2016

Constructing a Predictive Model for the Winner of Survivor

Danielle N. Dobie
Minnesota State University Mankato

Constructing a Predictive Model for the Winner of Survivor

by

Danielle Dobie

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Masters of Arts

In

Mathematics

Minnesota State University, Mankato

Mankato, Minnesota

May 2016

Constructing a Predictive Model for the Winner of Survivor

Danielle Dobie

This thesis has been examined and approved by the following members of the thesis committee.

_____
Dr. In-Jae Kim, Advisor

_____
Dr. Deepak Sanjel

_____
Daardi Sizemore

Constructing a Predictive Model for the Winner of Survivor

Dobie, Danielle Nicole M.A. in Mathematics, Minnesota State University, Mankato, Minnesota, May 2016.

Abstract. Throughout this paper, we discuss various predictive models that could be used to predict the winner of CBS's reality television show, Survivor, which is hosted by Jeff Probst. We first give an in-depth explanation to how the data were collected and sorted, and what the variables in the data mean. We then apply a series of predictive models to the data and analyze the results in order to determine whether the winner of Survivor can be predicted based on information the audience knows prior to the merge. If a model under consideration does not work, we explain why it fails. For the predictive model that we eventually propose for the show, we first apply Principal Component Analysis in order to achieve dimension reduction on the number of continuous variables of the collected data and then quantize them to construct a Naïve Bayes' Classifier model along with other categorical variables.

Table of Contents

1 Introduction 1
1.1 Overview of Survivor 1
1.2 Data Collection 3
2 Logistic Regression 13
2.1 Analysis of Logistic Regression Results 14
3 Dimension Reduction Techniques 16
3.1 Linear Discriminant Analysis 16
3.1.1 Analysis of Linear Discriminant Analysis Results 17
3.2 Principal Component Analysis 17
3.2.1 Analysis of Principal Component Analysis 19
3.2.2 Results when Combining Principal Component Analysis 20
4 Model Construction Using a Bayesian Approach 22
4.1 Bayes' Theorem 22
4.2 Naïve Bayes' Classification 24
4.2.1 Analysis of Naïve Bayes' Classifier Results 28
4.2.2 Further Analysis of Naïve Bayes' Classifier Model 30
5 Conclusion 33
5.1 Future Work 33
6 Appendix 34
6.1 Data Table 34
6.2 Naïve Bayes Classifier and PCA Code in SAS 37

Bibliography 43

1 INTRODUCTION

Reality television has been one of the biggest sensations of the 21st century. Many people feel as though Survivor ignited the flame that made reality television so popular (Yahr, Moore & Chow, 2015). During the 16 years that Survivor has been in existence, 32 seasons have been filmed, 31 of them have currently been televised in full. For the many regular viewers of the show, it is only natural to wonder who will win, before the season has been completed. With this in mind, our goal is to construct a predictive model to crown a winner of Survivor based on pre-merge information. We construct our proposed model using 18 seasons of Survivor. Then we use the 24th season to test the model.

1.1 OVERVIEW OF SURVIVOR

Survivor has evolved slightly since it premiered in 2000, but the basic premise of the show has not changed. In every season of Survivor, players are divided into tribes of an equal number of people. The number of tribes at the start of the season has varied among two, three, and four. Every few days, the tribes meet up to compete in a variety of challenges. The challenges are broken into two categories: reward challenges and immunity challenges. The tribe that loses the immunity challenge gets sent to tribal council where they have to vote off one member of their tribe. Each member of the tribe must cast one vote against someone else on their tribe. The tribe member who receives the most votes gets sent home and has no chance at winning the game. If there is a tie at tribal council, then the contestants who received the tied votes, are not allowed to vote in

the tie-breaker round. All other tribe members vote again, this time, they are only voting on the players who received the tied votes. The contestant who receives the most votes from the tie-breaker round would be sent home.

About a third to a half of the way through the season, the tribes merge to form one tribe. At this point in the game, the immunity challenges become individual challenges. The contestant who wins this challenge cannot be voted off at tribal council and is still able to vote while anyone else may be voted off. Starting in season three, contestants would often face a tribe swap before the merge. This gives the contestants an opportunity to meet members from the other tribes.

Starting in season 12, hidden immunity idols were introduced to the game. Hidden immunity idols are always hidden, as its name implies. Because they are hidden, they are typically found in private. Therefore, any contestant who finds a hidden immunity idol does not have to tell their tribe mates. Hidden immunity idols may be played at tribal council, and they are a form of individual immunity. A contestant can play a hidden immunity idol at any tribal council, excluding the last two councils. This idol can be used to save the finder or the finder may choose to play it to save another tribe mate. Any votes that are cast against the player for whom the idol is played will not count, and the contestant who receives the majority of the remaining votes would be voted off instead. Because of how powerful hidden immunity idols are, they have forced contestants to rethink their strategies.

1.2 DATA COLLECTION

In order to build a predictive model for the show, the only seasons we have considered in our study are the seasons in which every contestant is a first time player, there are no players returned to the game after being voted off, and no two people on the show knew each other before beginning of the game. Therefore, the seasons we collected data on are seasons 1, 2, 3, 4, 5, 6, 9, 10, 12, 13, 14, 15, 17, 18, 19, 21, 24, 28, 30. Seasons 8, 11, 16, 20, 22, 23, 25, 26, 27, and 31 have at least one player who is not playing Survivor for the first time. Therefore, to eliminate any advantages or disadvantages to those contestants, we exclude these seasons for our study. Seasons 7 is not part of this study because at some point in the game, two players who had been voted off of their tribe were brought back into the game. (Burnett, 2003 ep.7) Again, to eliminate any advantages or disadvantages to these contestants, we decided to not include this season in the study. Finally, season 29 is not part of this study because each contestant knew one other contestant before the show started filming. Therefore, to eliminate any player biases, we excluded this season from our study. We chose to use season 24 to test our model because season 24 started with two tribes. Out of the 31 seasons of Survivor, 25 of them started with two tribes. Seasons 28 and 30 both started with three tribes, which is not the norm on Survivor.

Our data collection process consists of watching each episode prior to the merge of each season and collecting the values of both categorical and continuous variables. A complete list of the variables can be found in the Appendix 6.1. The table for the list has each variable's name on the Excel spreadsheet, what that shorthand name stands for, and,

when applicable, what each of the categories represents. We now describe the process by which each variable was collected.

The following is a list of the continuous variables we collected, along with their collection process.

Percent of tribe in original alliance: This information was collected by watching the seasons. In order for tribe members to be considered in an alliance, that group of contestants verbally confirm that they are in an alliance together. They must also vote together at least one time, unless they agree to split the votes in case a hidden immunity idol is played. Season 19 is the only exception to this rule. During this season, they showed several people of the Foa Foa tribe grouping off and making an alliance, but never got that whole group of people together to state that they were in an alliance together (Burnett, 2009). This group of people all voted together at the first tribal council. Therefore, we took these actions as implication that they were in an alliance together. Other than season 19, there are two special cases for how we collected data for this category. The first took place during season 12. During this season, the tribes were split up into four tribes of four members each (Burnett, 2006 ep.1). After only one tribal council, the four original tribes were brought to an end and two new tribes were chosen. During this season, we computed the percentages of tribe in original alliance based on the tribes that were formed on day four, not the original four tribes having four members each. The other special case scenario took place during season 13. Here, there were four tribes of five members each. Because the number of people

on each tribe was so small, we started counting percentage of tribe in each contestant's alliance after the tribe swap which took place on day seven (Burnett, 2006 ep.3)

Percent of team immunity challenges won: This information was collected by watching the seasons. Each time an immunity challenge was played, the ratio of wins and losses was recorded so that it could later be converted into percentages. Whenever a tribe did not finish last during an immunity challenge, it was recorded that that tribe won the challenge. The only time the losing tribe got credit for winning is when the winning tribe decided to give up immunity to the losing tribe so that they could go to tribal council. This only happened twice; the first one happened during season 14 by the Moto tribe (Burnett, 2006 ep.4) and the second happened during season 24 by the Manono tribe (Burnett, 2012 ep.4).

Percent of team reward challenges won: This information was collected by watching the seasons. Each time a reward challenge was played, the ratio of wins and losses was recorded so that it could later be converted into percentages. Whenever a tribe did not finish last during the reward challenge, it was recorded that the tribe won the challenge.

Percent of votes received pre-merge: This information was collected by watching the seasons. Out of all the possible votes one could have received at tribal council, what percentage of those votes did they actually receive? For example, suppose that a contestant appears in only one tribal council before the merge and there are nine people in his/her tribe. If the contestant gets one vote, then the percentage of votes received pre-merge would be .125 since the contestant received one out of

the eight possible votes their tribe mates casted. This percentages is computed using only the votes that occur prior to a tie-breaker vote.

Percent of tribe going into merge: This information was collected by watching the seasons. This percentage is based on which tribe each contestant has been with the longest. Given that tribe, this variable is calculated by taking the number of people who are on that tribe and make the merge together, divided by the total number of people who make the merge. If there is a tie for which tribe a contestant stayed with the longest, then the percentage is based on the tribe that contestant was with first. We chose this tie-breaker by assuming that each contestant would be more loyal to the tribe they were first with, rather than the people they met later in the game.

Percent of challenges sat out of: This information was collected by watching the seasons. Tribe members typically only sit out of challenges if their tribe has more members than the other tribe(s). For that reason, some contestants do not have an opportunity to sit out of any challenges. In order to fill in this missing data, we looked at similar variables to make inference. The two variables we believe to be linked to the percentage of challenges sat out of are whether or not someone is considered as a leader to their tribe and whether they have been blamed for the loss of a challenge. For each contestant with a missing value for sitting out, we compared their scoring for the previous two categories with other people with the same scoring. We then took the average of what all those contestants had for the percentage of challenges sat out. We then used this value for the missing value. For example, anyone with a missing value who had not been blamed for the loss

of a challenge and was not a leader, received 25.033%. This is because, of everyone who didn't have a missing value for the percent of challenges sat out of who also were not a leader and had never been blamed for a challenge, had an average sit out percentage of 25.033. Other than the season in which there were missing values, there were only two other seasons with a special situation that we must make some clarifications about. During season 13, Nate was kidnapped by the other tribe and was forced to sit out of the challenge. (Burnett, 2006 ep.6) Because this decision was not made by him and his tribe, we did not count this towards his sit out percentage. The other exception was during season 19. On day one, tribes had to vote for a leader, knowing nothing about their tribe mates. (Burnett, 2009 ep.1) After leaders were selected, the leaders were then asked to pick the member on their tribe who they thought would be the smartest, the best swimmer, and the most agile member. These selections were all made before the tribe mates were given an opportunity to talk to each other. Therefore, the leader had to select these people based on first impressions. These contestants then had to compete in a challenge for their tribe. Everyone else sat out of the challenge. Because the selection of contestants who competed in the challenge took place before the tribes got to know each other, we excluded this challenge in the sit out percentage.

Percent of votes knew about: This information was collected by watching the seasons. In order to receive credit for knowing how the vote was going to go at tribal council, the constant had to have voted with the majority, or have agreed to split the votes in order to potentially flush out a hidden immunity idol. There was one

special case for this category, and it took place during season 12. During this season, the tribes were split up into four tribes of four members each. After only one tribal council, the four original tribes were brought to an end and two new tribes were chosen. Because only one contestant who made the merge was a part of that first vote, we felt like that tribal council did not significantly impact the game at all. Therefore, we left this tribal council vote out of the percentage (Burnett, 2006 ep.2).

Here is a list of the categorical variables we collected, along with their collection process.

Gender: This is pretty clear by watching the seasons.

Age: This information was collected by the CBS website (Survivor cast, 2015). Most seasons list the contestants' age. The seasons that do not state the age still state their dates of birth. For the seasons only with the dates of birth, we calculated their ages by taking the date filming began and subtracted their dates of birth.

State they are from: This information was collected by the CBS website (Survivor cast, 2015). The states in the U.S. were then broken into 9 regions which were taken off the U.S. Census Bureau website (Economic census, 2015).

Education level: This information was collected by the CBS website (Survivor cast, 2015). Everyone who did not have this information listed was assigned the High School Diploma degree, unless their occupation implied a higher level of education was required. If a higher level of education was implied, then we took the lowest possible degree and assigned it to that contestant. For example, in Season 18, Erinn was a hairdresser, yet no education was mentioned on the CBS website. For that reason, since many hairdressers go to a trade school, we

assigned her the trade school level or education. In Season 19, John was a rocket scientist. His highest education level received was also left off the CBS website, however since being a rocket scientist requires at least a Bachelor's degree, we assigned John the education level of Bachelor's degree. If CBS made a mentioning of the contestant going to college but did not specify that they graduated, then some college was implied.

Population of city from: The city each contestant lives in was collected by the CBS website (Survivor cast, 2015). Once the city information was collected, we used the 2010 census on the fact finders website (Fact finder,.n.d). We ran into a few special cases when we were collecting our data. The first was Kelly from season 3. Her current residence was not clear. Therefore, we went with her last known location which would be Durham, NC, since she went to college at Duke. The second special case was Jane from season 21. The CBS website stated that she was from Jackson Springs, NC, however we could not find a population for this city via the U.S. Census Bureau website. Instead, it said the town was unincorporated. We interpreted this to mean that the city was small, and hence we put it in the smallest population category. Kelly from season 1 was from Kernville, NV. For the city population, we used the population of Las Vegas. Amber from season 2 was from Beaver, PA. To find the population for this city, we had to go to the cities website (Business research reports, 2015). Zoe from season 4 was from Monhegan, ME. We had to go to that cities website as well to find the city's population (Welcome to Monhegan., 2015). Finally, there was Christy from season 6. She is from Basalt, CA. After trying to locate the city of

Basalt, we realized that Basalt is not actually a city. Instead, we went with the location of the basalt campgrounds which has an address in Gustine, CA. We then used the population of Gustine.

Marital status: Marital status was collected by the CBS website or was recorded if there was a mentioning of a spouse during the season (Survivor cast, 2015). If there was no mentioning of a spouse during the season or on the CBS website, then it was assumed that the contestant was single.

Occupation: This information was collected by watching the seasons. Each time a contestant does a one-on-one interview, their name appears on the screen along with their occupation. For the few contestants for whom this was not applicable, their occupation was found on the CBS website (Survivor cast, 2015).

In majority for first vote: This information was collected by watching the seasons. If a contestant voted with the majority of their tribe at the first tribal council they attended, then that contestant received a yes for this category. One exception to this would be if the majority of the tribe agreed to split the votes in case someone played a hidden immunity idol. The other special case for this category took place during season 12. During this season, the tribes were split up into four tribes of four members each. After only one tribal council, the four original tribes were brought to an end and two new tribes were chosen. Because the new tribes were chosen so soon, we based this category on the tribes that were formed on day 4 (Burnett, 2006 ep.2).

Blamed for loss of a challenge: This information was collected by watching the

seasons. If someone on their tribe said that it was his/her fault for losing the challenge, then this was recorded as being blamed.

Hidden immunity idol found: This was collected by watching the seasons. Not all seasons had hidden immunity idols. For the seasons without hidden immunity idols, every contestant received a no for this category.

Leader pre-merge: This was collected by watching the seasons. For a contestant to be considered as a leader before the merge, one of the contestant's tribe mates had to say the contestant was a leader or made any other similar comment.

One person trusted completely: This information was collected by watching the seasons. In order to get a marking of yes for this category, a contestant must have verbally confirmed that he/she trusted one person completely.

Were they lazy: This was collected by watching the seasons. To be considered lazy, one of the contestant's tribe mates had to complain about the lack of work that contestant did around camp.

Personality rub tribe wrong way: This was collected by watching the seasons. If a tribe mate complained about the contestant's personality, then this mark went against the contestant.

Did they get a nickname: This was collected by watching the seasons. To be considered a yes for this category, either a tribe mate, or Jeff Probst must call this contestant a nickname more than one time.

In majority for first vote after switch up: This information was collected by watching the seasons. Some seasons did not have tribal swaps. In order to fill in the missing data for these seasons, we based this category on whether or not they were in the

majority for the first tribal council they attended. If they were in the majority for their first tribal council, then we marked that contestant as a yes for this category as well.

Overheard information they were not supposed to: This was collected by watching the seasons. If a contestant overheard tribe mates talking about an alliance or future plan, then the contestant received this marking.

Same occupation as someone else: This information was collected by the CBS website (Survivor cast, 2015). If the contestant had the same occupation as another contestant who made the merge, then this was marked as a yes. Not all seasons had two people with the same occupation who made the merge. For these seasons, every contestant received a no for this category.

The data collection process has been completed. We now start constructing predictive models.

2 LOGISTIC REGRESSION

Logistic regression is used to analyze data where the dependent variable only has two possible outcomes. Predicted values from a logistic regression always lie between zero and one. This is due to the transformation process, called a logit. A logistic model will give a function $g(x) = a_0 + a_1x_1 + a_2x_2 + \cdots + a_nx_n$ where a_0 is the constant term, and $a_1, a_2, \cdots, a_n$ are the coefficients of the $x_1, x_2, \cdots, x_n$, respectively. To find the probability of success from this function, one computes the following:

$$P(success) = \frac{e^{g(x)}}{1+ e^{g(x)}}$$

(Hogg, McKean & Craig, 2013). Since $e^{g(x)}$ is always positive, then $P(success)$ is always greater than zero. The probability of success is capped at one due to the following:

$$\lim_{g(x)\to \infty} P(success) = 1.$$

Hence, $P(success)$ is always between zero and one.

When using a logistic regression, one must convert all categorical variables into dummy variables. If a certain categorical variable has n categories, then one must create $n - 1$ dummy variables. For example, we have five different categories for the education variable. In order to use this variable in a logistic regression, we had to convert the five categories into four dummy variables. Once the dummy variables are created, we ran the model.

2.1 ANALYSIS OF LOGISTIC REGRESSION RESULTS

To create the model, we used the built-in SAS (Statistical Analysis System) function PROC LOGISTIC. We used the forward selection which compares the dependent variable, winner, with each independent variable. The independent variable that has the highest correlation (highest F value) with the dependent variable is entered into the model first. After that, the next best variable is found similarly and entered into the model. This will be continued until no more variables meet the 0.05 significance level needed in order to be entered the model. To construct a predictive model, we used all 19 seasons of data, excluding season 24. Figure 2.1 shows an excerpt from the SAS output.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	3.4306	0.5682	36.4517	<.0001
Sex	1	-1.3355	0.6113	4.7726	0.0289
Blamed	1	-1.8205	0.8449	4.6426	0.0312
Sameocc	1	-1.4966	0.7205	4.3152	0.0378

Figure 2.1: The above figure shows the output for our PROC LOGISTIC command. Our logistic function can be constructed from this information.

This output gives us our logistic function. The logistic function is:

$$g(x) = 3.4306 - 1.3355 \cdot sex - 1.8205 \cdot blamed - 1.4966 \cdot sameocc.$$

To test the effectiveness of our model, we then used this function on the data from season 24. Table 2.1 shows the probabilities of winning for each contestant on season 24 of Survivor based on our logistic function.

Place	Name	Probability
1st	Christina	.968647
	Chelsea	.968647
	Kim	.968647
4th	Jonas	.890426
	Michael	.890426
	Jay	.890426
	Leif	.890426
	Troyzan	.890426
	Tarzan	.890426
10th	Kat	.833425
11th	Alicia	.528345
	Sabrina	.528345

Table 2.1: The above table shows the probability of winning for each contestant who made the merge in season 24 of Survivor, based on our logistic function g(x) above.

The winner of this season was Kim, who according to our model, was tied for the highest probability of winning. The problem with this model though, is that it only takes into account three different variables. Each of these variables only had two possible outcomes. That means, only eight different combinations of probabilities could have come from this model. Hence this model is not as effective as desired.

3 DIMENSION REDUCTION TECHNIQUES

When analyzing a large data set, there are frequently several variables of interest. The more variables that get introduced, the more complex the system becomes. Often times, many variables are related to one another. In order to reduce the number of variables using the hidden relationship among variables, we examined two dimension reduction techniques: Linear Discriminant Analysis and Principal Component Analysis.

3.1 LINEAR DISCRIMINANT ANALYSIS

Linear discriminant analysis is one form of dimension reduction. For this method, we are trying to best separate our contestants into two categories, winners and losers. We will do this using a linear combination of variables, which is similar to the logistic regression process. Because our variables are not normally distributed, nonparametric methods must be taken in order to estimate the group-specific densities.

We ran our model using the built-in SAS code PROC DISCRIM function which runs a variety of density estimation techniques. It also calculates a distance between the prior probabilities and the posterior probability (User's guide, n.d.). Prior probabilities are calculated based on what has happened in the past. Posterior probabilities are calculated based on conditional probabilities. These terms will be discussed in greater detail later in this paper (See Section 4.2). After a model has been built and a new set of data has been run to test its accuracy, the built-in function then calculates the probability that a certain input is misclassified as a winner or loser.

3.1.1 ANALYSIS OF LINEAR DISCRIMINANT ANALYSIS RESULTS

To construct our model, we used all 19 seasons of data, excluding season 24. We then used season 24 to test our model. Figure 3.1 shows an excerpt from the SAS output.

Number of Observations and Percent Classified into Winner			
From Winner	0	1	Total
0	5 45.45	6 54.55	11 100.00
1	1 100.00	0 0.00	1 100.00
Total	6 50.00	6 50.00	12 100.00
Priors	0.5	0.5	

Figure 3.1: The above figure shows how accurately linear discriminant analysis classifies the contestants on season 24 of Survivor who made the merge as winners and loser..

This output tells us that our model classified five out of the 11 loser accurately. However, it classified the true winner as a loser, and it classified six of the losers as potential winners. The linear discriminant method gave us large amounts of errors. Hence this method is not effective for our data set. Next, we try another form of dimension reduction, Principal Component Analysis.

3.2 PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis (PCA) is another form of dimension reduction. In order to use PCA on a data set, the data set must be centered; this means that the arithmetic means of all the variables must be zero. After converting a data set into a matrix, say X, one must center X by doing the following:

$$\tilde{X} = (I_n - \frac{1}{n} \cdot J_n) \cdot X$$

where n is the number of observations, hence, the number of rows in X, I_n is the identity matrix with n rows and n columns, and J_n is the $n \times n$ matrix each of whose entries is one.

Once the matrix has been centered, one can start creating the principal components. To do this, one must first construct the covariance matrix, C, which can be done as follows:

$$C = \frac{1}{n} \cdot \tilde{X}^T \cdot \tilde{X}$$

where $\tilde{X}^T$ is the transpose of $\tilde{X}$. From here, one should compute the eigenvalues of C, which are $[\lambda_1, \lambda_2, \dots, \lambda_l]$. Here, l is the number of variables collected. We then find their corresponding unit eigenvectors, $[\boldsymbol{u}_1, \boldsymbol{u}_2, \dots, \boldsymbol{u}_l]$. Once the unit eigenvectors are found, the principal component scores can be calculated as follows:

$$PrinScores = \tilde{X} \cdot [\boldsymbol{u}_1, \boldsymbol{u}_2, \dots, \boldsymbol{u}_l].$$

This will give all the principal component scores. I. J. Kim (personal communication, Month day, 2015).

3.2.1 ANALYSIS OF PRINCIPAL COMPONENT ANALYSIS RESULTS

To construct our model, we used all 19 seasons of data, excluding season 24. Season 24 was going to be used to test the model. When we first ran our data set through SAS, we used the built-in SAS function PROC PRINCOMP. When we did this, it took the 48 variables, and created 48 principal components. Ideally, one would want to select only a

few of these components to explain a large portion of the variance. Unfortunately, the first 12 components were able to account for only 50% of the variance amongst the variables. Figure 3.2 shows an excerpt from the SAS output.

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	3.21934947	0.55007130	0.0685	0.0685
2	2.66927818	0.34427334	0.0568	0.1253
3	2.32500483	0.22886322	0.0495	0.1748
4	2.09614162	0.07718527	0.0446	0.2194
5	2.01895634	0.16337554	0.0430	0.2623
6	1.85558080	0.14917415	0.0395	0.3018
7	1.70640665	0.04841136	0.0363	0.3381
8	1.65799529	0.09245341	0.0353	0.3734
9	1.56554188	0.06888342	0.0333	0.4067
10	1.49665846	0.03925085	0.0318	0.4385
11	1.45740760	0.00694253	0.0310	0.4695
12	1.45046507	0.07620212	0.0309	0.5004

Figure 3.2: The column on the right in the above figure shows what portion of the variance is explained by that variable in combination with the variables listed above it.

Due to these results, it is clear that principal component analysis on its own, would not lead to significant dimension reduction. Instead, we decided to try combining logistic regression with principal component analysis.

3.2.2 RESULTS WHEN COMBINING PRINCIPAL COMPONENT ANALYSIS WITH LOGISTIC REGRESSION

After the principal component analysis approach did not work on its own, we ran a logistic regression on the principal components that were significant. We did this using

the built-in SAS function for both principal component analysis and logistic regression which were stated before. By a significant principal component, we mean that its corresponding eigenvalue has to exceed 1. After adding on this restriction, we were left with 19 components. These 19 components explain 68% of the variance in our data set. To construct our model, we used all 19 seasons of data, excluding season 24. Figure 3.3 shows an excerpt from the SAS output.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.6236	0.3417	58.9556	<.0001
Prin5	1	0.4455	0.2014	4.8931	0.0270
Prin12	1	0.7388	0.2507	8.6824	0.0032

Figure 3.3: The above figure shows the output for our PROC LOGISTIC using the principal components as its new variables. Our logistic function can be constructed from this information.

This output gives us our logistic function. The logistic function is:

$$g(x) = -2.6236 + 0.4455 \cdot Prin5 + 0.7388 \cdot Prin12.$$

To test our model, we then used this function on the data from season 24. Table 3.4 shows the probabilities of winning for each contestant on season 24 of Survivor based on our logistic function.

The winner of this season was Kim, who according to our model had the third highest percentage of winning. Because the true winner of this season had a significantly lower probability to win than the contestant who placed first according to this model, this model was not that accurate either.

Place	Name	Probability
1st	Sabrina	.528929
2nd	Jonas	.338384
3rd	Kim	.313157
4th	Tarzan	.218149
5th	Alicia	.104381
6th	Chelsea	.095329
7th	Troyzan	.084679
8th	Michael	.074417
9th	Christina	.073575
10th	Kat	.051274
11th	Jay	.035508
12th	Leif	.023271

Table 3.4: The above table shows the probability of winning for each contestant who made the merge in season 24 of Survivor, based on our logistic function g(x) above.

4 MODEL CONSTRUCTION USING A BAYESIAN APPROACH

We now construct a model using a Bayesian approach. This means, we will be trying to maximize the chances of accurately predicting the winner of Survivor, by constantly updating our model with available information. In theory, our model should become more accurate with the addition of each new observation.

4.1 BAYES' THEOREM

The final model we construct applies Bayes' Theorem. Bayes' Theorem is formulated by using conditional probability. **Conditional probability** is defined to be the probability of event A occurring, given event B already occurred. Symbolically, this is represented as

$$P(A \mid B)$$

In order to calculate the conditional probability of A given B, one would need to know the probability of event B occurring, and the probability that event A and B would occur at the same time. Symbolically, these two are represented respectively as

$$P(B) \quad and \quad P(A \cap B)$$

Once these two pieces of information have been identified, one can find the conditional probability of A given B as follows

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

provided that $P(B) > 0$. Conditional probabilities must satisfy the following three properties:

1) $P(A \mid B) \geq 0.$

2) $P(\bigcup_{j=1}^{\infty} A_j \,|B) \;=\; \sum_{j=1}^{\infty} P(A_j \mid B)$ given A_1, A_2, A_3, ... are mutually exclusive events.

3) $P(B \mid B) \;=\; 1.$

Note that events are **mutually exclusive** if they are pairwise disjoint. Symbolically, this is represented as

$$P(A_i \cap A_j) \;=\; \emptyset$$

for all $i \;\neq\; j$. Note $P(A \mid B) \;=\; \frac{\mathrm{P(A \cap B)}}{P(B)}$ can be rearranged into

$$P(A \;\cap\; B) \;=\; P(B) \cdot \;P(A \mid B).$$

This rearrangement of terms is called the **multiplication rule.**

Now, let us take a deeper look at how one can find the probability of B. Let it be given that event A is formed by *n* mutually exclusive events, A_1, A_2, ... , A_n, where $P(A_j) > 0$. Hence, $A \;=\; \bigcup_{i=1}^{n} A_i$. Recall from earlier that $P(B) \;>\; 0$. Since A_1, A_2, ..., A_n are mutually exclusive, event B can only occur with one of the events A_j at a time. Hence

$$B \;=\; B \;\cap\; (A_1 \cup\; A_2 \;\cup \cdots \cup\; A_n)$$

$$=\; (B \;\cap\; A_1) \;\cup\; (B \;\cap\; A_2) \;\cup \cdots \cup\; (B \;\cap\; A_n).$$

Again, since A_1, A_2, … , A_n are mutually exclusive, then

$$P\;(B) \;=\; (B \;\cap\; A_1) \;+\; (B \;\cap\; A_2) \;+ \cdots +\; (B \;\cap\; A_n).$$

By the multiplication rule, $P\;(B \;\cap\; A_j) \;=\; P\;(A_j) \;\cdot\; P\;(B \mid A_j)$ for all j = 1, 2, …, *n*.

Therefore,

$$P\;(B) \;=\; P\;(A_1) \;\cdot\; P\;(B \mid A_1) \;+\; P\;(A_2) \;\cdot\; P\;(B \mid A_2) \;+ \cdots +\; P\;(A_n) \cdot\; P\;(B \mid A_n)$$

$$=\; \sum_{j=1}^{n} P\;(A_j) \cdot\; P\;(B \mid A_j).$$

Combining the definition of conditional probability and the law of total probability, we get **Bayes' Theorem** which states

$$P\,(A_j|\,B)\;=\;\frac{P\,(B\;\cap\;A_j)}{P\,(B)}\;=\;\frac{P\,(A_j)\cdot P\,(B\mid A_j)}{\sum_{j=1}^{n}P\,(A_j)\cdot P\,(B|A_j)}$$

(Hogg, McKean & Craig, 2013). The Naïve Bayes Classification uses Bayes' Theorem with an independence assumption among variables.

4.2 NAÏVE BAYES CLASSIFICATION

Naïve Bayes classification is based on both prior probabilities and likelihood probabilities in order to construct a posterior probability. **Prior probabilities** are calculated based on what has happened in the past. For example, when we construct a model using 18 seasons of Survivor, only 18 of the 182 contestants were able to win their season since only one person can win each season. Therefore, our prior probability of winning for each incoming contest that made the merge is 18/182. The prior probability of losing for each incoming contest that made the merge is 164/182. The **likelihood** of an event given winning or losing is calculated using the conditional probabilities discussed in Bayes Theorem. For example, the conditional probability of finding an idol, given the contestant was a winner is 3/18 ≈.167. The conditional probability of finding an idol, given the contestant was not a winner is 12/164 ≈ .073.

Because all of the variables collected are considered conditionally independent in a Naïve Bayes model, given any set of n variables, **x** = [x_1 x_2 ... x_n], the likelihood of getting a string of values X given a winner is

$$P(\mathbf{x} \mid winner) = P(x_1\, x_2 \ \ldots\ x_n \mid winner) = \prod_{i=1}^{n} P(x_i \mid winner)$$

and the likelihood of getting a string of values X given a loser is

$$P(\mathbf{x} \mid loser) = P(x_1\, x_2 \ \ldots\ x_n \mid loser) = \prod_{i=1}^{n} P(x_i \mid loser).$$

Therefore, applying Bayes' Theorem, the **posterior probabilities** can be found as follows:

$$P(winner \mid \mathbf{x}) = \frac{P(winner)\cdot\prod_{i=1}^{n} P(x_i \mid winner)}{P(\mathbf{x})}$$

and

$$P(loser \mid \mathbf{x}) = \frac{P(loser)\cdot\prod_{i=1}^{n} P(x_i \mid loser)}{P(\mathbf{x})}.$$

By considering the ratio of winning to losing,

$$\frac{P(winner \mid \mathbf{x})}{P(losing \mid \mathbf{x})} = \frac{P(winner)\cdot\prod_{i=1}^{n} P(x_i \mid \text{winner})}{P(loser)\cdot\prod_{i=1}^{n} P(x_i \mid \text{loser})}.$$

We can find the contestant with the highest ratio of winning to losing who would be the predicted winner. Note that in the computation of the ratio, it is not needed to compute $P(\mathbf{x})$. To find each contest's actual probability of winning in the Naïve Bayes model, we calculate

$$\frac{P(winner \mid \mathbf{x})}{P(winner \mid \mathbf{x}) + P(loser \mid \mathbf{x})}.$$

Before we apply Bayes' Naïve classification method to our data set, we first reduce the number of variables we were working with. First, we applied PCA to the seven continuous variables. These seven continuous variables were represented as [$x_1\ x_2$ ... x_7]. The values of x_1, x_2, ..., x_7 can be found in table 4.1

Instead of using the built-in SAS function like before, we worked the process out on our own using the steps described above in the principal component analysis section

Variable	What Variable Represents
x_1	Value of OrigAl
x_2	Value of TeamIm
x_3	Value of TeamRe
x_4	Value of SitOut
x_5	Value of votesknew
x_6	Value of tribeper
x_7	Value of votesrec

Table 4.1: The table above shows what each variable represents in the vector $[x_1\ x_2\ \cdots\ x_7]$.

of the paper (see Section 3.2). The code can be found in the Appendix. After creating these new components using the continuous variables, we selected the first two components to be part of our model, which account for over 69% of variance among the continuous variables of consideration.

The first principal component scores were calculated by taking the seven continuous variables $[x_1\ x_2\ \cdots\ x_7]$ for all 182 contestants. We first find the principal component that is the unit eigenvector [.0693884 .0440006 .9962552 .0042537 .231302 -.002798 -.012799]^T and then multiply this vector to the centered data matrix. The second principal component scores were obtained by using the unit eigenvector [.022694 .096413 .019339 .0161525 .9616313 .183538 .176447]^T.

Next, we had to calculate the conditional probabilities discussed above. To do this, we first quantize the continuous variables of consideration so that we can convert all the variables into categorical variables, described as follows. For the principal component scores that are continuous variables, we quantize the scores by creating three separate bins. These three bins were all created of equal length. Since all but one of the first principal component scores ranged from -0.5735 to 0.46221, each bin was of length 0.34524. There is one outlier score of 7.41396. We decided to exclude this value when

determining the equal length of each bin intervals. Hence, the first bin for the first principal component scores is the interval [-0.5735, -0.22826] and the second bin is the interval (-0.22826, 0.11697]. Since the third bin needs to include the score 7.41396, it is the interval (0.11687, 7.41396]. (Note that without the outlier 7.41396, the third bin would be the interval (0.11687, 0.46221].) We then analyzed the second principal component scores and created bins for them in a similar manner. Again, the second principal component scores were separated into three bins of equal length. The first bin contains the scores of [-0.9489, -0.52505]. The second bin contained the scores of (-0.52505, -0.1012]. The third bin contained the scores of (-0.1012, 0.32266]. We now

Explanation of new variables and what each category represents.			
Variable Name	**What Variable Stands For**	**Category Label**	**What Label Means**
Age	How old was the contest when show taped	1	Under 40
		2	40 and Above
Firstvotess	Did the contestant vote with the majority at the first tribal council he/she attended before and after the first tribal swap?	0	Contestant was not in majority for both of these tribal councils.
		1	Contestant was in majority for both of these tribal councils.
Lazy/personality	Was the contestant considered lazy or did his/her personality rub the tribe the wrong way?	0	Contestant was not considered lazy and personality did not rub others the wrong way.
		1	Contestant was considered lazy or personality rubbed others the wrong way.

Table 4.2: The table above gives an explanation as to how three of the new variables were constructed.

have the first two categorical variables in our model, obtains from the quantized principal component scores.

The next four variables are obtained from the categorical variables collected at the beginning of our research. The first of these variables we used was the hidden immunity idol category. The next three variables were variations of the original variables and can be found in Table 4.2.

Since all the variables are now categorical variables, we are ready to describe the model.

4.2.1 ANALYSIS OF NAÏVE BAYES' CLASSIFIER RESULTS

In order to create our model, we use the six new categorical variables discussed in the previous section. To calculate each $P(x_i|winner)$ and $P(x_i|loser)$, we compute the relative frequencies for each of these categories using all 19 seasons of data, excluding season 24. Figure 4.3 is a list of all the relative frequencies in our new model.

P(winner)	0.098901099			P(loser)	0.9011
P(Prin1=bin1 | winner)	0.222222222			P(Prin1=bin1 | loser)	0.22561
P(Prin1=bin2 | winner)	0.555555556			P(Prin1=bin2 | loser)	0.53049
P(Prin1=bin3 | winner)	0.222222222			P(Prin1=bin3 | loser)	0.2439
P(Prin2=bin1 | winner)	0			P(Prin2=bin1 | loser)	0.09756
P(Prin2=bin2 | winner)	0.166666667			P(Prin2=bin2 | loser)	0.15244
P(Prin2=bin3 | winner)	0.833333333			P(Prin2=bin3 | loser)	0.75
P(idol=0 | winner)	0.833333333			P(idol=0 | loser)	0.92683
P(idol=1 | winner)	0.166666667			P(idol=1 | loser)	0.07317
P(firstvotess=0 | winner)	0.166666667			P(firstvotess=0 | loser)	0.22561
P(firstvotess=1 | winner)	0.833333333			P(firstvotess=1 | loser)	0.77439
P(lazy/peronality=0 | winner)	0.833333333			P(lazy/peronality=0 | loser)	0.68902
P(lazy/peronality=1 | winner)	0.166666667			P(lazy/peronality=1 | loser)	0.31098
P(age=1 | winner)	0.833333333			P(age=1 | loser)	0.7622
P(age=2 | winner)	0.166666667			P(age=2 | loser)	0.2378

Figure 4.3: The above figure lists the relative frequencies of all our variables given the contestant was a winner and then given that the contestant was a loser.

After computing each of the conditional probabilities, we introduce the season 24 data in order to test our model. Note that before we convert the original values of the season 24 data into the categorical values corresponding to the categorical variables of consideration, we must center the season 24 data. To do this, we use the mean of each of the variables from our original data matrix X and subtract that value from each of the new data points from season 24. Once we get the new categorical variables, we use the conditional probabilities calculated previously in order to compute the winning percentage for each contestant. Figure 4.4 shows these results.

Just 24 data										
Prin1	Prin2	idol	irstvotessy/persona		age	Name	Numerator	Denominato	win/lose ratio	P(winner)
3	3	1.00	1	0	1.00	Kim	0.00176649	0.0049051	0.3601343	0.26478
3	3	1.00	1	0	1.00	Sabrina	0.00176649	0.0049051	0.3601343	0.26478
3	3	0.00	1	0	1.00	Chelsea	0.00883247	0.0621313	0.1421583	0.12446
2	2	0.00	0	1	1.00	Christina	0.00017665	0.0036116	0.0489123	0.04663
2	3	0.00	1	1	1.00	Alicia	0.00441624	0.0609904	0.0724088	0.06752
1	3	0.00	1	0	2.00	Tarzan	0.00176649	0.0179311	0.0985158	0.08968
3	3	0.00	1	0	1.00	Kat	0.00883247	0.0621313	0.1421583	0.12446
3	3	0.00	1	0	2.00	Troyzan	0.00176649	0.0193850	0.0911271	0.08352
1	3	0.00	1	1	1.00	Leif	0.00176649	0.0259384	0.0681034	0.06376
3	3	0.00	1	0	1.00	Jay	0.00883247	0.0621313	0.1421583	0.12446
3	2	0.00	0	0	1.00	Michael	0.0003533	0.0036791	0.0960283	0.08761
1	2	0.00	0	0	1.00	Jonas	0.0003533	0.0034032	0.1038144	0.09405

Figure 4.4: The figure above shows both the winning to losing ratio for each contestant that made the merge on season 24 of Survivor, along with their overall probability of winning.

To summarize this output, we ordered the contestants from the highest probability of winning to lowest one. These probabilities are listed in Table 4.5.

This model predicts the contestant who receives the highest winning probability as the winner. The true winner of the season, Kim, is tied for first place in this model. It should be noted that Sabrina, who is tied for first place in this model, received second place during this season. This is good evidence that the Naïve Bayes' Classifier in conjunction with Principal Component Analysis would be a reasonable predictive model for the show.

Place	Name	Probability
1st	Kim	0.26478
	Sabrina	0.26478
3rd	Chelsea	0.12446
	Kat	0.12466
	Jay	0.12466
6th	Jonas	0.09405
7th	Tarzan	0.08968
8th	Michael	0.08761
9th	Troyzan	0.08352
10th	Alicia	0.06752
11th	Leif	0.06376
12th	Christina	0.04663

Table 4.5: The table above shows the probability of winning for each contestant who made the merge in season 24 of Survivor, based on our Naïve Bayes' Classifier model above.

4.2.2 FURTHER ANALYSIS OF NAÏVE BAYES' CLASSIFIER MODELS

To further test our model, we used to most recent season of Survivor, season 31. Originally, we did not use season 31 in our research because all of the contestants were returning players. Each contestant was selected by the fans to get another chance at playing Survivor. Therefore, this season did not fit our original criteria that each contestant must be a first time player, no one was brought back into the game after being voted out, and no one knew each other before the game started.

Now that we built a Naïve Bayes' Classifier model that is accurate for seasons that meet these original criterion, we wanted to see if our model can also predict the winner of the seasons that did not fit the original criterion. For that reason, we choose the most recent season of Survivor to test our model with.

The data collection procedure was the same as mentioned in Section 1.2. The only unique scenario that happened in season 31 is that two contestants never went to tribal council before the merge. Both of these contestants were shown to be in a verbal agreement with other tribemates that they were in an alliance. For that reason, we assumed that these two contestants would have known how each vote at tribal council would have went down. Therefore, to fill in the missing values for the variable *Votesknew*, we gave each contestant a value of 1.

Once all the data was filled in, we ran our Naïve Bayes' Classifier model, just as we did in Section 4.2 and 4.2.1. Figure 4.6 shows the results.

Season 31	Prin1	Prin2	idol	irstvotes:	lazy/personality	age	Numerato	Denomir	win/lose ratio	P(winner)
Jeremy	2	3	1	1	0	1	0.00442	0.01067	0.413947456	0.29276
Tasha	3	3	0	1	0	1	0.00883	0.06213	0.142158271	0.124465
Spencer	1	3	0	0	0	1	0.00177	0.01674	0.105502413	0.095434
Kelley	2	3	1	1	0	1	0.00442	0.01067	0.413947456	0.29276
Keith	3	3	0	1	0	2	0.00177	0.01938	0.091127097	0.083516
Kimmi	1	3	0	1	0	2	0.00177	0.01793	0.09851578	0.089681
Abi-Maria	2	3	0	1	1	1	0.00442	0.06099	0.072408766	0.06752
Joe	3	3	0	1	0	1	0.00883	0.06213	0.142158271	0.124465
Stephen	1	3	0	1	0	1	0.00883	0.05747	0.153684617	0.133212
Ciera	3	3	0	1	0	1	0.00883	0.06213	0.142158271	0.124465
Kelly	1	2	0	0	0	1	0	0.00218	0	0
Andrew	3	2	0	0	0	2	0	0.00073	0	0
Kass	3	3	0	1	0	2	0.00177	0.01938	0.091127097	0.083516

Figure 4.6: The figure above shows both the winning to losing ratio for each contestant that made the merge on season 31 of Survivor, along with their overall probability of winning.

To summarize this output, we order the contestants from the highest probability of winning to lowest one. These probabilities are listed in Table 4.7.

This model predicts the contestant who receives the highest winning probability as the winner. The true winner of the season, Jeremy, is tied for first place in this model.

Place	Name	Probability
1st	Jeremy	0.29276
	Kelley	0.29276
3rd	Stephen	0.133212
4th	Tasha	.124465
	Joe	.124465
	Ciera	.124465
7th	Spencer	.095434
8th	Kimmi	0.089681
9th	Keith	0.083516
	Kass	0.083516
11th	Abi-Maria	0.06752
12th	Kelly	0
	Andrew	0

Table 4.7: The table above shows the probability of winning for each contestant who made the merge in season 31 of Survivor, based on our Naïve Bayes' Classifier model above.

This again confirms that the Naïve Bayes' Classifier in conjunction with Principal Component Analysis would be a reasonable predictive model for the show.

5 CONCLUSION

Throughout this paper, we have looked at several predictive modeling techniques. Along the way, we had to refer back to principal component analysis as the dimension reduction technique in order to reduce the number of variables we used on our model. Eventually, we were able to construct a predictive model using Naïve Bayes' Classification along with Principal Component Analysis.

5.1 FUTURE WORK

In order to better test our model, we would like to apply our predictive model to other seasons of Survivor in order to see if we can get similar results. If we want to test our model on another season of Survivor, then we would have to apply our model to a season in which not all contestants are first time players, at least one player returned to the game after being voted off, or all contestants knew at least one other contestant on the show before the show began filming. This is because no other season of Survivor has aired that excludes all of the above restrictions. Because these variables may have a significant impact on the outcome of the season, we feel as though applying our model to seasons 7, 8, 11, 16, 20, 22, 23, 25, 26, 27, and 29 may not give us as consistent results. However, based on the results for season 31, our model may have just as much predictive power with these seasons. To better determine the predictive power of our model, we could apply the remaining seasons of Survivor to our model.

6 APPENDIX

6.1 DATA TABLE

Explanation of variables and what each category represents.			
Variable Name	**What Variable Stands For**	**Category Label**	**What Label Means**
Winner	Did this contestant win?	0	No
		1	Yes
Sex	Is the contestant a male or female?	0	Female
		1	Male
Age	How old was the contestant when show taped	1	Age 29 and younger
		2	Age 30-39
		3	Age 40-49
		4	Age 50 and older
State	What states is the contestant from?	1	CT, ME, MA, NH, RI, VT
		2	NJ, NY, PA
		3	IL, IN, MI, OH, WI
		4	IA, KS, MN, MO, NE, ND, SD
		5	DE, DoC, FL, GA, MD, NC, SC, VA, WV
		6	AL, KY, MS, TN
		7	AR, LA, OK, TX
		8	AZ, CO, ID, MT, NV, NM, UT, WY
		9	AK, CA, OR, HI, WA
Edu	What is the highest level of education achieved?	1	High School
		2	Some College
		3	Trade School/Associates Degree
		4	Bachelors

		5	MA, MS or PhD
Pop	What is the population of the city the contestant lives in?	1	4,999 and under
		2	5,000-14,999
		3	15,000-54,999
		4	55,000-149,999
		5	150,000-499,999
		6	500,000 and over
Mar	Is the contestant married?	0	No
		1	Yes
Occ	What is the contestants' occupation?	1	Anything in the medical field
		2	Student
		3	Military & Protective Services
		4	General Labor (Construction worker, athlete, etc.)
		5	Anything in the business field
		6	Science/Technology
		7	Social Sciences (entertainer, model, designer, etc.)
		8	Service Jobs (hairstylist, waiter, photographer, etc.)
		9	Teacher, Coach, Principal
firstvote	Did the contestant vote with the majority at the first tribal council he/she attended?	0	No
		1	Yes
OrigAl	What percent of the contestants' tribe was he/she in an alliance with?	Percentage in decimal form	

TeamIm	What percent of the team immunity challenges did the contestant win?	Percentage in decimal form	
TeamRe	What percent of the team reward challenges did the contestant win?	Percentage in decimal form	
SitOut	What percent of the challenges did the contestant sit out of when the oFpresented itself?	Percentage in decimal form	
Blamed	Was the contestant blamed for the loss of a challenge?	0	No
		1	Yes
Idol	Did the contestant find a hidden immunity idol?	0	No
		1	Yes
Leader	Was the contestant considered a leader by their tribe mates?	0	No
		1	Yes
Trust	Did the contestant completely trust at least one other contestant?	0	No
		1	Yes
Lazy	Was the contestant considered lazy by their tribe mates?	0	No
		1	Yes
Person	Did the contestants' personality rub at least one of their tribe mates the wrong way?	0	No
		1	Yes
Nickname	Did the contestant receive a nickname by either his/her tribe mates or Jeff Probst?	0	No
		1	Yes
Heard	Did the contestant overhear a conversation their tribe mates	0	No

	were secretly having?		
		1	Yes
firstvotesw	Did the contestant vote with the majority at the first tribal council after a tribal swap that he/she attended?	0	No
		1	Yes
Votesknew	What percent of tribal council results did the contestant know was going to happen?	Percentage in decimal form	
Tribeper	What percent of the people who made the merge were on the contestants' tribe the longest?	Percentage in decimal form	
Votesrec	What percent of votes did the contestant receive that he/she could have been received?	Percentage in decimal form	
Sameocc	Did any of the other contestants have the same occupation as the contestant of interest?	0	No
		1	Yes

6.2 NAÏVE BAYES CLASSIFIER AND PCA CODE IN SAS

```
*** Converting all my excel files into matrices in SAS;

proc iml;

use Survivor.All1s182; read all into Jn;

use Survivor.No24nocategorical; read all var {OrigAl TeamIm TeamRe SitOut

votesknew tribeper votesrec} into X;

*** centering my origional matrix X must use

([nxn identity matrix - (1/n)*(nxn matrix of all ones)]*X
```

```
where n is number of competitors, X is my original nxc matrix,
and c is number of continuous variables;
Tildax = (I(182) - (1/182)*(Jn))*X;
print Tildax;
***creating sas data set out of matrix tildax;
*** I couldn't get this to work so I copied the matrix tildax into an excel fie
and imported it as Survivor.Centeredno24;
***constructing covariance matrix, then finding maximum eigenvalue.;
Tildaxtrans= Tildax`;
C=(1/182)*Tildaxtrans*Tildax;
print C;
***finding eigenvalues and eigenvectors using built-in sas code;
Eval=eigval(C);
print eval;
Evect=eigvec(C);
print Evect;
***Checking to make sure eigenvectors are right;
Check1= (C- 0.3592926#I(7))*{.0693884, .0440006, .9962552, .0042537, .0231302, -
.002798, -.012799};
Print Check1;
Check2= (C- 0.1008337#I(7))*{-.022694, -.096413, -.019339, .0161525, .9616313, -
.183538, -.176447};
Print Check2;
```

```
*** Checking to see if these Eigenvectors were UNIT Eigenvectors;
unit1=
sqrt(.0693884*.0693884+.0440006*.0440006+.9962552*.9962552+.00425378*.004253
7+.0231302*.0231302+(-.002798)*(-.002798)+(-.012799)*(-.012799));
print unit1;
unit2=
sqrt(.022694*.022694+.096413*.096413+.019339*.019339+.0161525*.0161525+.96163
13*.9616313+.183538*.183538+.176447*.176447);
print unit2;
***Create Principal Components;
Princomps=Tildax*Evect;
Print Princomps;
***Centering Continuous Variables from Season 24
by first finding means of original matrix, then subtracting these from new matrix;
proc means data=Survivor.No24nocategorical;
        var OrigAl TeamIm TeamRe SitOut votesknew tribeper votesrec;
run;
proc iml;
Meanno24={0.4200331 0.5716248 0.5708464 0.2506364 0.8260073 0.5499362
0.0664145,
0.4200331 0.5716248 0.5708464 0.2506364 0.8260073 0.5499362 0.0664145,
0.4200331 0.5716248 0.5708464 0.2506364 0.8260073 0.5499362 0.0664145,
0.4200331 0.5716248 0.5708464 0.2506364 0.8260073 0.5499362 0.0664145,
```

```
0.4200331 0.5716248 0.5708464 0.2506364 0.8260073 0.5499362 0.0664145,
0.4200331 0.5716248 0.5708464 0.2506364 0.8260073 0.5499362 0.0664145,
0.4200331 0.5716248 0.5708464 0.2506364 0.8260073 0.5499362 0.0664145,
0.4200331 0.5716248 0.5708464 0.2506364 0.8260073 0.5499362 0.0664145,
0.4200331 0.5716248 0.5708464 0.2506364 0.8260073 0.5499362 0.0664145,
0.4200331 0.5716248 0.5708464 0.2506364 0.8260073 0.5499362 0.0664145,
0.4200331 0.5716248 0.5708464 0.2506364 0.8260073 0.5499362 0.0664145,
0.4200331 0.5716248 0.5708464 0.2506364 0.8260073 0.5499362 0.0664145};
use Survivor.Catonly24; read all var{OrigAl TeamIm TeamRe SitOut votesknew tribeper
votesrec} into matrix24;
centered24= matrix24-meanno24;
use Survivor.All1s182; read all into Jn;
use Survivor.No24nocategorical; read all var {OrigAl TeamIm TeamRe SitOut
votesknew tribeper votesrec} into X;
Tildax = (I(182) - (1/182)*(Jn))*X;
Tildaxtrans= Tildax`;
C=(1/182)*Tildaxtrans*Tildax;
Evect=eigvec(C);
***Creating Principal Components of Season24;
PrinComps24=centered24*Evect;
print PrinComps24;

***Testing Season 31;
```

```
proc iml;
Meanno24={0.4200331 0.5716248 0.5708464 0.2506364 0.8260073 0.5499362
0.0664145,
0.4200331 0.5716248 0.5708464 0.2506364 0.8260073 0.5499362 0.0664145,
0.4200331 0.5716248 0.5708464 0.2506364 0.8260073 0.5499362 0.0664145,
0.4200331 0.5716248 0.5708464 0.2506364 0.8260073 0.5499362 0.0664145,
0.4200331 0.5716248 0.5708464 0.2506364 0.8260073 0.5499362 0.0664145,
0.4200331 0.5716248 0.5708464 0.2506364 0.8260073 0.5499362 0.0664145,
0.4200331 0.5716248 0.5708464 0.2506364 0.8260073 0.5499362 0.0664145,
0.4200331 0.5716248 0.5708464 0.2506364 0.8260073 0.5499362 0.0664145,
0.4200331 0.5716248 0.5708464 0.2506364 0.8260073 0.5499362 0.0664145,
0.4200331 0.5716248 0.5708464 0.2506364 0.8260073 0.5499362 0.0664145,
0.4200331 0.5716248 0.5708464 0.2506364 0.8260073 0.5499362 0.0664145,
0.4200331 0.5716248 0.5708464 0.2506364 0.8260073 0.5499362 0.0664145,
0.4200331 0.5716248 0.5708464 0.2506364 0.8260073 0.5499362 0.0664145};
use Survivor.Season31; read all var{OrigAl TeamIm TeamRe SitOut votesknew tribeper
votesrec} into matrix31;
centered31= matrix31-meanno24;
use Survivor.Identity182; read all into In;
use Survivor.All1s182; read all into Jn;
use Survivor.No24nocategorical; read all var {OrigAl TeamIm TeamRe SitOut
votesknew tribeper votesrec} into X;
Tildax = (In - (1/182)*(Jn))*X;
```

```
Tildaxtrans= Tildax`;

C=(1/182)*Tildaxtrans*Tildax;

Evect=eigvec(C);

***Creating Principal Components of Season31;

PrinComps31=centered31*Evect;

print PrinComps31;
```

BIBLIOGRAPHY

Burnett, M. *Survivor.* New York, NY: Central Broadcasting Service.

Business research reports. (2015). Retrieved from http://www.beaverpa.us/documents/Demographic_Business_Report.pdf

Economic census. (2015). Retrieved from http://www.census.gov/econ/census/help/geography/regions_and_divisions.html

Fact finder. (n.d). *American FactFinder - Community Facts.* Retrieved from http://factfinder.census.gov/faces/nav/jsf/pages/community_facts.xhtml

Hogg, R., McKean, J., & Craig, A. (2013). *Introduction to mathematical statistics* (7th ed.). Boston, MA: Pearson.

Naïve Bayes classifier. (2015). Retrieved from http://documents.software.dell.com/Statistics/Textbook/Naive-Bayes-Classifier

Survivor cast. (2015). Retrieved from http://www.cbs.com/shows/survivor/cast/

User's guide (n.d.). *SAS/STAT(R) 12.1 user's guide.* Retrieved from http://support.sas.com/documentation/cdl/en/statug/65328/HTML/default/viewer.htm#statug_discrim_overview.htm

Welcome to Monhegan. (2015). Retrieved from http://monheganwelcome.com/

Yahr, E., Moore, C., & Chow, E. (2015, May 29). How we went from 'Survivor' to more than 300 reality shows: A complete guide. *The Washington Post.* Retrieved from www.washingtonpost.com/graphics/entertainment/reality-tv-shows/