Jill Mercer (jmm1037)

Stats 660

Titanic Stat 660 Final

Summary

One of the most interesting events in history was the sinking of the Titanic in 1912. As can be noted in several documentaries on the subject, women and children were evacuated first and several other males and fathers were left onboard and went down with the ship. Over 1500 people died in this tragic accident while 705 individuals were saved. There are several interesting factors that can be investigated with a dataset such as this.

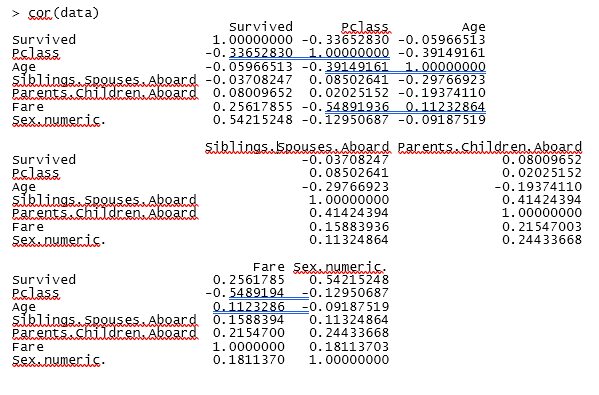
The Titanic data set is a supervised set and consists of eight variables: survived, pclass, name, sex, age, siblings, parents and fare. Each of the 887 columns in the .csv file corresponds to a passenger that was on the ship. The survived and sex columns are binomial in nature and the pclass variable contains values of 1-3 related to the class of the passenger. The siblings and parents’ variables indicate the number of each the passenger had on board the ship. The dataset was adapted to fit the model building format. The sex column was adjusted to indicate a 0 for male and 1 for female.

One question that I have wondered is whether myself and my fiancé would have survived. By running a logistic regression model, a multiple logistic regression and linear regression on the data with the variables of survived, age, sex and class I hope to answer the question: what the probability of death is given someone’s age and sex. The goal in this is to be able to predict a dummy variable given either age and sex to determine the likelihood of survival and ultimately see if we would have survived the accident.

After looking at the dataset it was determined that the most effective approaches would be using linear regression to determine the variables of importance and logistic regression for classification. Multiple logistic regression would not be useful as the variable of survival is a binomial and thus logistic regression would be able to handle this classification better.

The first step is to determine the correlation between different variables with the dataset and thus a pairwise correlation matrix was developed:

Figure : Pairwise Correlations



Based on the results in Figure 1, a significant correlation can be seen between Survived and Sex.numeric. This shows that the strongest indicator as to whether a person lives, or dies is due to the sex of the person. Three regression models are fit based on this information, Survived vs Sex and Age, Survived vs Age and Class and finally Survived vs Sex, Age and Class. The goal in this to determine which model is the best predictor of survival rate.

After running the regression models, the best performance was with variables Age, Sex and Pclass. All variables contribute to the model, with the worst p-value being Age at 1.46e-06 which is considerably lower than 1. The regression model on Age and Sex had a higher significant p-value on Sex at <2e-16 whereas age was 0.725, this indicates that Age does not contribute greatly to the model prediction. Lastly the regression model with Age and Class saw significance on both variables with a higher weight on Pclass.

The next step was to determine the success rate of the models. A probability estimate was performed on each regression and each observation was given a likelihood of survival. The probability of surviving was then considered to be true if its value was higher than 0.5 and alternatively false if lower. The accuracy of each model is given as follows:

Figure : Age+Sex classification

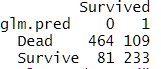


Figure 2 accuracy: 78.58%

Figure : Age+Class classification

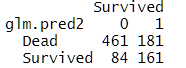


Figure 3 accuracy: 70.12%

Figure : Sex+Age+Class classification

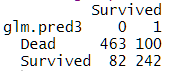


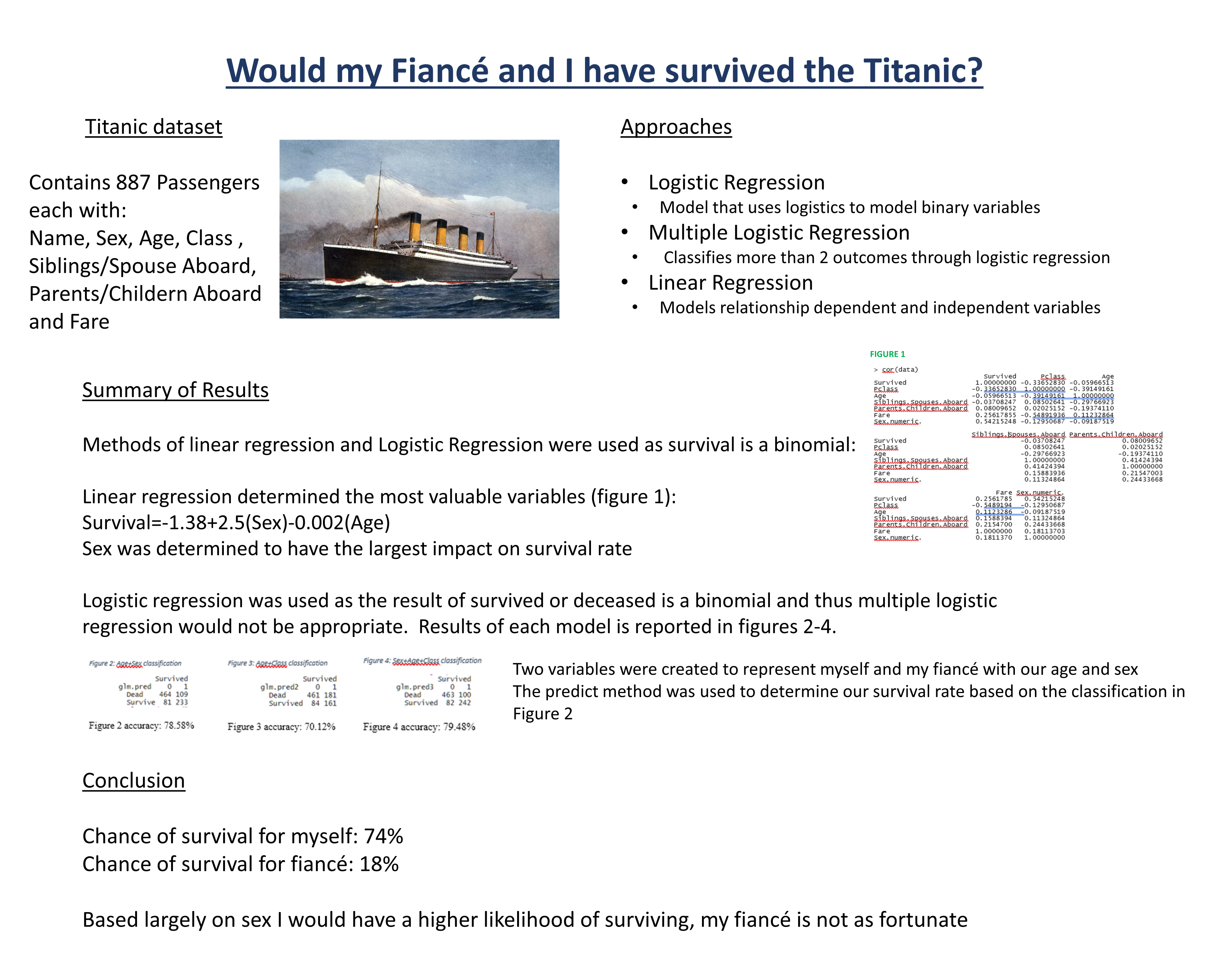
Figure 4 accuracy: 79.48%

The most accurate model was with variables sex, age and class although, it can be noted that the addition of class improves the model just 1% more than the age and sex prediction model.

After developing the models, it was evident that Sex of the person played a major role into who lived or survived the accident. Creating 2 variables that demonstrated myself and my fiancé with our sex and age, I was able to determine the chance of survival between us of being 74% and 19% respectively.

After these results were calculated, I would have a larger chance of survival at 74% due to being female than my fiancé would at 19% as being male. Age does play a slight role as the results tend to decrease the older the person, as can be seen through the regression model Survival=-1.38+2.5(Sex)-0.002(Age).

Poster slide:



Code:

titanic=read.csv("titanicformated.csv")

newdata=read.csv("morevars.csv")

data2=newdata[c(1:2,4:8)]

data=titanic[c(1:2,4:8)]

#pairwise correlation

cor(data)

attach(data)

#regression models

fit=glm(Survived~Sex.numeric.+ Age, data=data, family=binomial)

fit2=glm(Survived~Age+Pclass, data=data, family=binomial)

fit3=glm(Survived~Sex.numeric.+Age+Pclass, data=data, family=binomial)

summary(fit)

summary(fit2)

summary(fit3)

coef(fit)

#probability of survival based on sex age and class

prob=predict(fit,type="response")

prob2=predict(fit2,type="response")

prob3=predict(fit3,type="response")

#accuracy of model

glm.pred=rep("Dead",887)

glm.pred[prob>.5]="Survived"

table(glm.pred,Survived)

glm.pred2=rep("Dead",887)

glm.pred2[prob2>.5]="Survived"

table(glm.pred2,Survived)

glm.pred3=rep("Dead",887)

glm.pred3[prob3>.5]="Survived"

table(glm.pred3,Survived)

#percent chance myself and fiance will survive

mySelf=data.frame(Sex.numeric.=1, Age=25)

fiance=data.frame(Sex.numeric.=0, Age=23)

predict(fit,mySelf, type="response")

predict(fit,fiance, type="response")