5-8-2018

Exploring the Use of Predictive Analytics in Banking and Finance Decision-Making

Melanie Tummino

Exploring the use of Predictive Analytics in Banking and Finance Decision-Making

Melanie Tummino

Submitted in Partial Completion of the
Requirements for Commonwealth Honors in Mathematics

Bridgewater State University

May 8, 2018

Dr. Wanchunzi Yu, Thesis Advisor
Dr. Vignon Oussa, Committee Member
Dr. John Pike, Committee Member

Exploring the use of predictive analytics in banking and finance decision-making

Melanie Tummino

May 8, 2018

Abstract

Predictive analytics is a branch of advanced analytics that is composed of various statistical techniques where each contributes in making predictions about future scenarios and outcomes. Some of these techniques include machine learning, artificial intelligence, data mining, predictive modeling, logistic regression, etc., and the patterns found in the results can be used to identify risks and opportunity. Predictive analytics is often associated with meteorology and weather forecasting due to the fact there are many attributes to contribute to a response, but generally, it has many applications in existing growing or established businesses, especially when it comes to decision-making about revenue, customers, and productivity (Siegel, 2016). This project is focused on the banking and financing area, and I analyzed two different datasets – one in which I generated data using software, and the other from a Portuguese bank that is available to the public online. The purpose of this project was to create a list of targeted customers that are more likely to sign up for a credit card and more likely to be issued a checking account as the binary responses by using predictive analytics. I investigated the relationship between the binary response, and the predictor variables, the characteristics of customers within the dataset. Analysis procedures and logistic regression are employed which allowed me to create the most accurate model for targeting these specific group of customers.

1 Introductory Information

Predictive analytics is most commonly used in banking and finance, especially when making decisions as to which customers a promotion should be given to. One of the main fields to focus on when using predictive analytics in a business will reside within the customer databases considering a business can't exist without its customers. To make decisions based on productivity, a company tends to focus on the existing, content customers to ensure that each decision made is in light of them. A common prediction in this field, and also the topic in which my research is focused on, is which customers a bank should target in order to increase profit. I used a number of predictive variables based on characteristics presumed to be most predictive in the analysis to target these specific customers. The statistical techniques implemented allow for more strategic and profitable decisions that are based on specific predictor variables that correlate to a binary response (Siegel, 2016). First, I generated my own data given these conditions to complete the analysis and integrate results, and then I implemented the same analysis procedures to a real-life bank dataset from Portugal. In addition, I compared and tested the performance of different models using statistical software including R and SAS, and BigML is used for the analysis and visualization aspect of the research.

2 Methodology

Although there are multiple statistical techniques to put into consideration while using predictive analytics, my research is focused on three main techniques, which include logistic regression, predictive modeling, and machine learning. I learned how to simulate data, consider which variables in action, identify the principle analysis and create a logistic regression model.

I used the logistic regression model,

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_i x_i \quad (1)$$

to analyze the data, where p is the probability that the outcome will occur, the βi's are the estimators, and the Xi's are the predictor variables. The binary response Y, whether or not a customer is more likely to sign up for a credit card or more likely to be issued a checking account in the logistic regression model follows a binomial distribution. We assume a normal distribution, the dependent variable is dichotomous in nature, there are no outliers, there are no high intercorrelations among the predictors (no two variables are the same) and there is a linear relationship between the odds ratio, $\log \frac{p}{1-p}$ and each predictor variable (Johnson & Wichern, 2007).

In this study, I had several parameters, which included my binary response and predictor variables. The binary response is the list of customers that will or will not sign up for a credit card or should or should not be issued a checking account, and the predictor variables are the characteristics within the dataset such as age, GPA, major, sex, education, job, account balance, amount of hours working per week, etc. There is a connection between these variables because the predictor variables are what influence the binary response. The model follows a normal distribution in both datasets, meaning the error terms are normal, and the estimators are generated using a parameter estimate table, and additionally through R and SAS after implementing the statistical technique of Principle Components Analysis, or PCA (this is explained in more detail later). Predictive modeling is used to create a model given the variables that are most likely to influence the binary response, or outcome. I built different predictive models based on the data reduction procedure of PCA. Machine learning is an application of artificial intelligence that provides a system the ability to identify patterns and make logical predictions without human intervention (Tan, Steinbach, & Kumar, 2006). These models are analyzed and compared using three types of machine learning software: R, SAS and BigML. After drawing conclusions and making interpretations, the goal was to

choose the most useful predictive model for the analysis.

3 Simulation Studies and Results

The first step was to generate my own data using R and an example dataset from the textbook, *Multivariate Normal Regression Analysis*, which contained 170 observations. This example dataset guided me in creating my own dataset by using the same binary response and chosen predictor variables. I generated both categorical and quantitative variables, where the categorical variable was generated with *c* classes and *c-1* dummy variables (Agresti, 2007), and the quantitative variables followed a normal distribution with mean and standard deviations following the example from the textbook. I calculated accurate means and standard deviations for these variables using the information in the example dataset. I generated one categorical variable for major (business, humanities, sociology or science), and four quantitative variables for age ranging from 20-27, GPA ranging from 1.5-4.5, hours of work ranging from 0-30 hours per week, and sex (although sex is considered a categorical variable, I generated it as a quantitative variable for simplicity reasons within the R code, where 1 signified female and 0 signified male), and then of course, the binary response, or the list of customers that will or will not sign up for a credit card. The final dataset contained a total of 1,000 observations and six parameters. Once the dataset was generated, I created a regression model in R using the "glm()" function, or general linear model function and the parameter estimate table to help with my estimates. Please refer to Table 1 for the table of estimates found in the textbook, *Multivariate Normal Regression Analysis*.

The LOGISTIC Procedure	
Parameter	Estimate
Intercept	33.2158
NSEX	-0.6174
DUM_MAJ1	0.8180
DUM_MAJ2	5.4874
DUM_MAJ3	5.0109
GPT	-3.9438
AGE	-0.5519
HRS	-1.2542

Table 1: Parameters table, where FEMALE is considered the referral and NSEX = MALE, and BUS is considered the referral and DUM_MAJ1 = HUM, DUM_MAJ2 = SCI and DUM_MAJ3 = SOC.

Estimators and Corresponding P-values : R vs SAS				
Parameter	R	P-value	SAS	P-value
Intercept	35.28885	$1.32e^{-}12$	38.0312	<.001
Age (20-27)	-0.72808	$1.89e^{-}05$	-0.7280	<.001
Hours (0-30)	-1.14731	$< 2e^{-}16$	-1.1472	<.001
GPA (1.5-4.5)	-3.68707	$3.42e^{-}15$	-3.6867	<.001
HUM	0.70867	.0978	-2.1674	<.001
SCI	5.80275	$2e^{-}16$	2.9290	<.001
SOC	4.98266	$2e^{-}16$	2.1090	<.001
MALE	-0.25672	0.0478	-0.1284	0.0478

Table 2: Estimates and p-values of both models generated through R and SAS.

Generating this data took a lot of trials and errors within the model to ensure the estimates

were close enough to the estimates in the parameter estimate table, and that the model considered a fair amount of the variables to be significant. Through software such as R and SAS, I was able to determine which variables had the most significant impact on the model by looking at its corresponding p-value. If its p-value is very close to 0, then the variable is significant and should remain in the model, but if the p-value is closer to 1, then it should be removed from the model to make it stronger. Please refer to Table 2 for the p-values of each variable generated in R and SAS. I uploaded the same exact dataset to both R and SAS and generated linear models in both to compare. Each of the model's estimates were the same except for the estimates of the categorical variables, which you will notice are slightly different. Refer to Table 2 for a comparison of estimates generated through R and SAS. The discrepancy isn't obvious, but after experimentation between the models, I chose to include the estimates generated through R because the values were closer to those of the estimate table, which was essential. Once I was satisfied with this decision, I exported the data as an Excel file which made the dataset easier to read and interpret. Through pivot tables created in Excel, I discovered an idea as to how I could organize the data to make the decision-making aspect of the model more simple. For example, variables that had a wide range of data contained in them such as age or hours of work per week, I grouped into buckets to cut down decision-making options. To my surprise, the model was stronger without buckets, so I decided to keep the data the way it initially was. It was discovered that existing clients who are credit card users within the dataset did not work more than 14 hours per week, had a GPA that is less than a 3.0, and had a major in either Science or Social Science. These three variables were considered most significant within the models generated through R and SAS due to their very small p-values, and this was later confirmed through BigML as well. These three attributes affected the logistic regression model (Equation 1) most, hence making the strongest model for the project.

BigML is a cloud-based predictive analytics software that creates a visualization of the pos-

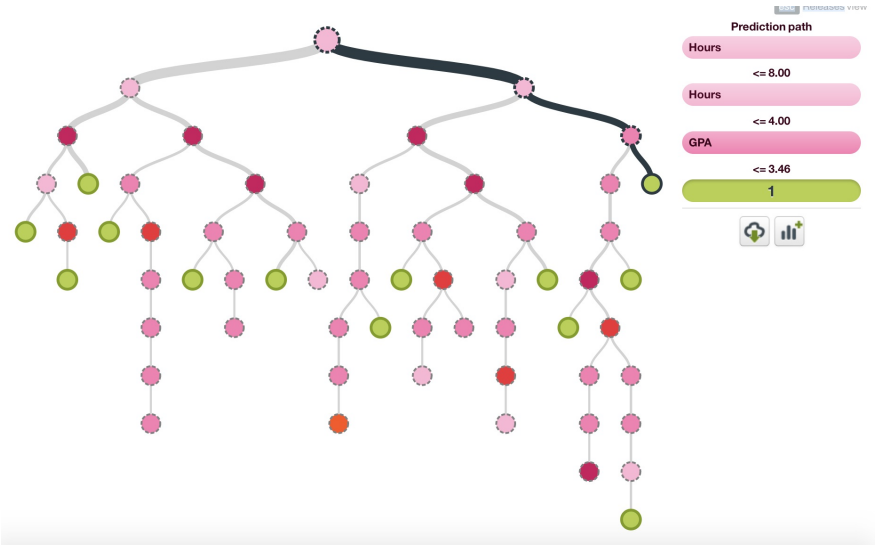

Figure 1: Screenshot of the tree diagram for the generated dataset that was generated through BigML. The bold path that leads to a green dot is the path that the model took in order to predict a specific group of customers as a "1." The individual steps that were taken in that exact instance are shown in the far right upper corner.

sible scenarios and outcomes within a dataset, utilizing the method of a tree diagram. This visually helps in understanding the process of the requirements a customer must meet in order to be predicted as more likely to sign up for a credit card while following a tree of decisions and predicted outcomes (1: customer will most likely sign up for a credit card, and 0: customer will most likely not sign up for a credit card). Please refer to Figure 1 to see a tree diagram of the generated dataset using BigML. This software also includes a model summary report, which behaves similarly to a summary in R and SAS, but instead of a numeric value to indicate the significance of the variable, it displays a bar chart. Each bar represents the percentage of the predictor variable being used within the model's decision-making, thus allowing for interpretation as to which predictor variables are more/less likely to influence the binary response. The variable with the most significant percentage will have the largest bar, thus influencing the model with higher significance than the variables that follow with smaller percentages. Please refer to Figure 2 for a screenshot of the summary

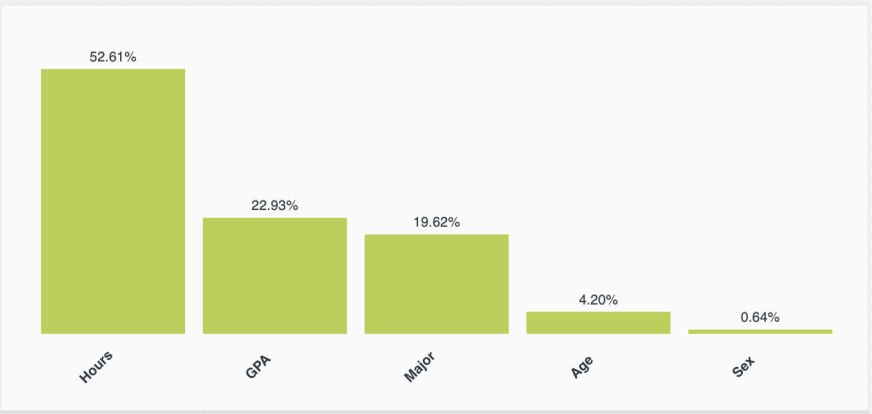

Figure 2: Screenshot of the summary report for generated dataset generated through BigML.

report generated through BigML. It was concluded that the amount of hours spent working per week, GPA and major were most significant within the decision-making of the model. Additionally, BigML can split the dataset into "test/training" datasets, meaning it separates 80% of data as the "test" and 20% of the data as the "training," and then it runs an evaluation against both of the models. This process provides honest assessments of the performance of the predictive model. The evaluation includes an accuracy percentage, which is determined by precision and recall and allows for interpretation as to how strong the model is, or how accurately it can predict outcomes. Precision is defined by the number of true positives over the total number of positive predictions, and recall is defined by the number of true positives over the number of positive instances. The number of false positives, false negatives, true positives and true negatives, logistic regression graphs, etc. are also given through the evaluation.

It was not surprising that the model had a very high accuracy percentage considering this data was generated using very accurate numbers from a textbook. Within the model, the precision percentage was a 91.5% and the recall percentage was a 90.8%, meaning that the model had a meager count of false positives and negatives, hence the model was more correct than incorrect while deciding whether or not a customer would sign up for a credit card. The

dataset can later be downloaded as a batch file through Excel, in which all 1000 observations and predictions can be viewed. In conclusion, if this dataset was obtained from an actual bank and they wanted to better understand which customers were more likely to sign up for a credit card, I would recommend that they pay close attention to the customers who work 14 hours per week, have a GPA that is less than a 3.0, and a major in either Science or Social Science.

4 Real-Life Application and Results

The real-life dataset from a Portuguese bank was available to the public and included a heavy amount of data, which is why I decided to use it for the real-life application part of my research. This dataset contains 45,211 observations and has a total of 17 parameters. There are seven quantitative variables including age, balance, day, duration, campaign, pdays and previous, and nine categorical variables including job, marital, education, default, housing, loan, contact, month and poutcome, and the binary response is whether or not a customer should be issued a checking account. Refer to Table 3 to see the statistics of the quantitative variables including the mean and standard deviations of each of the seven variables. This dataset is not only a lot larger than the dataset I previously generated, but there's a lot more variability due to the fact that it is real-life data. Hence it will be a lot more challenging to analyze. Although the two datasets are very different, similar analysis procedures took place when determining the binary response.

	Age	Balance	Day	Duration	Campaign	Pdays	Previous
Mean	41.17	1422.66	15.92	263.96	2.79	39.77	0.54
Standard Deviation	10.58	3009.64	8.25	259.86	3.11	100.12	1.69

Table 3: Means and standard deviations of each of the seven quantitative variables in the

Portuguese bank dataset.

At first, I went through all of the attributes in the dataset to make sure I understood what each meant. After eliminating the attributes I did not find useful, I also eliminated the ones I found to be insignificant in the decision-making of the model. The main goal was to reduce the number of variables considering there was a very large amount of them and identify the more significant ones. I was able to eliminate a total of four, which brought the count of predictor variables down to twelve.

Due to the increased amount of quantitative variables in this dataset in comparison to my generated dataset, I had to take a different approach. I decided to implement a new statistical technique called Principal Component Analysis (PCA), a dimension-reduction tool that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables, called principal components, or PCs (Johnson & Wichern, 2007). This technique is specifically used for reducing the number of quantitative variables only, in which my dataset has a total of seven. PCA allows one to express the data matrix in a new coordinate system (i.e. defining new variables as linear combinations of the originals) in which the variables are uncorrelated and arranged in order of 'importance' in the sense of observed variance. This new coordinate system is given by an orthonormal basis of eigenvectors of the correlation matrix, and the dimension reduction is accomplished by projecting onto the top eigenspace. The process of PCA included turning each quantitative variable into a PC, and I then used the PCs to generate another regression model by using the set of generated eigenvectors as the 'transformed' estimators in each PC for my predictor variables, along with the categorical variables and binary response. If I chose to make the regression model using all seven PCs, no information about the data would be lost or omitted, which would defeat the purpose of this technique – to reduce the number of variables. This is done by looking at the cumulative proportion,

which is the amount of the information being used in the dataset. The number of PCs used depends on the cumulative proportion, in which generally speaking, it is considered "good" to lose around 20% of the data, so the cumulative proportion should be approximately .80 (Johnson & Wichern, 2007). Based off this idea, my first thought was to use five PCs, but it was discovered that too much data was omitted, so it wasn't clear which PCs were significant in the model. I then decided to increase the number of PCs to six, which made the cumulative proportion about .94 (meaning 6% of the data was omitted), and this allowed for more clarity in terms of which variables were more significant in the model. Please refer to Figure 3 for a list of coefficients and corresponding cumulative proportions of the five PCs. Although I was only able to eliminate one of the quantitative variables, this still contributed to reducing the number of variables, which satisfied the goal.

Eigenvalues of the Correlation Matrix				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	1.63401450	0.49913921	0.2334	0.2334
2	1.13487529	0.04922252	0.1621	0.3956
3	1.08565277	0.10839297	0.1551	0.5506
4	0.97725980	0.06192206	0.1396	0.6903
5	0.91533774		0.1308	0.8210

Figure 3: Screenshot of the eigenvalues and their corresponding cumulative proportions generated through SAS.

I implemented the PCA function in both R and SAS to compare the strength of the models using each of the eigenvectors that the software generated, and I realized the model became

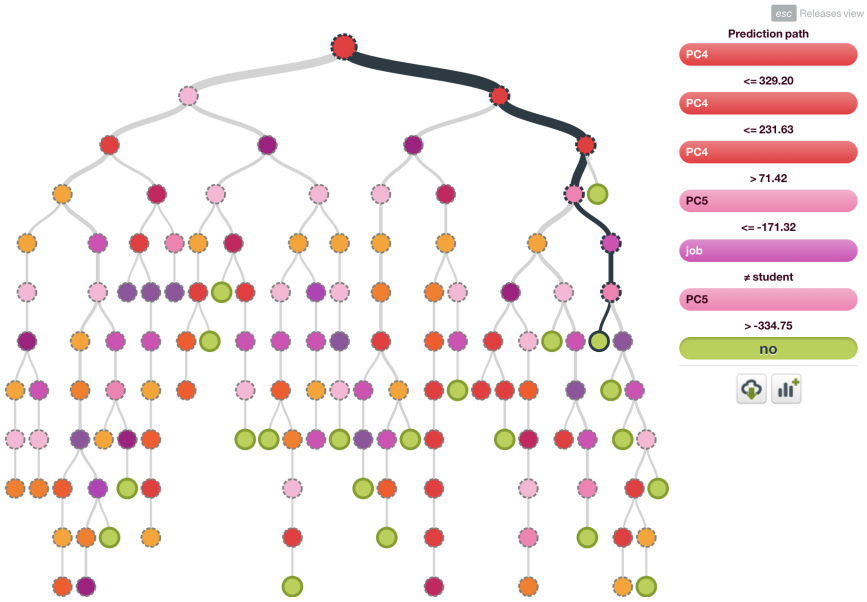

Figure 4: Screenshot of the tree diagram for the Portuguese bank dataset generated through BigML.

stronger when using the SAS eigenvectors. Again, the reasoning behind this discrepancy isn't clear. It is possible that it could be because R has multiple PCA methods without any original code behind the function provided, whereas SAS only has one method and allows users to see the process behind the original code, making it more detailed and clear as to how these eigenvectors are being generated. SAS is also a paid software whereas R is not, so some might argue the software is stronger in some aspects due to this. Although there wasn't much variable reduction, PCA still helped in eliminating at least one of the quantitative variables. It was discovered that components, or PCs 4, 6, and 1 were the strongest factors in the model. This helped in understanding why the model wasn't clear in variable significance when there were only five PCs – considering PC6 is such a significant variable, it is crucial to have it in the model.

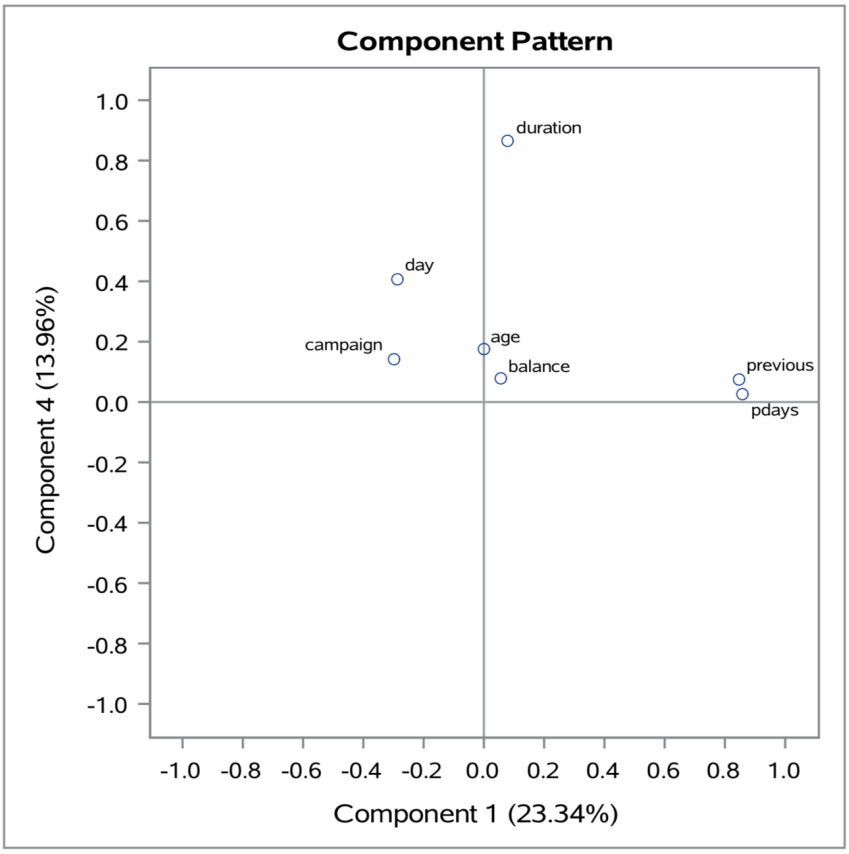

Figure 5: One of the many graphs SAS provides within its Procomp procedure. The variables closer to the center (closer to 0) have more significance in rotating the variable space.

Once I finished the analysis procedures for reducing the number of variables, I uploaded the manipulated dataset (the dataset including the PCs) to BigML. The results find that balance, age, and job are most significant in banking and finance decision-making when pertaining to this particular real-life application. The summary report in BigML (refer to Figure 6) showed that PC4, PC6, PC1, and job were the most significant variables in the decision-making of the model (each of these variables also had the smallest p-values). This also confirmed the reasoning behind why the tree diagram in Figure 4 uses the following PCc as its first determining factors when evaluating the likeliness of a customer being issued a credit card (as seen in the bold path). SAS showed that each of the PCs used quantitative variables, balance and age most in determining the rotated variable space. This is visually

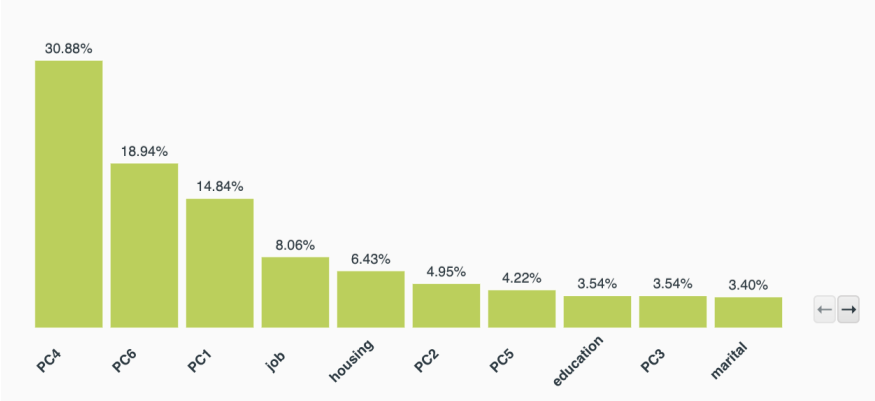

Figure 6: Screenshot of the summary report for the Portuguese bank dataset generated through BigML.

demonstrated in Figure 5, which is the orthogonal variable space of PC4 and PC1, for those two variables are the closest to the center, meaning their p-values are closest to 0. Thus, balance, age, and job are the most influential variables in the model respectively. I uploaded both the raw dataset and the manipulated dataset to BigML to compare results, and the evaluation showed that both accuracy percentages were relatively high, but the accuracy percentage for the manipulated dataset was slightly lower. This was expected because although the model was stronger using the PCs, some data was still omitted in the process. In conclusion, if this bank in Portugal wished to promote the issue of a checking account to their clients, I would recommend they ask their clients who work in management, are 33-38 years in age and their balance is equal to or greater than $3,000.

5 Conclusion and Future Results

Throughout my research, I learned how to simulate my own data and analyze it, and I also learned how to analyze data from a real-life application. During this process I learned how to create a regression model, consider which variables in action and learn in depth how to use three reliable statistical techniques to make the model stronger. I would recommend each

of these techniques to anyone trying to analyze a dataset, regardless of its sizes, although I would not recommend using PCA specifically if the dataset does not hold more than six quantitative variables. Each of these techniques contributed to helping me understand the significance of a variable, and how much that can affect the overall regression model.

During this process, there was some discrepancy between R and SAS output when it came to the estimators of the categorical variables. R output generated estimators that were a lot more similar to the original parameter estimate table, which was what we wanted to happen, while SAS output generated values that were some-what off. There was also some discrepancy between the PCA functions in both software, where SAS is known to have a stronger and more detailed method behind its function. A good idea for future results would be to investigate the exact reasoning behind the discrepancy in software. I believe it is critical to understand which software is stronger during which functions while implementing predictive analytics. I also noticed that every time I generated a new test/training in BigML for my manipulated dataset, the accuracy percentage would be either slightly higher or slightly lower. As explained previously, it was not surprising to receive a lower accuracy percentage, but it is a little more concerning to see a higher accuracy percentage, although the difference is very small. An excellent way to further investigate this in the future would be to regenerate a training/test one hundred or so times and find an accurate confidence interval to better interpret this accuracy percentage.

References

Agresti, A. (2007). *An introduction to categorical data analysis*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis, sixth edition*. Upper Saddle River, New Jersey: Pearson Education, Inc.

Siegel, E. (2016). *Predictive analytics*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Tan, P., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining*. Boston, MA: Pearson Education, Inc.

Appendix
Main R code

```
z = 2.657e+01 + -1.234e-7*x1 + 4.374e-08*x2 + -1.146e-06*x3 + 5.462e-07*x4 + 1.619e-07*x5 + -4.1
pr = 1/(1+exp(-z))          # probability obtained from logit function

total.cut <- cut(pr, c(0, 0.5, 1))
total.count <- table(total.cut)
y = rbinom(100,1,pr)       # binary response variable
df = data.frame(y=y,x1=x1, x2=x2, x3=x3, x4=x4, x5=x5, x=x6)  # Create a data frame structure
glm(y~x1+x2+x3+x4+x5+x6,data=df,family="binomial")


Bank <- read.csv("~/Desktop/Senior '18/Honors Thesis/Part II 2018/Bank.txt", sep=";")

mod <- glm(y~age+balance+campaign+contact+day+default+duration+education+housing+job+loan+marita
summary(mod)

Qvar = Bank[c("age","balance","day","duration","campaign","pdays","previous")]

#Mean <- as.numeric(apply(Qvar,2,mean))
#StandardDeviation <- as.numeric(apply(Qvar,2,sd))
#library(xtable)
#print(xtable(rbind(Mean,StandardDeviation),digits=2))

prcomp(Qvar)
library(FactoMineR)
PCAName=PCA(Qvar)
PCAName$var #for the betas

#using SAS eigenvectors
PC1 = .000508*Bank$age+.042574*Bank$balance+-.223982*Bank$day+.062709*Bank$duration+-.233727*Ban
PC2 = -.165040*Bank$age+-.080778*Bank$balance+.584159*Bank$day+-.361582*Bank$duration+.624431*Ba
PC3 = .673358*Bank$age+.694411*Bank$balance+.028128*Bank$day+-.230598*Bank$duration+.099304*Bank
PC4 = .0178661*Bank$age+.078178*Bank$balance+.409640*Bank$day+.876043*Bank$duration+.143047*Bank
PC5 = .693790*Bank$age+-.706073*Bank$balance+-.067857*Bank$day+-.065448*Bank$duration+.095881*Ba
PC6 = -.077848*Bank$age+-.066114*Bank$balance+-.658714*Bank$day+.200538*Bank$duration+.717913*Ba
#PC7 = .005079*Bank$age+.020836*Bank$balance+.038555*Bank$day+.014183*Bank$duration+.023258*Bank

SASData = data.frame(PC1=PC1, PC2=PC2, PC3=PC3, PC4=PC4, PC5=PC5, PC6=PC6, Bank[c("y","contact",

SASmod2 <- glm(y~PC1+PC2+PC3+PC4+PC5+PC6+education+housing+job+loan+marital,data=SASData,family=
summary(SASmod2)

write.csv(Bank, file="RAWPortugueseBankData.csv")
```

Main SAS code

```
%web_drop_table(WORK.Bank);
FILENAME REFFILE '/folders/myfolders/Bank.txt' TERMSTR=CR;

PROC IMPORT DATAFILE=REFFILE
        DBMS=DLM
        OUT=WORK.Bank;
        DELIMITER=";";
        GETNAMES=YES;
        DATAROW=2;
RUN;

PROC CONTENTS DATA=WORK.Bank; RUN;
%web_open_table(WORK.Bank);
proc princomp data = bank plots=all n=5;
run;
```

Reference for BigML

http://bigML.com