STAT 449/549 Homework 5

Casey Freimund

Due 10/25/2019

Overview

During this homework you will produce a brief report of the data analysis described below. The report should include:

1. Presentation of statistical summary of data
    - Methods: describe what decisions were made and why they were made. If there is any challenge in analysis, describe your approach to tackle the problem.
    - Results: No computer output, but a small number of tailored tables and graphics would be appropriate.
2. Predictions: Make and describe predictions for the test data.

There is no page limit for this report, but loquaciousness will not be considered fondly.

You will also need to submit a file that provides your predictions for the testing set. This can be a text file or a .csv file.

The Titanic

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships.

One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others.

In this homework, your mission to complete the analysis of what sorts of people were likely to survive. In particular, your goal is to predict which passengers (in the test data) survived the tragedy based on your model (using the training data) and present your findings.

The Data

You will be working with a reduced data set. The data has been split into two groups:

- training set (`training.csv`)
- test set (`testing.csv`)

The *training set* should be used to build your classification models. For the training set, you are provided the outcome (also known as the "ground truth") for each passenger. Your model will be based on "features" like passengers' gender and class. You can also use feature engineering to create new features.

The *test set* should be used to see how well your model performs on unseen data. For the test set, we do not provide the ground truth for each passenger. It is your job to predict these outcomes. For each passenger in the test set, use the model you trained to predict whether or not they survived the sinking of the Titanic.

###Data Dictionary

Variable	Definition	Key
PassengerId	Index of passengers	
Survived	Survival	0 = No, 1 = Yes
Pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
Name	Name	
Sex	Sex	
Age	Age in years	
SibSp	# of siblings / spouses aboard the Titanic	
Parch	# of parents / children aboard the Titanic	
Ticket	Ticket number	
Fare	Passenger fare	
Cabin	Cabin number	
Embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

###Variable Notes

- Pclass: A proxy for socio-economic status (SES)
  - 1st = Upper
  - 2nd = Middle
  - 3rd = Lower
- Age: Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5
- Sibsp: The dataset defines family relations in this way:
  - Sibling = brother, sister, stepbrother, stepsister
  - Spouse = husband, wife (mistresses and fiancés were ignored)
- Parch: The dataset defines family relations in this way…

- Parent = mother, father
- Child = daughter, son, stepdaughter, stepson
- Some children travelled only with a nanny, therefore parch=0 for them.

Important Considerations

- There are missing data in this data set. Make a decision on how to handle these observations, and justify your decision (how does your decision affect your analysis?).
- `pclass`, `sex`, and `embarked` are factors, and should be treated as such.
- Several of the variables may be highly correlated. How should these be handled?
- You are not limited to methods we have covered in class; however, you must be able to concisely explain the methods used.

Report

When the Titanic sank, it took with it 1502 (67.5%) of its passengers, due to the insufficient amount of lifeboats on board. But what about the 722 that survived? Did those 722 people have something in common that aided in their survival, or was it purely luck? I'm going to take a deeper look at the passengers and find out if I can determine who would survive based on their individual travelling circumstances.

To start my analysis, I had to split the data into two sets; a training set to explore the data and build a model and a testing set to use my model on. Following the data splitting, I looked for variables that were missing observations, 'NA'. Age was the only variable that contained 'NA' values. Due to only 154 of my 791 training observations missing age, I decided I could keep age as a variable in my analysis, if I replaced the 'NA' values with the median age. Next, I found that more than half of the observations were missing cabin number, so I decided to drop that variable. I also came across two passengers that didn't have a recorded port of embarkation, so I decided to get rid of those two observations. I came to this conclusion after finding out those two passengers were travelling together and that there would be no other way for me to determine where they could have came from.

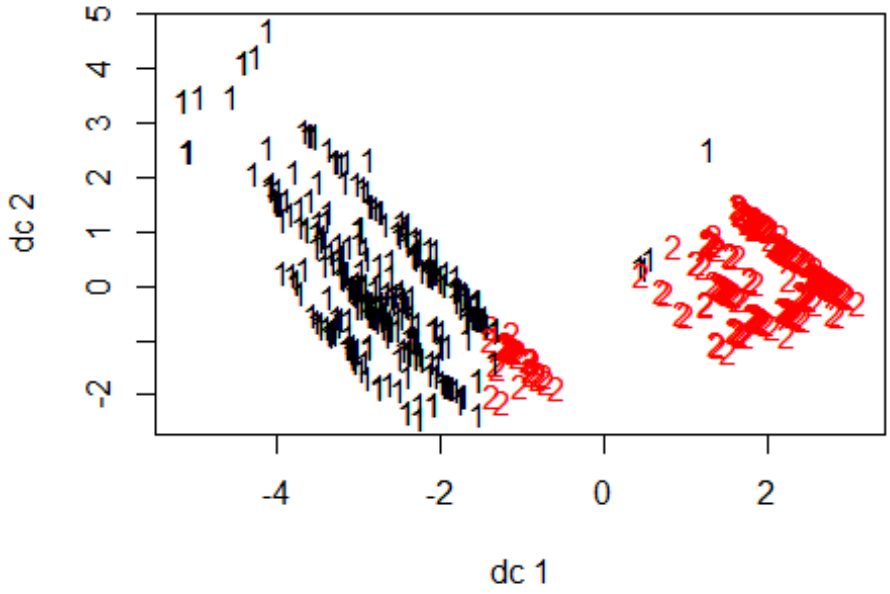
After handling variables with missing values, I moved on to data manipulation. I first changed the sex variable so I could use it in my model, by making males equal to 0 and females equal to 1. Next, I noticed that some ticket variables had just numbers and some had numbers and letters. After digging into the data I found that some passengers had the same ticket number, as shown below.

```
##     PassengerId Survived Pclass    Sex Age SibSp Parch Ticket  Fare Cabin
## 163         263        0      1   male  52     1     1 110413 79.65   E67
## 459         559        1      1 female  39     1     1 110413 79.65   E67
## 486         586        1      1 female  18     0     2 110413 79.65   E68
##     Embarked
## 163        S
## 459        S
## 486        S
```

Based on the variables in this example, I found it safe to assume that this is a family of 3 riding on the titanic. From this and other observations, I concluded that any group of people with the same ticket ID were travelling together. I also came to the conclusion that if a passengers Sibsp and/or their Parch measurement was greater than 1, they also weren't travelling alone. By taking all of these variables into consideration, I was able to account for people travelling with family, and childeren travelling with nannies. To help reduce variables, I added a variable to every observation called 'alone'. If a passenger shares a ticket ID with at least 1 more passenger, alone will equal 0, indicating that they are travelling with someone. If the passenger has a unique ticket ID, they are travelling alone and therefore alone is set equal to 1. Likewise, if SibSp and/or Parch is greater than or equal to 1, alone will be set to 0, and 1 otherwise.

My final step in data manipulation involved changing the port of embarkation to a numeric variable so it could be used in creating my predictive model. Doing so, I made passengers who travelled from Cherbourg equal to 1, Queenstown equal to 2, and Southhampton equal to 3.

The next step in my data analysis was to determine if there was a visual difference between those who survived the titanic, and those who did not. I used k-means clustering to try and divide the data into two groups, survived and didn't survive. Below you can see there are two distinct groups within our data, with some overlap, when the survived variable is removed.



From here, I started my variable reduction process to try and explain as much of the variability in the data as possible, while still allowing my model to be easily interpretable. I

started off with testing to see if exploratory factor analysis would be useful in reducing my number of variables. By using Bartlett's test for sphericity, I was able to conclude that our population correlations among our pairs of variables are not all zero, and that our data matrix is not the identity matrix. To take this test one step further, I used KMO to measure the sampling adequacy of my variables, results are shown below.

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = mat)
## Overall MSA =  0.67
## MSA for each item =
## Survived   Pclass      Sex      Age    SibSp    Parch     Fare Embarked
##     0.61     0.59     0.65     0.58     0.69     0.76     0.71     0.72
##    alone
##     0.73
```

Based on Hutcheson and Sofroniou, our overall MSA value is mediocre, and all of our variables are acceptable to keep. But, I still believed that we could reduce our number variables to aid in our model making process. By looking at the clusters produced by my k-means cluster analysis, I noticed a pattern amongst those who survived. A majority of survivors appeared to be female, travelling in first class, and/or travelling with someone. Next, I decided to group the variables together using a varimax factor rotation to determine if sex, class, and alone could stand alone as variables, or if there were other variables that could be grouped with them.

By splitting the variables into 4 groups, I was able to explain about 75% of the total variability in the data. All of our variables have a commonality factor greater than .6, and none of our variables appeared in multiple groups. From here I determined that embarked and sex could be categorized as their own variables, but I still had to address the other 6 variables that were divided into their respective group.

```
##        Pclass    Age   Fare
## Pclass  1.000 -0.326 -0.550
## Age    -0.326  1.000  0.094
## Fare   -0.550  0.094  1.000

##        SibSp  Parch  alone
## SibSp  1.000  0.419 -0.510
## Parch  0.419  1.000 -0.501
## alone -0.510 -0.501  1.000
```

Although age was grouped with class and fare in our factor analysis, it has a very weak relationship with class and fare, so I determined it could stand alone as its own predictive variable. Still looking at the first correlation matrix, we can see that fare and class have a stronger negative correlation, because when you pay more to travel, the higher class you'll be put in, i.e. 1st class. I found it best to keep class as a variable of interest and get rid of fare because what class a passenger ends up in, is dependent on the fare they paid.

In my second correlation matrix, you can see that sibling/spouse on board and parents/childeren on board have a negative correlation with alone. Although its a

somewhat strong correlation, I chose to keep alone in my set of variables because I created the alone variable based on SibSp, Parch, and Ticket ID, therefore alone is dependent on those 3 variables.

After reducing my number of variables from 12 down to 6, I first made a full model using all 6 variables. My full model gave me a cross-validation misclassification rate of 0.2231. Below, you can see a majority of passengers that were misclassified were passengers that didn't survive, but our equation predicted that they did. I took a closer look at my full equation using the summary function (shown below), and found that alone wasn't significant to the model when $\alpha = .05$. As you can also see, embarked isn't significant to the model when $\alpha = .01$.

```
##
## mypred.s   0   1
##        0 372  63
##        1 113 241

##             Estimate Pr(>|z|)
## (Intercept)   2.8699   0.0000
## Pclass       -1.1938   0.0000
## Sex           2.5269   0.0000
## Age          -0.0324   0.0001
## Embarked     -0.2793   0.0198
## alone         0.0429   0.8295
```

From here, I decided to make a reduced model with my original 6 variables, but get rid of embarked and alone. My reduced model gave me a cross-validation misclassification rate of 0.2243, which doesn't appear to be a significant difference. By looking at the chart below, again we can see the majority of people were misclassified as surviving, when they did not. To determine if there's a significant difference between both models, I conducted an ANOVA test with my reduced model, against my full model.

```
##
## mypred.s   0   1
##        0 370  62
##        1 115 242

##             Estimate Pr(>|z|)
## (Intercept)   2.2255        0
## Pclass       -1.2184        0
## Sex           2.5328        0
## Age          -0.0321        0
```

Based on my ANOVA test, I was able to determine that there was not a significant difference in results between my full and reduced model (p=0.065901). Therefore, my model for predicting if a passenger survived the titanic or not was;

$\hat{y} = 2.23 - 1.2Pclass + 2.53Sex - 0.03Age - 0.28Embarked + 0.09Alone.$

From this, I was able to conclude that the majority of people who survived were either in 1st or 2nd class, Female, and/or travelling with someone.