Modeling Titanic Survival

Qiushi Yan[a]

[a]Beijing, China

ARTICLE HISTORY
Compiled October 22, 2020

ABSTRACT
This case study showcases the development of a binary logistic model 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 statistical decisions 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

Who Survived Titanic? A Logistic Regression Analysis: https://sci-hub.do/https://journals.sagepub.com/doi/pdf/10.1177/084387140401600205

- Do human beings behave more in line with the selfish *homo oeconomicus*, where everybody is out for himself or herself and possibly even puts other people's lives in danger

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. After the British passenger liner struck an iceberg in her maiden voyage on 15 April 1912 and was eventually wrecked, more than 1500 people perished. Decades of effort has been devoted to the study of the historic event, in which one major interest is to predict possibility of survival given a number

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

Table 1. Cleaned data with 2208 rows and 11 columns

.	Variable	Definition	Note
1	survived	Survival Status	0 = Lost, 1 = Saved
2	age	Age	In years, some infants had fractional values
3	gender	Gender	
4	class_dept	Class or Department	Passengers, Crew or Staff
5	joined	Port of Embarkation	Cherbourg, Queenstown or Southampton
6	nationality	Motherland	from wiki passenger list
7	title	Title	Extracted from name
8	spouse	Number of spouse on board	
9	sibling	Number of siblings on board	
10	parent	Number of parents on board	
11	children	Number of children on board	

of characteristics, since there was clear account that some people (woman, children) were allowed to get on the lifeboat first.

There are several variants of Titanic data existed on the web, with primary source based on Encyclopedia Titanica (1999) founded by Philip Hind. This project is based on the most recent version with following columns available (table 1). After appropriate formatting and cleaning, the data at hand recorded the survival status 2208 Titanic travelers alongside his/her gender, age, family relations on board, nationality, etc. There were 1496 victims and 712 survivors in total. Preprocessing steps are detailed in the data section in the appendix.

- Did socioeconomic advantage mean better chance of survival? If this is the case, people with higher financial means, i.e. who live in the first class are more likely to survive. Similarly, passengers from second class will have a higher change of survival than third class people.
- people in their prime
- Quantify predictive ability of each predictors, i.e. which predictor is most dominant in determining whether a passenger will survive
- Find Interactions between predictors. Specifically, there are important interactions that need extra notice. For example, it has been widely studied in sociology and anthropology that human are sometimes driven by *procreation instinct* so that social norms would entail needs to protect females of reproductive age (Frey, Savage, and Torgler 2009).[1] Therefore, we could specify and study the interaction between age and gender. Another typical interaction is between offspring and gender. *Parental investment* suggest that women on average invest more in caring for their offspring than males. In times of a disaster, higher opportunity cost will alert females with offspring more than others, and make them seek more aggressively for changes to secure the children as well as themselves.
- 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.[2]
- For those who traveled alone with no family relations on the vessel, is their survival possibility greater or less? On one hand, they are more likely to be in shortage of psychological and physical support. On the other hand, they would may be able to reach a life-saving decision faster without transaction cost and

[1] The average peak reproductive period in females is between the ages of 16 and 35.

[2] Though there is no international maritime law enforcing this kind of chivalry.

negotiation.

- Did English subjects receive any special care or given priority to aboard lifeboats? After all, Titanic was perated by British crew, and managed by British captain, masters and officers. Conversely, British elite were once incarnation for chivalry.

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 ensure reproducibility, all the analysis is done in R (R Core Team 2020) and RStudio with code and text made public in this repo. A brief summary of each section is listed below

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

2. Exploration

2.1. *Data processing and descriptive statistics*

A nice summary of of the data is given by the `Hmisc::describe` function. For numerical variables, a inline histogram is produced alongside summary measures such as the number of missing values and the mean. For discrete variables, we focus on the number of categories and their relative frequency.

11 Variables 2208 Observations

---

survived

n	missing	distinct
2208	0	2

```
Value          0     1
Frequency   1496   712
Proportion 0.678 0.322
```

---

age

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
1497	711	71	0.999	30.18	14.31	8	17	22	29	38	47	54

```
lowest :  0.8  1.0  2.0  3.0  4.0, highest: 67.0 69.0 70.0 71.0 74.0
```

---

gender

n	missing	distinct
2208	0	2

```
Value      Female  Male
Frequency     489  1719
Proportion  0.221 0.779
```

---

class_dept

n	missing	distinct
2208	0	5

```
lowest : 1st    2nd    3rd    crew   staff, highest: 1st    2nd    3rd    crew   staff

Value        1st   2nd   3rd  crew staff
Frequency    321   270   709   822    86
Proportion 0.145 0.122 0.321 0.372 0.039
```

---

joined

n	missing	distinct
2208	0	4

Value	Belfast	Cherbourg	Queenstown	Southampton
Frequency	200	271	123	1614
Proportion	0.091	0.123	0.056	0.731

---

nationality

n	missing	distinct
2208	0	7

`lowest : American English  Finnish  Irish    Other   , highest: Finnish  Irish    Other    Swedish  Syrian`

Value	American	English	Finnish	Irish	Other	Swedish	Syrian
Frequency	246	1002	58	168	549	99	86
Proportion	0.111	0.454	0.026	0.076	0.249	0.045	0.039

---

title

n	missing	distinct
2208	0	4

Value	Miss	Mr	Mrs	other
Frequency	267	1590	212	139
Proportion	0.121	0.720	0.096	0.063

---

spouse

n	missing	distinct	Info	Sum	Mean	Gmd
2198	10	2	0.087	66	0.03003	0.05828

---

sibling

n	missing	distinct	Info	Mean	Gmd
2198	10	4	0.139	0.05778	0.1108

Value	0	1	2	3
Frequency	2091	91	12	4
Proportion	0.951	0.041	0.005	0.002

---

parent

n	missing	distinct	Info	Mean	Gmd
2198	10	3	0.08	0.03822	0.07474

Value	0	1	2
Frequency	2138	36	24
Proportion	0.973	0.016	0.011

---

children

n	missing	distinct	Info	Mean	Gmd
2198	10	5	0.077	0.03913	0.0767

`lowest : 0 1 2 3 4, highest: 0 1 2 3 4`

Value	0	1	2	3	4
Frequency	2140	37	16	3	2
Proportion	0.974	0.017	0.007	0.001	0.001

There are several interesting patterns to notice.[3] First, `age` has roughly 30% missingness. On the other hand, the variable has a nice distribution with 80% known observations falling between 14 and 50. Distributions of subject's family relation on Titanic are all too narrow, as shown in figure 1. This motivates categorization since we will not lose too much information. Lastly, nearly half of the subjects are English. And if we focus on crew, the number rise to 85%.

Given this results, the final step in data munging is to dichotomize `spouse`, `parent`, `children` and `sibling` to denote if there is such relation. Thus we no longer have to deal with continuous predictors with poor distribution.

Univariate relationship between each independent variable and survival status is

[3] 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.

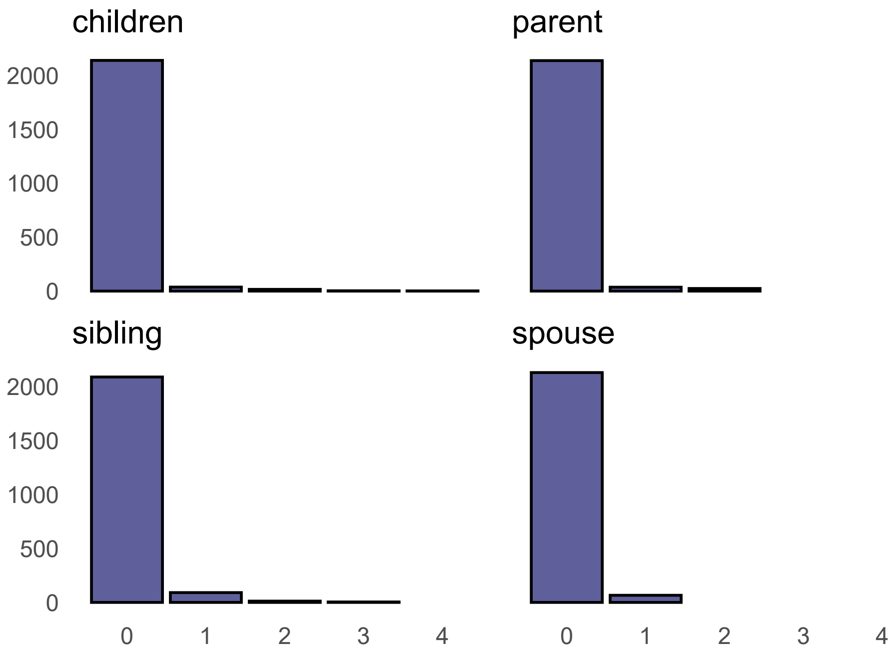


Figure 1. Few subjects have more than one companion in any of the 4 relations.

presented in figure 2. For each column, this is a anova-type plot with no control over confounding variables, though it may still assist us in determining how to spend degrees of freedom. If a predictor's effect on the response is strong, it's more likely that we need to spend more parameters on it. However, if a variable's effect appears to be weak, it could either result from a flat relationship with the response, or from nonlinearity and interaction among variables this plot fails to detect.

Finally, redundant analysis

2.2. *Data missing patterns*

```
naplot(naclus(t), "na per var")
```

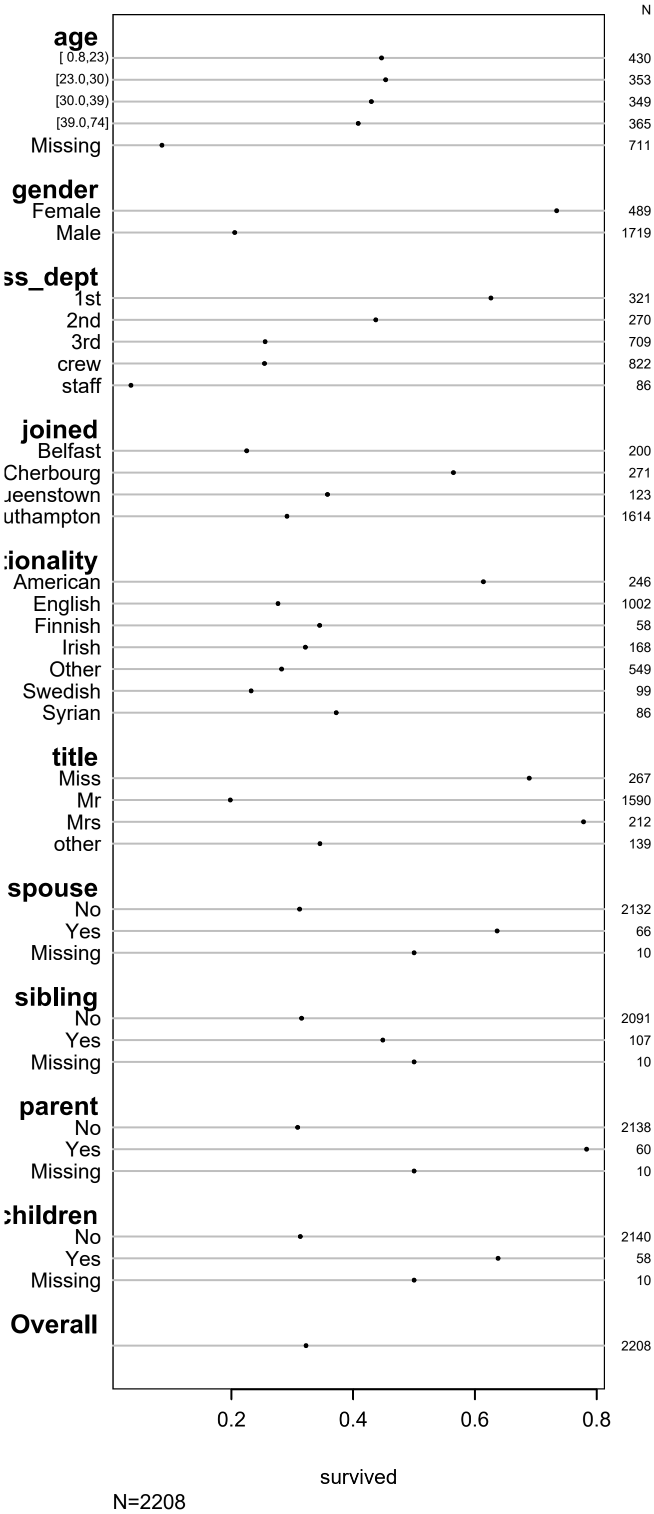
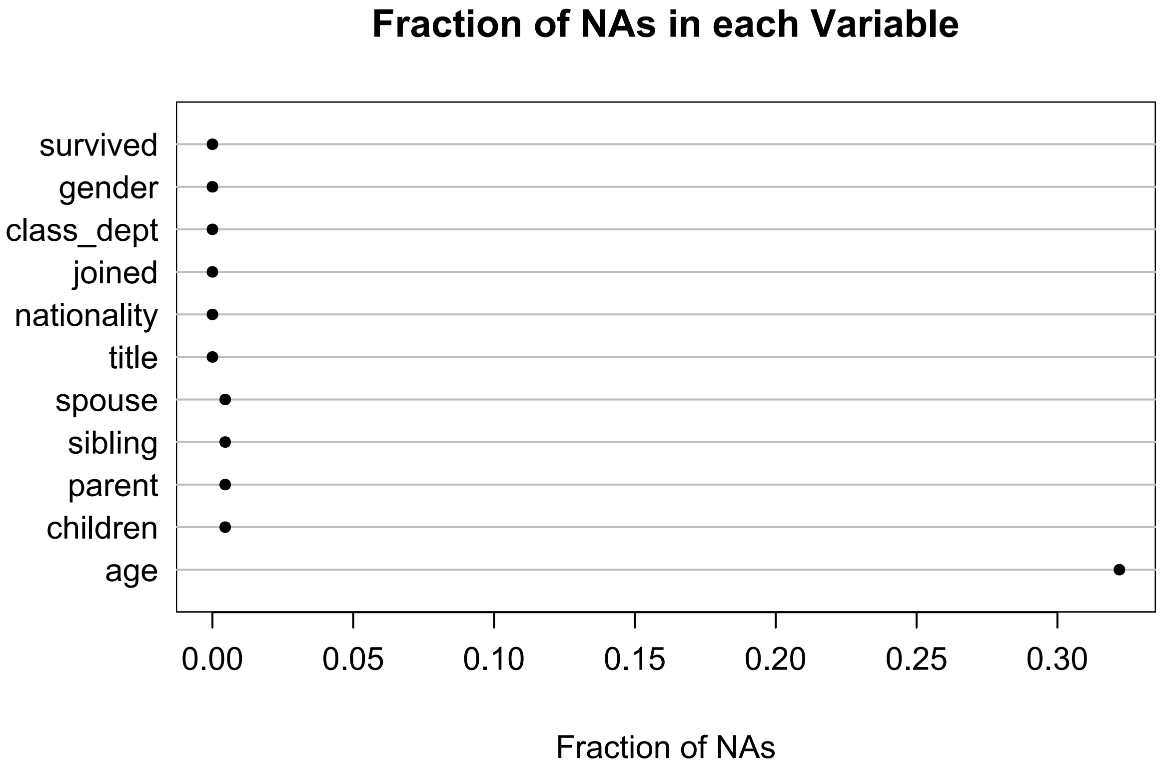
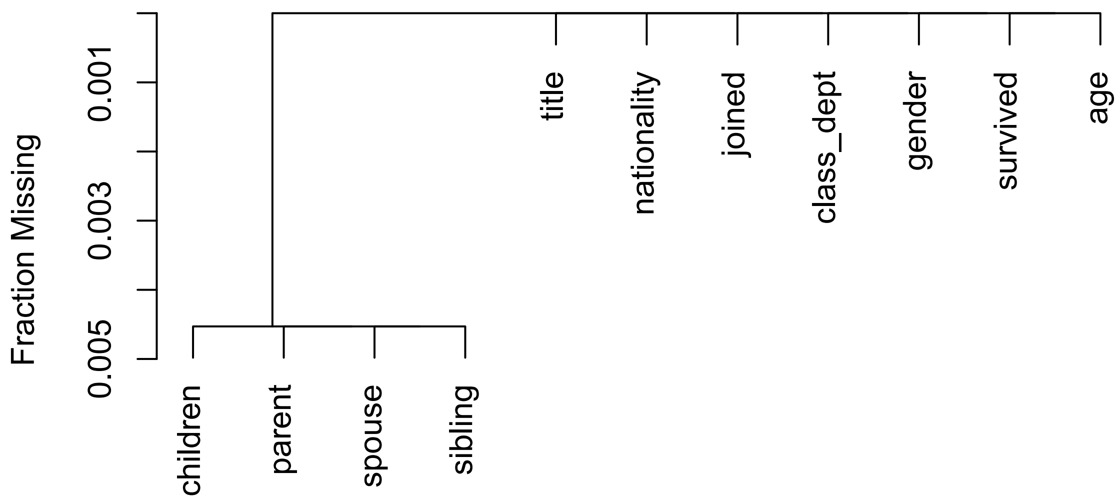


Figure 2. Summary of relationship between survival and each predictor




```
plot(naclus(t))
```



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 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

2.3. *Loess regression for nonlinear pattern*

The loess is a common nonparametric regression method to study nonlinear relationship. In the case of binary response, the fitted value at $x = x_0$ is the proportion of positive cases near the neighborhood of x_0 [4]. If the trend of a loess curve shows non-monotoncity, it is reasonbale to include that nonlienarity relationship in the model, e.g., modeling the predicotr with polynomial transformation or with splines.

figure 2

[4]with varying weights according to their distance to x_0

Table 2. Wald Statistics for `survived`

	χ^2	d.f.	P
age (Factor+Higher Order Factors)	75.38	24	<0.0001
All Interactions	50.41	20	0.0002
Nonlinear (Factor+Higher Order Factors)	38.24	18	0.0036
gender (Factor+Higher Order Factors)	221.89	5	<0.0001
All Interactions	32.22	4	<0.0001
class_dept (Factor+Higher Order Factors)	116.17	20	<0.0001
All Interactions	24.77	16	0.0740
age × gender (Factor+Higher Order Factors)	32.22	4	<0.0001
Nonlinear	8.29	3	0.0404
Nonlinear Interaction : f(A,B) vs. AB	8.29	3	0.0404
age × class_dept (Factor+Higher Order Factors)	24.77	16	0.0740
Nonlinear	22.80	12	0.0295
Nonlinear Interaction : f(A,B) vs. AB	22.80	12	0.0295
TOTAL NONLINEAR	38.24	18	0.0036
TOTAL INTERACTION	50.41	20	0.0002
TOTAL NONLINEAR + INTERACTION	58.16	23	<0.0001
TOTAL	275.41	29	<0.0001

3. Modeling

the choice of model. In this setting, it is obvious that we would prefer probabilistic predictions to classification with output label 0 and 1, since we are placing emphasis upon the *tendency* of survival. And the true value of our model consist not in the decision on who will survive, but in what characteristics would increase or decrease the possibility of survival. The notion has ruled out most of the black box machine learning models for classification, say, random forest, support vector machines and neural network. Not only are they not intrinsically probability oriented, it is hard to interpret main effects and interactions as everything seems to be interacted with one another.

3.1. *Saturated model*

First and foremost,

The limiting sample size for binary outcome would be the number of minority class, in our case 712. Using the 15:1 rule, that will give us some confidence spending roughly 47 parameters or degrees of freedom.

This plot is useful in identifying weather the relationship between survival status and any predictor is flat. [5]

likely shrinkage

[5] A misuse of this plot would be checking nonlinearity. Even with spline transformation and large corrected χ^2 there is no guarantee for nonlinearity.

3.2. *Multiple imputation and the final model*

3.3. *Diagonostics and interpretation*

3.4. *Validation*

There will not be another Titanic, and any model on Titanic will not be used for prediction. Therefore, the goal of model validation is primarily to provide quantify the degree of overfitting with various bias-corrected measures.

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/`

4. Discussion

The most decisive explanation for such effect is that first-class passengers had better access to information about the imminent danger and were aware that the lifeboats were located close to the first class cabins. Thus, their marginal effort costs to survive were lower. In contrast, most third-class passengers had no idea where the lifeboats were located (safety drills for all passengers were introduced after the Titanic disaster), and they did not know how to reach the upper decks where the lifeboats were stowed.

Wyn Craig Wade: there was a class culture on Titanic akin to the notion of a "culture of poverty

> Undoubtedly, the worst barriers were the ones within the steerage passengers themselves. Years of conditioning as third-class citizens led a great many of them to give up hope as soon as the crisis became evident ... Barriers to steerage? Yes, but of a kind less indictable to the White Star Line than to the whole of civilization.

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

5. Conclusion

Appendix A. Data

The source data is accessed on Encyclopedia Titanica, a leading archive on titanic facts. In contrast to the the famous titanic dataset (known as `titanic3`) distributed by kaggle for introductory level machine learning practices, the case study uses a more up-to-date and complete dataset in the following ways

- **larger sample size**: Our data includes crew and staff members alongside passengers, while titanic3 only incorporate passenger information. We do not use a separate test set approach for validation either. As a result, the sample size is about 2.5 times larger.
- **more columns**: Additional variables such as role on the ship, nationality and occupation are added. A major difference is made by separating the travel companion data into four distinctive columns: number of parents, children, sibling and spouses that each passenger traveled with. These were combined into two columns before.
- **more accurate**: `titanic3` was an effort to study Titanic in the 20th century, lastly updated and improved by Thomas Cason in 1999. The data has been

constantly revised, many errors corrected, many missing ages filled in, and new variables created. Now it reflects the state of the data as of 21 October 2020.

The data cleaning process involves using appropriate data types, creating new features, adjusting levels for categorical variable and excluding irrelevant columns. Code can be found at clean.R.

`title` is extracted through each person's name with regular expressions and then collapsed into 4 levels.[6]

Passengers are classified according to their cabin class. Others on the vessel fall into one of crew and staff members. Crew includes victualling crew[7], engineering crew, deck crew and officers, substitute crew and guarantee group. Staff members include restaurant staff and orchestra.

Rare nationality (lower than 50 people) is collapsed.

There is an indicator column telling when a person's age is only approximate and cannot be determined from current facts. These inaccurate age have been assigned NA.

Variables we do not utilize in this project includes name, date of birth and death, lifeboat number[8], fare, and cabin number.[9]

Appendix B. Measures used in valiation

Appendix C. Original Computing Environment

```r
sessionInfo()
```

```
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:
```

[6] For example, the title for passenger "Abbing, Mr Anthony" is "Mr".

[7] crew in charge of food, housekeeping, laundry, room service, etc.

[8] 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. In this sense `boat` is more the result of survival, rather than a cause.

[9] While some study used this attribute to find cabin locations, its large amount of missingness could be a source of major complexity.

```
[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_1.2-4   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_2.5
[22] labeling_0.4.2      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_0.7.2         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_5.74
```

References

Allaire, JJ, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. 2020a. *rmarkdown: Dynamic Documents for R*. R package version 2.5, https://github.com/rstudio/rmarkdown.

Allaire, JJ, Yihui Xie, R Foundation, Hadley Wickham, Journal of Statistical Software, Ramnath Vaidyanathan, Association for Computing Machinery, et al. 2020b. *rticles: Article Formats for R Markdown*. R package version 0.16.1, https://github.com/rstudio/rticles.

Frey, Bruno S, David A Savage, and Benno Torgler. 2009. "Surviving the Titanic disaster: economic, natural and social determinants." .

Harrell, Frank E, Jr. 2020. *Hmisc: Harrell Miscellaneous*. R package version 4.4-1, https://CRAN.R-project.org/package=Hmisc.

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

Harrell, Jr., Frank E. 2020. *rms: Regression Modeling Strategies*. R package version 6.0-1, https://CRAN.R-project.org/package=rms.

Hind, Philip. 1999. https://www.encyclopedia-titanica.org/.

Koenker, Roger, and Pin Ng. 2019. *SparseM: Sparse Linear Algebra*. R package version 1.78, http://www.econ.uiuc.edu/~roger/research/sparse/sparse.html.

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

Roecker, Ellen B. 1991. "Prediction error and its estimation for subset-selected models." *Technometrics* 33 (4): 459–468.

Sarkar, Deepayan. 2008. *Lattice: Multivariate Data Visualization with R*. New York: Springer. ISBN 978-0-387-75968-5, http://lmdvr.r-forge.r-project.org.

Sarkar, Deepayan. 2020. *lattice: Trellis Graphics for R*. R package version 0.20-41, http://lattice.r-forge.r-project.org/.

Terry M. Therneau, and Patricia M. Grambsch. 2000. *Modeling Survival Data: Extending the Cox Model*. New York: Springer.

Therneau, Terry M. 2020. *survival: Survival Analysis*. R package version 3.1-12, https://github.com/therneau/survival.

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

van Buuren, Stef, and Karin Groothuis-Oudshoorn. 2011. "mice: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software* 45 (3): 1–67. https://www.jstatsoft.org/v45/i03/.

van Buuren, Stef, and Karin Groothuis-Oudshoorn. 2020. *mice: Multivariate Imputation by Chained Equations*. R package version 3.11.0, https://CRAN.R-project.org/package=mice.

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, and Dewey Dunnington. 2020a. *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. R package version 3.3.2, https://CRAN.R-project.org/package=ggplot2.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2020b. *dplyr: A Grammar of Data Manipulation*. R package version 1.0.2, https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2014. "knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman and Hall/CRC. ISBN 978-1466561595, http://www.crcpress.com/product/isbn/9781466561595.

Xie, Yihui. 2015. *Dynamic Documents with R and knitr*. 2nd ed. Boca Raton, Florida: Chapman and Hall/CRC. ISBN 978-1498716963, https://yihui.org/knitr/.

Xie, Yihui. 2016. *bookdown: Authoring Books and Technical Documents with R Markdown*. Boca Raton, Florida: Chapman and Hall/CRC. ISBN 978-1138700109, https://github.com/rstudio/bookdown.

Xie, Yihui. 2020a. *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.21, https://github.com/rstudio/bookdown.

Xie, Yihui. 2020b. *knitr: A General-Purpose Package for Dynamic Report Generation in R*. R package version 1.30, https://yihui.org/knitr/.

Xie, Yihui, J.J. Allaire, and Garrett Grolemund. 2018. *R Markdown: The Definitive Guide*. Boca Raton, Florida: Chapman and Hall/CRC. ISBN 9781138359338, https://bookdown.org/yihui/rmarkdown.

Xie, Yihui, Christophe Dervieux, and Emily Riederer. 2020. *R Markdown Cookbook*. Boca Raton, Florida: Chapman and Hall/CRC. ISBN 9780367563837, https://bookdown.org/yihui/rmarkdown-cookbook.

Zeileis, Achim, and Yves Croissant. 2010. "Extended Model Formulas in R: Multiple Parts and Multiple Responses." *Journal of Statistical Software* 34 (1): 1–13.

Zeileis, Achim, and Yves Croissant. 2020. *Formula: Extended Model Formulas*. R package version 1.2-4, https://CRAN.R-project.org/package=Formula.