Warm-up Project: Titanic: Machine Learning From Disaster

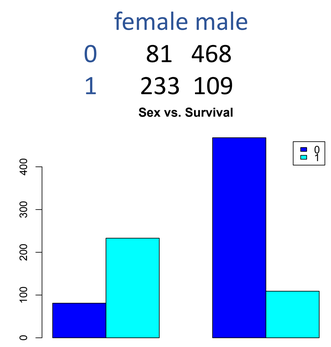
Kaitlin Sun
Jul 19th, 2019

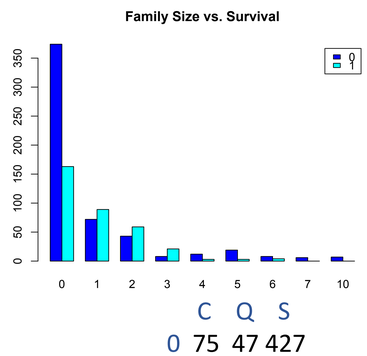
Introduction

On April 15, 1912, the Titanic sank after colliding with an iceberg, leading to the death of 1502 out of 2224 passengers and crew. For the project, we tried to find out what sorts of people were likely to survived. we analyzed the survival dataset of 891 people and built models to predict survival of other 481 people.

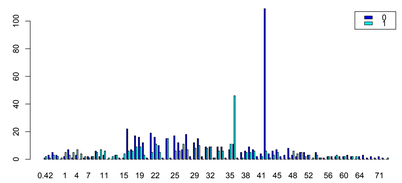
Relationship Between "Survived" and Variables

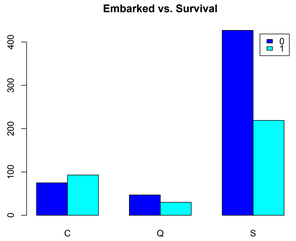
female male
0 81 468
1 233 109



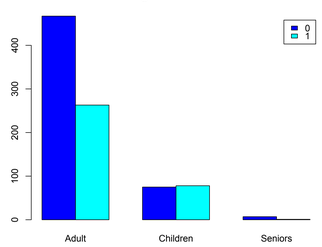


C Q S
0 75 47 427
1 93 30 219

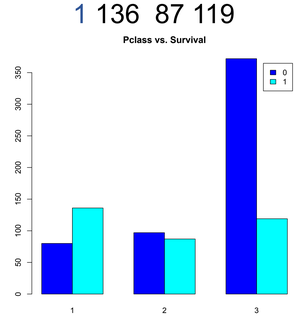




Adult Children Seniors
0 467 75 7
1 263 78 1



1 2 3
0 80 97 372
1 136 87 119



Data Analysis

- 74% female survived and 81% male died
- Survival rate is only 50% only for family size of 2-3
- Survival rate for "Embarked"="S" is 33%
- 1st class has survival rate over 50%

Data Processing

In both train and test dataset, ticket number, cabin and name are information that do not repeat for any passenger, which makes them less significant comparing to other data. Besides, title information from name could be strongly correlate to sex and pclass(a proxy of socio-economic status). As a result, I decided not to use these data in my prediction.

- Missing data: For training dataset, there are missing data on "Age" and "Embarked".
  - Embarked: There are over 50% of "S" for this dataset and only 2 missing data, so I imputed "S" for both missing data.
  - Age: Making a linear regression for age and found out the significant variables for age. Then, I used Predictive Mean Matching(PMM) to impute all the missing data.
- Class: changing the class of "Pclass" and "Survived" to factor
- New columns: for my prediction, I made two new columns: "Fnum" and "Agegr"
  - "Fnum": combining "SibSp"(number of siblings on the Titanic) and "Parch"(number of siblings on the Titanic) into "Fnum"(number of family number on the Titanic).
  - "Agegr": variable "Age" contains too many different numbers, so I made another column classifying "Age" into children(age<18), adults(18<=age<=64) and seniors(age>65).

First Prediction

For the first prediction, I used logistic regression without data processing.

- gf=glm(Survived~Pclass+Sex+Age+SibSp+Parch+Fare+Embarked,family=binomial)

By checking the model with summary(lr), the standard error of "Embarked" is huge(around 670) and p-value is also relatively large. I decided to not include "Embarked" in my model.

- gf=glm(Survived~Pclass+Sex+Age+SibSp+Parch+Fare,family=binomial)
- The result is 76%

Modification Predictions

With the first model, "SibSp" and "Parch" are not significant variables, and they have similar effect to people on a sinking ship. For the second prediction, I used "Fnum" instead.

- gf=glm(Survived~Pclass+Sex+Age+SibSp+Parch+Fare+Embarked,family=binomial)
- This modification give a result of 77%

Then, I discussed with other people who are doing this same project and decided to put "Embarked" in my model.

- gf=glm(Survived~Pclass+Sex+Age+Fnum+Fare+Embarked,family=binomial)
- The result is 78%

Further Predictions

The 3rd try is the highest score. Even though there was no improvement on accuracy, I tried more predictions with different changes.

For the first three predictions, I did not impute any missing data.

For my 4th prediction, as mentioned in "Data Processing", I made a linear regression for Age and imputed all the missing data for "Age"with PMM. Also, I impute "S" for missing data on "Embarked".

- gf=glm(Survived~Pclass+Sex+Age+Fnum+Fare+Embarked,family=binomial)
- The result is 77%

Then, I tried random forest to predict survival rate.

- Rf=randomForest(Survived~.,data=t,mtry=3,importance=TRUE)
- The result is 77%

Afterwards, I used "Agegr" instead of "Age" to reduce the error from imputing missing data; however, this also make the data vaguer.

- gf=glm(Survived~Pclass+Sex+Agegr+Fnum+Fare+Embarked,family=binomial)
- The result reduced to 75%

Conclusions

By all the models I made, the final result is 78%, with logistic regression using "Pclass", "Sex", "Age", "Fnum", "Fare" and "Embarked".

As my first machine learning project, the result accuracy is not high. However, I spent lots of time learning how to clean data, how to build models and fix errors. It helped me to get familiar with machine learning and R language.

References

1. Chow, Wing Ho. "Math 4432 Final Project Chow Wing Ho 20279607.Pdf." 2018.
2. James, Gareth, et al. *An Introduction to Statistical Learning with Application in R*. Springer Science+Business Media, 2017.
3. Jonge, Edwin de, and Mark van der Loo. "An Introduction to Data Cleaning with R." *Statistics Netherlands*, 2013, cran.r-project.org/doc/contrib/de_Jonge+van_der_Loo-Introduction_to_data_cleaning_with_R.pdf.
4. Kabacoff, Robert. "Bar Plots." *Quick-R: Bar Plots*, 2017, www.statmethods.net/graphs/bar.html.
5. Mi, Jeffrey, Mi, Jonathan (2019)
6. Shailesh. *R Snippets*. 2 Nov. 2017, buildmedia.readthedocs.org/media/pdf/r-snippets/latest/r-snippets.pdf.
7. /@harshithamekala08. "Dealing with Missing Data Using R - Coinmonks." *Medium*, Coinmonks, 30 June 2018, medium.com/coinmonks/dealing-with-missing-data-using-r-3ae428da2d17.