B.TECH HONS. PROJECT

Titanic Machine Learning From Disaster

Name: Francis Xavier

Year: 2nd Year

Objective:

Predicting survival rate of people involved in a shipwreck by RMS Titanic dataset.

The objective of this project is to analyze the dataset of titanic survivors and use machine learning algorithms to predict the type of passengers that are likely to survive.

Data set Link: <https://www.kaggle.com/c/titanic>

About the dataset:

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.

We analyze as to what sort of people survived the shipwreck and use Machine Learning tools to find which passengers survived.

We are given 2 datasets-

1. A training set, complete with the outcome (or target variable) for a group of passengers as well as a collection of other parameters such as their age, gender, etc. This is the dataset on which we train your predictive model.
2. A test set, for which we predict the now unknown target variable based on the other passenger attributes that are provided for both datasets.

Libraries used:

**library**('ggplot2') *: Data visualization using*

*Graphics*

**library**('ggthemes') *: Visualization; Themes and*

*Scales for ggplot2*

**library**('scales') *: Scale function for visualization*

**library**('dplyr') *: For data manipulation*

**library**('caret') *: Classification and Regression*

*Training*

**library**('rpart') *: Recursive Partioning and*

*Regression Trees*

**library**('rpart.plot') *: Plotting ‘rpart’ Models*

**library**('RColorBrewer') *: ColorBrewer Palettes*

Loading and Checking Data:

> train <- read.csv("F:/Francis/R-Studio/Titanic

/train.csv", stringsAsFactors = F)

> test<- read.csv("F:/Francis/R-Studio/Titanic/

test.csv", stringsAsFactors = F)

> full <- bind\_rows(train, test)

> str(full)

'data.frame': 1309 obs. of 12 variables:

$ PassengerId: int 1 2 3 4 5 6 7 8 9 10 ...

$ Survived : int 0 1 1 1 0 0 0 0 1 1 ...

$ Pclass : int 3 1 3 1 3 3 1 3 3 2 ...

$ Name : chr "Braund, Mr. Owen Harris" "Cumings

, Mrs. John Bradley (Florence Briggs

Thayer)" "Heikkinen, Miss. Laina" "

Futrelle, Mrs. Jacques Heath

(Lily May Peel)" ...

$ Sex : chr "male" "female" "female" "female"

...

$ Age : num 22 38 26 35 35 NA 54 2 27 14 ...

$ SibSp : int 1 1 0 1 0 0 0 3 0 1 ...

$ Parch : int 0 0 0 0 0 0 0 1 2 0 ...

$ Ticket : chr "A/5 21171" "PC 17599" "STON/O2. 3

101282" "113803" ...

$ Fare : num 7.25 71.28 7.92 53.1 8.05 ...

$ Cabin : chr "" "C85" "" "C123" ...

$ Embarked : chr "S" "C" "S" "S" ...

# Note: Training set : train, Test set: test , Full Combined Set: full

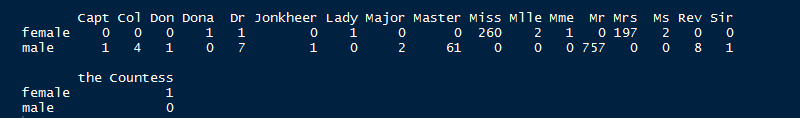
# Variables used and their descriptions:

# 

# The Name Variable:

# The Passenger Name can be broken down into additional meaningful variables.

> full$Title <- gsub('(.\*, )|(\\..\*)', '', full$Name)

> table(full$Sex, full$Title) ****

# Gives us a brief idea on how many Captains, Colonels, Doctors, Mr and Mrs etc. were present aboard the ship.

# Fates of Passengers:

> table(train$Survived)

0 1

549 342

# We see that 342 Passengers survived and 549 died.

# Proportion of Survivors:

> prop.table(table(train$Survived))

0 1

0.6161616 0.3838384

# The prop.table() function gives us the proportion of survivors with (0) being dead and (1) being survived. The output tells us that 38% of passengers survived the Disaster and 62% of passengers did not.

# Number of Survivors(Gender Wise):

> table(train$Sex,train$Survived)

0 1

female 81 233

male 468 109

# Shows us that in training set 81 females and 468 males died in the tragedy .

# Proportion of Survivors(Gender Wise):

> prop.table(table(train$Sex,train$Survived))

0 1

female 0.09090909 0.26150393

male 0.52525253 0.12233446

# The Outcome shows us that 233 females constitute 26% and 109 males constitute 12% of survivors in the tragedy.

# Number of Children who Survived:

> train$child<-0

> train$child[train$Age<18]<-1

> aggregate(Survived ~ child + Sex, data=train,FUN=sum)

child Sex Survived

1 0 female 195

2 1 female 38

3 0 male 86

4 1 male 23

# Tells us the number of child survivors where child=0 (for age above 18) and child=1(for below age 18, ie. Children).

# Using Fare rates to predict survival of passengers:

> train$Fare2 <- '30+'

> train$Fare2[train$Fare<30 &train$Fare>=20]<-'20-30'

> train$Fare2[train$Fare <20 &train$Fare>=10]<-'10-20'

> train$Fare2[train$Fare < 10] <- '<10'

> aggregate(Survived~Fare2 + Pclass + Sex, data=train

,FUN=function(x) {sum(x)/length(x)})

Fare2 Pclass Sex Survived

1 20-30 1 female 0.8333333

2 30+ 1 female 0.9772727

3 10-20 2 female 0.9142857

4 20-30 2 female 0.9000000

5 30+ 2 female 1.0000000

6 <10 3 female 0.5937500

7 10-20 3 female 0.5813953

8 20-30 3 female 0.3333333

9 30+ 3 female 0.1250000

10 <10 1 male 0.0000000

11 20-30 1 male 0.4000000

12 30+ 1 male 0.3837209

13 <10 2 male 0.0000000

14 10-20 2 male 0.1587302

15 20-30 2 male 0.1600000

16 30+ 2 male 0.2142857

17 <10 3 male 0.1115385

18 10-20 3 male 0.2368421

19 20-30 3 male 0.1250000

20 30+ 3 male 0.2400000

# While the majority of males, regardless of class or fare still don’t do so well, we notice that most of the class 3 women who paid more than $20 for their ticket actually also miss out on a lifeboat.

# Do Families Swim Together?

# We make a family size variable based on number of siblings/spouse(s) and number of children/parents. And using that variable to predict survival of passengers if they have a family.

> full$Fsize <- full$SibSp + full$Parch + 1

>ggplot(full[1:891,],aes(x=Fsize,fill=factor(Survived)))

+ geom\_bar(stat='count', position='dodge') +

+ scale\_x\_continuous(breaks=c(1:11)) +

+ labs(x = 'Family Size') +

+ theme\_few()

# Output:

# 

# Decision Trees

# A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm.

# Decision Trees are Machine Learning Techniques that we can use to see what decisions will be made for unseen data and predictions.

# Code:

>my\_tree<-rpart(Survived~Sex+Age,data=train,

method="class")

> new.fit <- prp(my\_tree,snip=FALSE)$obj

# Output:

# 

This decision tree representation does not provide any necessary information such as what the node stands for; survivors or non-survivors.

# Better Representation of Decision Trees:

# Code:

> fit<-rpart(Survived ~Sex + Age,data=train,

method="class")

> rpart.plot(fit,extra=104, box.palette="GnBu",

branch.lty=3, shadow.col="gray", nn=TRUE)

# Output:

# 

# The root node at the top, shows that 62% of passengers die, while 38% survive. The number above these proportions indicates the way that the node is voting (that everyone would die, coded as zero) and the number below indicates the proportion of the population that resides in this node(at the top level it is everyone, 100%).

# Then the decision Tree splits the node into Sex=male (yes) and if Sex not male then (no). If yes then it puts up another node there are only 19% male survivors and rest 81% dead and that is why the bucket is represented by a zero for higher dead. And for node 3 , there are 74% female survivors with the bucket being represented with 1.

# The final nodes at the bottom of the decision tree are known as terminal nodes, or leaf nodes. The boolean choices made for a given passenger, will end up in one of the leaf nodes, and the majority vote of all passengers in the bucket determines how we predict for new passengers with unknown fates.

# Drawbacks of Decision Trees:

# Decision trees do have some drawbacks though, they are greedy. They make the decision on the current node which appear to be the best at the time, but are unable to change their minds as they grow new nodes.

# There are a huge number of decisions that could be made, and exploring every possible version of a tree is extremely computationally expensive. This is why the greedy algorithm is used.

For survivors also separated by Pclass, Sex, Age, SibSp, Parch, Fare, Embarked:

Code:

> fit2<-rpart(Survived ~Pclass + Sex + Age + SibSp +

Parch + Fare + Embarked,data=train,

+ method="class")

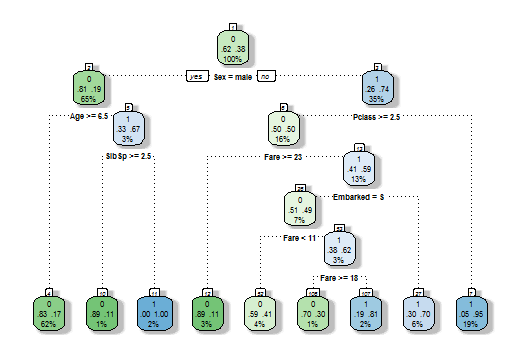
> rpart.plot(fit2,

+ extra=104, box.palette="GnBu",

+ branch.lty=3, shadow.col="gray",

tweak = 1.3, nn=TRUE)

Output:



Predictions for Null values:

1.For null values of age:

* Number of Null values in full.

Code:

> sum(is.na(full$Age))

[1] 263

* Making the Prediction

Code:

> Agefit <- rpart(Age ~ Pclass + Sex + SibSp + Parch +

Fare + Embarked + Title + Fsize,

+ data=full[!is.na(full$Age),],

+ method="anova")

> full$Age[is.na(full$Age)] <- predict(Agefit,

full[is.na(full$Age),])

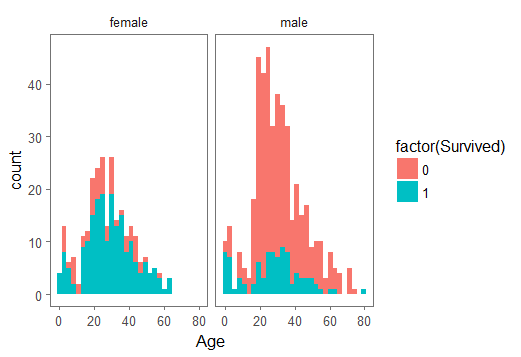
> ggplot(full[1:891,], aes(Age,fill=factor(Survived)))

+ geom\_histogram() +

+ facet\_grid(.~Sex) +

+ theme\_few()

Output (Before Prediction) :



2. After finding Null values for Age.

* Number of Null Values

Code:

> sum(is.na(full$Age))

[1] 0

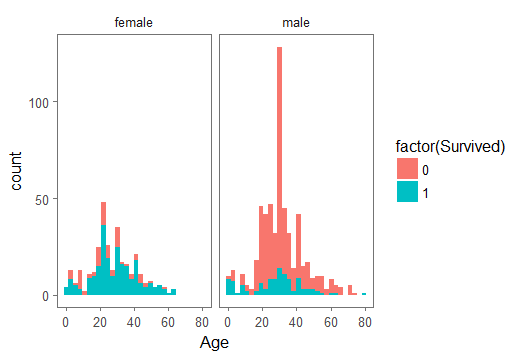
> ggplot(full[1:891,],aes(Age,fill =factor(Survived)))

+ geom\_histogram() +

+ facet\_grid(.~Sex) +

+ theme\_few()

Output (After Prediction):



We can see the difference between both ggplots for age before and after prediction of null values.  
And it suggests that Over 100 males of age group 20-40 were the maximum number of persons who died.

Conclusion:

It is clear from our analysis of data that there were a majority of men who were left behind that belonged to the lower classes and despite of the fewer deaths of women most of whom were carried to safety.

Many Factors such as having a family or being single mattered too because many may be reluctant to leave their families whereas a vast number of singletons were left behind, but comparatively the number of people having a family were bought to safety mainly when family size was more than 2.

Predictions of number of passengers and their proportions also gave us an idea about the percentage of survivors , survivors for male and female and children who were bought to safety.

Data Mining and machine learning techniques such as Decision trees also paved ways for Prediction of Data.