Modeling Titanic Survival

Qiushi Yan[a]

[a]Beijing, China

ARTICLE HISTORY
Compiled October 14, 2020

ABSTRACT
This short analysis showcases the development of a binary logistic model with spline transformations on predictors, to predict the possibility of survival in the loss of Titanic. I demonstrate the overall modeling process, including preprocessing, exploratory analysis, model fitting, adjustment, bootstrap validation and interpretation as well as other relevant techniques such as redundancy analysis and multiple imputation for missing data. The motivation and justification behind critical statistial decision are explained. This analysis is also made fully reproducible with R code and text provided.

KEYWORDS
logistic regression; multiple imputation; model validation

http://www.crema-research.ch/papers/2009-03.pdf
https://www.insider.com/titanic-secrets-facts-2018-4#at-the-memorial-of-frederick-fleet-one-of-the-lookouts-a-prankster-left-a-pair-of-bino
http://rpubs.com/edwardcooper/titanic1
https://www.kaggle.com/mrisdal/exploring-survival-on-the-titanic/report
https://www.kaggle.com/startupsci/titanic-data-science-solutions/comments
https://www.newscientist.com/article/dn22119-sinking-the-titanic-women-and-children-f

1. Introduction

The sinking of RMS Titanic brought to various machine learning competitions a quintessential dataset among others, in which one major interest is to predict possibility of survival given sex, age, class, etc. There are several variants of this data existed on the web, the one I will be using is accessed on Encyclopedia Titanica, namely `titanic3`, courtesy of Philip Hind, with the following variables (table 1)

This data frame recorded the survival status 1309 Titanic passengers[1] alongside his/her gender, age, family relations on board, ticket fare, etc. There were 809 victims and 500 survivors in total.

This case study has been greatly inspired by Dr. Frank Harrell's similar one in his *Regression Modeling Strategies* (2015, Chapter 12) book, here I attempt to propose my own idea and interpretation of model development that is as original as possible. To

CONTACT Qiushi Yan. Email: qiushi.yann@gmail.com, website: https://qiushi.rbind.io

[1]The data does not involve crew members, and the total number of passengers is said to be 1317

Table 1. Data with 1309 passengers and 14 columns

.	Variable	Definition	Note
1	pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
2	survival	Survival Status	0 = No, 1 = Yes
3	name	Name	
4	sex	Sex	
5	age	Age	In years, some infants had fractional values
6	sibsp	Number of Siblings/Spouses Aboard	
7	parch	Number of Parents/Children Aboard	
8	ticket	Ticket Number	
9	fare	Passenger Fare	in Pre-1970 British Pounds
10	cabin	Cabin	
11	embarked	Port of Embarkation	Cherbourg, Queenstown or Southampton
12	boat	Lifeboat	
13	body	Body Identification Number	
14	home.dest	Home/Destination	

ensure reproducibility, all the analysis is done in R (R Core Team 2020) and RStudio with code and text made public in this repo.

- Quantify predictive ability of each predictors, i.e. which predictor is most dominant in determining whether a passenger will survive
- Find Interactions between predictors
- Whether the *Women and children first* policy is respected. After the collision, The captain explicitly issued an order for women and children to be saved first.

Here is a brief summary of the following sections

- Exploration, data preprocessing based on descriptive statistics and visualization, finish with a redundancy analysis

2. Exploration

2.1. *Exploratory analysis on survival patterns*

Before any analysis, we'll start by some data munging. First exclude those variables that bring little insight to prediction: `name`, `ticket`, `embarked`, `body`, `cabin`[2] and `home.dest`. The `boat` column is left out for another reason, because an non-missing entry in `boat` basically means survival and missing means death[3].

Next, for purposes of interpretation we will transform `fare` into today's US dollars with correction for inflation. According to discussion here, we make the transformation

[2]Because this was primarily an identification for class and most were missing

[3]More precisely, there were 9 recorded passengers who got on the lifeboat yet died before reaching Carpathia, another RMS which spearheaded the rescue of Titanic survivors. There were also 13 passengers who survived with no boat information documented, and this is most likely due to data quality issues after looking up on Encyclopedia Titanica. Even with these exceptions, whether a passenger got on a lifeboat yields perfect prediction on his/her survival. If one fits a logistic regression model on survival based on whether `boat` is missing, the apparent accuracy will be nearly 1

$$\frac{\text{today's US dollar}}{\text{fare in 1912}} \approx \underbrace{5}_{\text{exchange rate then}} \times \underbrace{26}_{\text{inflation index from 1912 to 2020}}$$

There were 17 passengers whose `fare` is zero, all of whom males boarding in Southampton, the start of the voyage. It is suspected that some of them may be falsely included crew members, or this could be an error in data collection. I will treat these anomalies as missing entries for simplicity. Moreover, one passenger in the 3rd class had missing value in `fare`. Given the small amount of missingness, single imputation with median `fare` conditional on class is used.

Finally, a nice summary of all existing variables in the data is given by the `Hmisc::describle` function.

```
# print a summary for the data
titanic %>%
  mutate(survived = factor(survived)) %>%
  describe() %>%
  latex(file = "", size = "small", center = "none")
```

7 Variables 1309 Observations

---

pclass

n	missing	distinct
1309	0	3

Value	1st	2nd	3rd
Frequency	323	277	709
Proportion	0.247	0.212	0.542

---

survived

n	missing	distinct
1309	0	2

Value	0	1
Frequency	809	500
Proportion	0.618	0.382

---

sex

n	missing	distinct
1309	0	2

Value	female	male
Frequency	466	843
Proportion	0.356	0.644

---

age

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
1046	263	98	0.999	29.88	16.06	5	14	21	28	39	50	57

lowest : 0.1667 0.3333 0.4167 0.6667 0.7500, highest: 70.5000 71.0000 74.0000 76.0000 80.0000

---

sibsp

n	missing	distinct	Info	Mean	Gmd
1309	0	7	0.67	0.4989	0.777

lowest : 0 1 2 3 4, highest: 2 3 4 5 8

Value	0	1	2	3	4	5	8
Frequency	891	319	42	20	22	6	9
Proportion	0.681	0.244	0.032	0.015	0.017	0.005	0.007

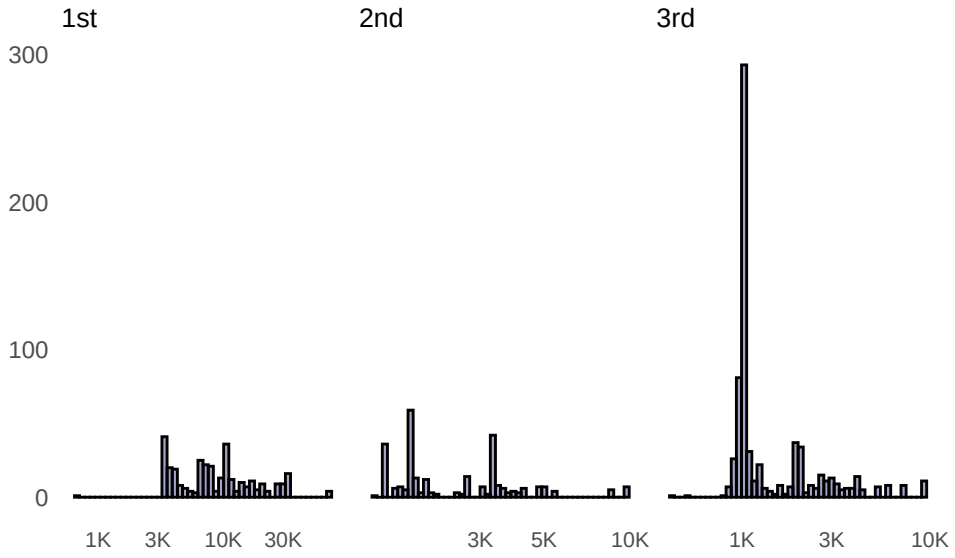


Figure 1. More than 75% of the third class passengers (700-plus in total) purchased tickets with price lower than $2000, while the median fare for second and first class is $3861. X axis is on log 10 scale

parch

n	missing	distinct	Info	Mean	Gmd
1309	0	8	0.549	0.385	0.6375

```
lowest : 0 1 2 3 4, highest: 3 4 5 6 9
```

Value	0	1	2	3	4	5	6	9
Frequency	1002	170	113	8	6	6	2	2
Proportion	0.765	0.130	0.086	0.006	0.005	0.005	0.002	0.002

fare

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
1309	0	283	1	4381	5019	939.8	1005.3	1030.3	1885.0	4080.4	10142.6	17374.5

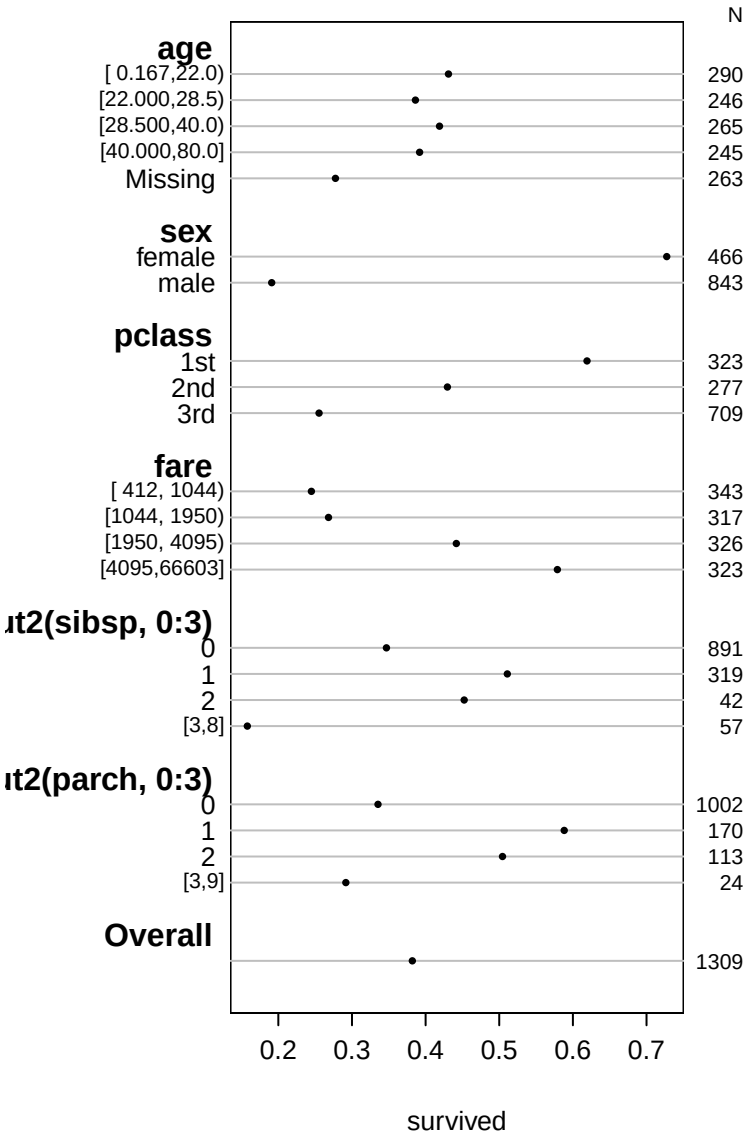
```
lowest :   412.204   521.625   650.000   810.875   836.875
highest: 29578.249 32177.704 34108.750 34190.000 66602.799
```

There are several interesting patterns to notice [4]

- `age` has roughly 20% missingness. On the other hand, the variable has a nice distribution with 80% known observations falling between 14
- Both `sibsp` and `parch` rarely have instances larger than 3, and can be readily categorized without losing too much information.
- The distribution of `fare` is heavily right skewed, caused by the large amount of third class passengers who make do with cheap cabins, as shown in figure 1. This may suggest a log transformation in the model.

- Approximately 20% `age` is missing. This calls for a necessary imputation method as we will see later, that survival can exhibit

[4]Though this may not be relevant to the model, it is still an surprising discovery that it wasn't until the late 19th century that the idea of women traveling alone gained ground. As a result, there were nearly twice as many males passengers as females on Titanic. In fact, only 40% female passengers have no family accompanies on the ship



Finally, redundant analysis

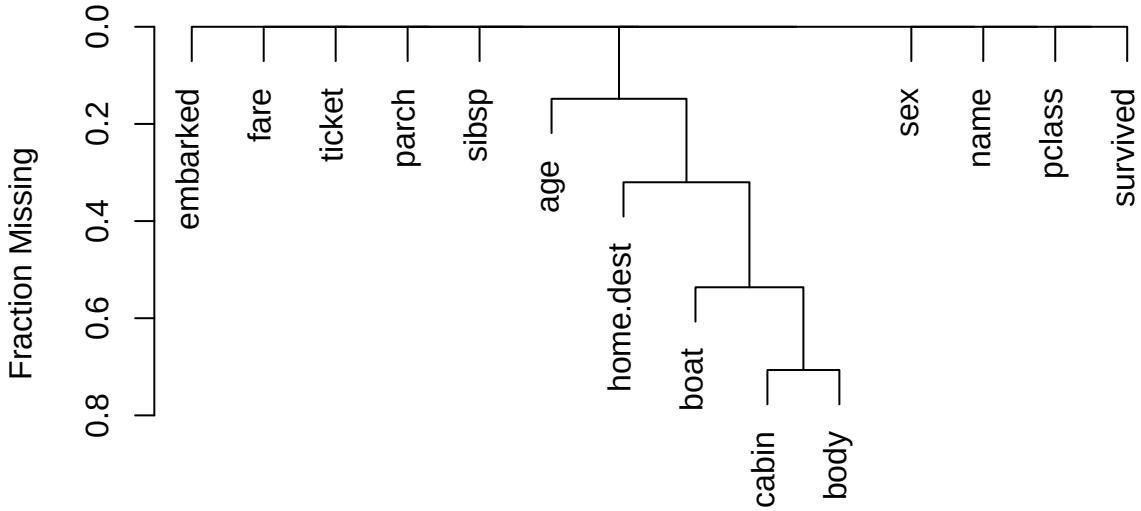
```
Redundancy Analysis

redun(formula = ~pclass + sex + age + cut2(sibsp, 0:3) + cut2(parch,
    0:3) + fare, data = titanic, minfreq = 40)

n: 1046     p: 6    nk: 3
```

```
Number of NAs:   263
Frequencies of Missing Values Due to Each Variable
          pclass              sex              age cut2(sibsp, 0:3)
               0                0              263                0
cut2(parch, 0:3)             fare
               0                0


Transformation of target variables forced to be linear

Minimum category frequency required for retention of a binary or
categorical variable: 40

R-squared cutoff: 0.9   Type: ordinary

R^2 with which each variable can be predicted from all other variables:

          pclass              sex              age cut2(sibsp, 0:3)
           0.797            0.102            0.304            0.414
cut2(parch, 0:3)             fare
           0.439            0.451

No redundant variables
```

2.2. *Data missing patterns*

Here we use the data before exluding irrelevant columns

```
plot(naclus(titanic_raw))
```



There are some simple workarounds

- complete-case analysis: That is, we delete all incomplete observations. Needless to say this will translate into a major harm on sample size since over 60% of `boat` are missing, not to mention other columns. Even if we remove `boat` and then delete rows with missing `age` we still lose over 1/5 of data. Moreover, figures in 2.1 have shed light on the relatively strong influence of `age` on survival. Also, the deletion of incomplete observations assumes date are missing completely at random (MCAR). When it's not the case, this could severely bias estimates of coefficients (Van Buuren 2018)
- single imputation:
- multiple imputation

For demonstration purposes I will fit two decision trees to predict

3. Modeling

3.1. *Initial model*

First and foremost,

3.2. *Multiple imputation*

3.3. *Validation*

There will not be another Titanic, so what is the point in validation since the model will not be used in prediction in all likelihood?

In the award-winning solution to this legendary dataset presented by IBM Watson, they used a holdout sample to validate their model. https://www.fharrell.com/post/split-val/

3.4. *The final model*

4. Discussion

A more detailed explanation of some of these measures is presented in the appendix.

5. Conclusion

Appendix A. Measures used in valiation

This will be Appendix A.

Appendix B. Techinical information

```
R version 4.0.2 (2020-06-22)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 18362)

Matrix products: default

locale:
[1] LC_COLLATE=English_United States.1252
[2] LC_CTYPE=English_United States.1252
[3] LC_MONETARY=English_United States.1252
[4] LC_NUMERIC=C
[5] LC_TIME=English_United States.1252
system code page: 936

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] mice_3.11.0     rms_6.0-1       SparseM_1.78    Hmisc_4.4-1
[5] Formula_1.2-3   survival_3.1-12 lattice_0.20-41 ggplot2_3.3.2
[9] dplyr_1.0.2

loaded via a namespace (and not attached):
 [1] Rcpp_1.0.5          mvtnorm_1.1-1       tidyr_1.1.2
 [4] png_0.1-7           zoo_1.8-8           assertthat_0.2.1
```

```
 [7] digest_0.6.25       R6_2.4.1            backports_1.1.10
[10] MatrixModels_0.4-1  evaluate_0.14       pillar_1.4.6
[13] rlang_0.4.8         multcomp_1.4-14     rstudioapi_0.11
[16] data.table_1.13.0   rticles_0.16.1      rpart_4.1-15
[19] Matrix_1.2-18       checkmate_2.0.0     rmarkdown_2.4
[22] labeling_0.3        splines_4.0.2       readr_1.4.0
[25] stringr_1.4.0       foreign_0.8-80      htmlwidgets_1.5.2
[28] munsell_0.5.0       broom_0.7.1         compiler_4.0.2
[31] xfun_0.18           pkgconfig_2.0.3     base64enc_0.1-3
[34] htmltools_0.5.0     nnet_7.3-14         tidyselect_1.1.0
[37] tibble_3.0.4        gridExtra_2.3       htmlTable_2.1.0
[40] bookdown_0.21       codetools_0.2-16    matrixStats_0.57.0
[43] fansi_0.4.1         crayon_1.3.4        conquer_1.0.2
[46] withr_2.3.0         MASS_7.3-51.6       grid_4.0.2
[49] nlme_3.1-148        polspline_1.1.19    gtable_0.3.0
[52] lifecycle_0.2.0     magrittr_1.5        scales_1.1.1
[55] cli_2.1.0           stringi_1.5.3       farver_2.0.3
[58] latticeExtra_0.6-29 ellipsis_0.3.1      generics_0.0.2
[61] vctrs_0.3.4         sandwich_3.0-0      TH.data_1.0-10
[64] RColorBrewer_1.1-2  tools_4.0.2         glue_1.4.2
[67] purrr_0.3.4         hms_0.5.3           jpeg_0.1-8.1
[70] yaml_2.2.1          colorspace_1.4-1    cluster_2.1.0
[73] knitr_1.30          quantreg_5.73
```

References

Harrell Jr, Frank E. 2015. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Van Buuren, Stef. 2018. *Flexible imputation of missing data*. CRC press.