005	AL	KCR	C	81	701	5503	1445	289	34	126	1008	63	50	935	862	5.49	4	4	25	178
2005	AL	ANA	W	81	761	5624	1520	278	30	147	848	29	39	643	598	3.68	7	11	54	158
2005	NL	LAD	W	81	685	5433	1374	284	21	149	1094	67	33	755	695	4.38	6	9	40	182
2005	NL	MIL	C	81	726	5448	1413	327	19	175	1162	73	38	697	635	3.97	7	6	46	169
2005	AL	MIN	C	81	688	5564	1441	269	32	134	978	59	42	662	604	3.71	9	8	44	169
2005	AL	NYY	E	81	886	5624	1552	259	16	229	989	73	43	789	718	4.52	8	14	46	164
2005	NL	NYM	E	81	722	5505	1421	279	32	175	1075	48	38	648	599	3.76	8	11	38	135
2005	AL	OAK	W	81	772	5627	1476	310	20	155	819	52	40	658	594	3.69	9	12	38	154
2005	NL	PHI	E	81	807	5542	1494	282	35	167	1083	56	46	726	672	4.21	4	6	40	189
2005	NL	PIT	C	81	680	5573	1445	292	38	139	1092	72	49	769	706	4.42	4	14	35	162
2005	NL	SDP	W	81	684	5502	1416	269	39	130	977	49	48	726	668	4.13	4	8	45	146
2005	AL	SEA	W	81	699	5507	1408	289	34	130	986	48	37	751	712	4.49	6	7	39	179
2005	NL	SFG	W	81	649	5462	1427	299	26	128	901	49	44	745	695	4.33	4	8	46	151
2005	NL	STL	C	81	805	5538	1494	287	26	170	947	62	35	634	560	3.49	15	14	48	153
2005	AL	TBD	E	81	750	5552	1519	289	40	157	990	69	51	936	851	5.39	1	4	43	194
2005	AL	TEX	W	81	865	5716	1528	311	29	260	1112	48	32	858	794	4.96	2	6	46	159
2005	AL	TOR	E	81	775	5581	1480	307	39	136	955	89	56	705	653	4.06	9	8	35	185
2005	NL	WSN	E	81	639	5426	1367	311	32	117	1090	89	45	673	627	3.87	4	9	51	140
2006	NL	ARI	W	81	773	5645	1506	331	38	160	965	67	53	788	727	4.48	8	9	34	168
2006	NL	ATL	E	81	849	5583	1510	312	26	222	1169	52	44	805	736	4.6	6	6	38	183
2006	AL	BAL	E	81	768	5610	1556	288	20	164	878	73	41	899	843	5.35	5	9	35	216
2006	AL	BOS	E	81	820	5619	1510	327	16	192	1056	66	56	825	773	4.83	3	6	46	181
2006	AL	CHW	C	81	868	5657	1586	291	20	236	1056	58	57	794	743	4.61	5	11	46	200
2006	NL	CHC	C	81	716	5587	1496	271	46	166	928	43	37	834	758	4.74	2	7	29	210
2006	NL	CIN	C	82	749	5515	1419	291	12	217	1192	59	38	801	725	4.51	9	10	36	213
2006	AL	CLE	C	81	870	5619	1576	351	27	196	1204	54	43	782	698	4.41	13	13	24	166
2006	NL	COL	W	81	813	5562	1504	325	54	157	1108	60	45	812	749	4.66	5	8	34	155
2006	AL	DET	C	81	822	5642	1548	294	40	203	1133	45	36	675	618	3.84	3	16	46	160

App.Fig.3 Scatterplot between runs scored and total estimated payroll colored by playoff teams.



Pearson Correlation Coefficients, N = 330
Prob > |r| under H0: Rho=0

	R	AB	X2B	X3B	HR	SO	HBP	SF	RA	ERA
Prin1	0.51193 <.0001	0.50780 <.0001	0.52944 <.0001	0.17477 0.0014	0.28437 <.0001	-0.20918 0.0001	0.07598 0.1685	0.22952 <.0001	0.90320 <.0001	0.89967 <.0001
Prin2	0.77814 <.0001	0.54936 <.0001	0.55405 <.0001	0.08579 0.1195	0.46015 <.0001	-0.38151 <.0001	0.14880 0.0068	0.45508 <.0001	-0.17796 0.0012	-0.16349 0.0029
Prin4	-0.01095 0.8429	-0.01772 0.7484	0.01892 0.7321	0.33872 <.0001	-0.06039 0.2740	0.34491 <.0001	-0.06420 0.2448	0.06525 0.2372	-0.24197 <.0001	-0.24016 <.0001
Prin5	-0.26863 <.0001	-0.15345 0.0052	-0.32179 <.0001	0.15884 0.0038	-0.17611 0.0013	-0.07878 0.1533	-0.37687 <.0001	-0.31513 <.0001	0.01252 0.8207	0.03823 0.4839
Prin6	0.06828 0.2160	-0.20405 0.0002	-0.11413 0.0383	-0.49936 <.0001	0.45966 <.0001	-0.06621 0.2303	0.28777 <.0001	-0.19493 0.0004	-0.02460 0.6562	-0.00257 0.9629
Prin8	0.03580 0.5169	-0.33154 <.0001	-0.05844 0.2899	0.36668 <.0001	0.11372 0.0390	0.12899 0.0191	0.42455 <.0001	-0.09694 0.0787	0.05688 0.3029	0.01969 0.7215
Prin10	0.04760 0.3868	0.07514 0.1733	-0.17427 0.0015	0.22250 <.0001	0.07411 0.1793	-0.10415 0.0588	0.02273 0.6807	-0.07563 0.1705	0.02980 0.5895	0.06096 0.2695
Prin11	-0.02562 0.6428	-0.18203 0.0009	0.04061 0.4623	0.17752 0.0012	-0.22830 <.0001	-0.27396 <.0001	0.36327 <.0001	0.51256 <.0001	0.05924 0.2833	0.08907 0.1063
Prin13	0.06839 0.2153	-0.13387 0.0149	-0.23733 <.0001	-0.01453 0.7926	0.05457 0.3230	0.00090 0.9870	-0.28235 <.0001	0.51563 <.0001	0.02287 0.6790	0.02065 0.7035
Prin15	-0.02587 0.6397	-0.03936 0.4761	0.15542 0.0047	-0.04312 0.4350	0.10919 0.0475	0.17475 0.0014	-0.17282 0.0016	0.04241 0.4426	0.06228 0.2592	0.04839 0.3809

Pearson Correlation Coefficients, N = 330
Prob > |r| under H0: Rho=0

	OPS	BABIP	WHIP	BatAge	PAge	ZstPayroll	AvgAttendance
Prin1	0.54219 <.0001	0.42963 <.0001	0.86376 <.0001	-0.15692 0.0043	-0.09045 0.1010	-0.35504 <.0001	-0.19323 0.0004
Prin2	0.75193 <.0001	0.39380 <.0001	-0.20695 0.0002	0.47855 <.0001	0.55923 <.0001	0.46605 <.0001	0.66087 <.0001
Prin4	0.00570 0.9178	0.36353 <.0001	-0.19695 0.0003	-0.59877 <.0001	-0.48329 <.0001	-0.34120 <.0001	-0.26880 <.0001
Prin5	-0.25722 <.0001	-0.17430 0.0015	0.02762 0.6172	0.38574 <.0001	0.23262 <.0001	0.43612 <.0001	0.35013 <.0001
Prin6	0.11111 0.0437	-0.52716 <.0001	-0.04902 0.3747	0.01437 0.7949	-0.16836 0.0021	-0.25117 <.0001	-0.14531 0.0082
Prin8	0.08235 0.1355	-0.09519 0.0843	0.07449 0.1770	-0.05199 0.3464	0.26859 <.0001	-0.06729 0.2228	0.11372 0.0390
Prin10	0.02054 0.7101	-0.11101 0.0439	0.09104 0.0988	-0.03709 0.5019	0.06718 0.2235	-0.18789 0.0006	0.06009 0.2764
Prin11	-0.03797 0.4918	-0.07037 0.2023	0.04185 0.4486	0.24945 <.0001	-0.04487 0.4166	-0.00821 0.8819	-0.04143 0.4532
Prin13	0.01509 0.7848	0.05143 0.3517	-0.00405 0.9415	-0.14478 0.0084	-0.03772 0.4947	0.11014 0.0456	0.08930 0.1054
Prin15	0.01980 0.7201	-0.10658 0.0531	0.09915 0.0720	0.21212 0.0001	0.17080 0.0018	-0.02235 0.6858	-0.28288 <.0001

Pearson Correlation Coefficients, N = 330
Prob > |r| under H0: Rho=0

	CG	SHO	SV	HRA	SOA	DP	FP	BPF	PPF	SBperc
Prin1	-0.18904 0.0006	-0.67613 <.0001	-0.50031 <.0001	0.62291 <.0001	-0.64030 <.0001	0.43698 <.0001	-0.33060 <.0001	0.35750 <.0001	0.42077 <.0001	-0.00337 0.9513
Prin2	0.22221 <.0001	0.21396 <.0001	0.27433 <.0001	-0.04597 0.4052	0.06442 0.2432	-0.00002 0.9998	0.29888 <.0001	0.35248 <.0001	0.23434 <.0001	0.22985 <.0001
Prin4	-0.05103 0.3554	0.12247 0.0261	0.18453 0.0008	-0.24096 <.0001	0.21931 <.0001	0.04870 0.3779	0.03116 0.5727	0.55240 <.0001	0.55094 <.0001	-0.03956 0.4739
Prin5	-0.01667 0.7629	0.00291 0.9580	-0.17338 0.0016	-0.00016 0.9977	0.11644 0.0345	-0.12906 0.0190	0.07645 0.1659	0.48866 <.0001	0.49845 <.0001	0.14646 0.0077
Prin6	0.40027 <.0001	0.01480 0.7888	0.09524 0.0841	0.20211 0.0002	-0.16123 0.0033	0.22969 <.0001	0.25867 <.0001	0.28869 <.0001	0.28932 <.0001	-0.12473 0.0234
Prin8	0.01831 0.7403	-0.04691 0.3956	0.11235 0.0414	-0.03071 0.5783	-0.07508 0.1737	-0.12601 0.0221	-0.58959 <.0001	0.05274 0.3395	0.05177 0.3485	0.18739 0.0006
Prin10	-0.47409 <.0001	-0.28068 <.0001	0.45218 <.0001	0.14031 0.0107	-0.15610 0.0045	0.02155 0.6965	0.37468 <.0001	-0.08130 0.1405	-0.08555 0.1209	0.11966 0.0298
Prin11	0.06082 0.2706	0.03623 0.5119	0.00546 0.9212	0.15934 0.0037	0.02300 0.6771	-0.22368 <.0001	0.12879 0.0193	0.07784 0.1583	0.08803 0.1104	-0.24176 <.0001

Pearson Correlation Coefficients, N = 330
Prob > |r| under H0: Rho=0

	CG	SHO	SV	HRA	SOA	DP	FP	BPF	PPF	SBperc
Prin13	0.10232 0.0634	0.04584 0.4065	0.18935 0.0005	0.10009 0.0694	-0.11369 0.0390	0.33257 <.0001	-0.06494 0.2394	-0.04059 0.4624	-0.04195 0.4475	-0.09959 0.0708
Prin15	-0.22842 <.0001	0.35463 <.0001	0.13109 0.0172	-0.07932 0.1505	-0.02447 0.6578	0.10198 0.0643	-0.07911 0.1516	0.00518 0.9253	0.02712 0.6235	0.00593 0.9145

App.Fig.4 Correlation Matrix of the Significant Principal Components

Assess Accuracy

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.PRINCOMP_TRAIN
Resubstitution Summary using Linear Discriminant Function

Number of Observations and Percent Classified into Playoffs

From Playoffs	N	Y	Total
N	211 90.17	23 9.83	234 100.00
Y	16 16.67	80 83.33	96 100.00
Total	227 68.79	103 31.21	330 100.00
Priors	0.7	0.3	

Error Count Estimates for Playoffs

	N	Y	Total
Rate	0.0983	0.1667	0.1188
Priors	0.7000	0.3000	

(a) Confusion Matrix of the LDA for the training data (2005-2015).

Assess Accuracy

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.PRINCOMP_TRAIN
Cross-validation Summary using Linear Discriminant Function

Number of Observations and Percent Classified into Playoffs

From Playoffs	N	Y	Total
N	206 88.03	28 11.97	234 100.00
Y	20 20.83	76 79.17	96 100.00
Total	226 68.48	104 31.52	330 100.00
Priors	0.7	0.3	

Error Count Estimates for Playoffs

	N	Y	Total
Rate	0.1197	0.2083	0.1463
Priors	0.7000	0.3000	

(b) Confusion Matrix of the LDA for the cross-validated training data.

Classification Summary for Test Data: WORK.PRINCOMP_TEST
Classification Summary using Linear Discriminant Function

Observation Profile for Test Data

Number of Observations Read	60
Number of Observations Used	60

Number of Observations and Percent Classified into Playoffs

From Playoffs	N	Y	Total
N	39 97.50	1 2.50	40 100.00
Y	9 45.00	11 55.00	20 100.00
Total	48 80.00	12 20.00	60 100.00
Priors	0.7	0.3	

Error Count Estimates for Playoffs

	N	Y	Total
Rate	0.0250	0.4500	0.1525
Priors	0.7000	0.3000	

(c) Confusion Matrix of the LDA for the testing data (2018-2019).

App.Fig.5 Confusion matrices that show the classification rate from our Linear Discriminant Analysis.

Obs	yearID	franchID	probsucc	predPlayoffs	Playoffs
41	2018	COL	0.17072	No	Y
42	2018	ATL	0.18639	No	Y
43	2019	ATL	0.18651	No	Y
44	2019	WSN	0.20937	No	Y
45	2018	OAK	0.28836	No	Y
46	2019	TBD	0.29109	No	Y
47	2019	OAK	0.40169	Yes	Y
48	2018	CLE	0.47809	Yes	Y
49	2018	MIL	0.51615	Yes	Y
50	2018	TBD	0.64285	Yes	N
51	2018	CHW	0.64419	Yes	Y
52	2019	NYM	0.68312	Yes	Y
53	2019	STL	0.79702	Yes	Y
54	2019	MIN	0.82139	Yes	Y
55	2018	BOS	0.91994	Yes	Y
56	2018	LAD	0.96449	Yes	Y
57	2018	NYM	0.97242	Yes	Y
58	2018	HOU	0.98443	Yes	Y
59	2019	HOU	0.99509	Yes	Y
60	2019	LAD	0.99890	Yes	Y

(a)

Obs	yearID	franchID	N	Y	Playoffs	_INTO_
37	2018	OAK	0.83671	0.16329	Y	N
38	2019	MIL	0.83357	0.16643	Y	N
39	2019	CLE	0.83230	0.16770	N	N
40	2019	TBD	0.83019	0.16981	Y	N
41	2019	OAK	0.78355	0.21645	Y	N
42	2018	CLE	0.76638	0.23362	Y	N
43	2018	ATL	0.73529	0.26471	Y	N
44	2019	WSN	0.72980	0.27020	Y	N
45	2019	CHW	0.72294	0.27706	N	N
46	2018	COL	0.70548	0.29452	Y	N
47	2019	ATL	0.59156	0.40844	Y	N
48	2018	STL	0.52058	0.47942	N	N
49	2018	MIL	0.48940	0.51060	Y	Y
50	2018	TBD	0.48649	0.51351	N	Y
51	2018	CHW	0.48615	0.51385	Y	Y
52	2019	MIN	0.34020	0.65980	Y	Y
53	2019	NYM	0.24482	0.75518	Y	Y
54	2019	STL	0.13432	0.86568	Y	Y
55	2018	BOS	0.12446	0.87554	Y	Y
56	2018	LAD	0.05827	0.94173	Y	Y
57	2018	HOU	0.05634	0.94366	Y	Y
58	2018	NYM	0.04435	0.95565	Y	Y
59	2019	HOU	0.03824	0.96176	Y	Y
60	2019	LAD	0.00297	0.99703	Y	Y

(b)

App.Fig.6 Misclassification results from our Logistic (a) and LDA (b) models on our test (2018-2019) data.