Modeling Titanic Survival

Qiushi Yan

[a]Beijing, China

ARTICLE HISTORY
Compiled November 10, 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 fully reproducible with all source R code and text.

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

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-binocu

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-fir

1. Introduction

The sinking of RMS Titanic brought to numerous machine learning competitions a quintessential dataset among others. After the "unsinkable" 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 for statistical inquiries is to model and predict survival given a number of characteristics, since there was clear account that some people were allowed to get on the lifeboat first.

There are numerous variants of Titanic data existed on the web, with primary source based on Encyclopedia Titanica (1999), a site started in 1996 as an attempt to tell the story of every person that traveled the Titanic as a passenger or crew member. This project is based on the most recent version as of October 2020, with following columns available (table 1). Source data and steps of data cleaning are elaborated in the data section in the appendix.

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
survived	Survival Status	0 = Lost, 1 = Saved
age	Age	In years, some infants had fractional values
gender	Gender	
class	Cabin class	1st, 2nd, 3rd or Crew
nationality	Motherland	from wiki passenger list
title	Title	Extracted from name
spouse	# of spouse on board	
sibling	Number of siblings on board	
parent	Number of parents on board	
children	Number of children on board	

After appropriate formatting and cleaning, the data at hand recorded the survival status 2208 Titanic travelers alongside his/her gender, age, companions on board, title, nationality, etc. There were 1496 victims and 712 survivors in total.

It is essential for every fruitful task of data analysis to first identify key questions of investigation that facilitates interpretation, however vague they are at the beginning. Then we can approach the core problem, filtering out trivialities, with statistical expression by abstraction. For our purposes, we could establish the following questions for which to quest

- To which degree is *Women and children first* policy respected? After the collision, the captain explicitly issued an order for women and children to be saved first.[1] Thus we should expect significantly higher proportion of females and children rescued than that in males and adults. If the opposite is true, that Titanic subjects behave more in line with the selfish *homo oeconomicus*, where everybody looked out for himself or herself and possibly even puts other people's lives in danger, then people in their prime with physical superiority would see higher probability of survival. This requires us to study gender and age effect.
- Did socio-economic advantages mean better chance of survival? If this is the case, passengers 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. Cabin class's impact on survival status needs special notice here.
- For those who traveled alone with no companions (spouse, sibling, parent, children) 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 negotiation.
- Did English subjects receive any special care or given priority to aboard lifeboats? After all, Titanic was operated by British crew, and managed by British captain, masters and officers. Conversely, British nobility and elite
- Quantify interactions among various characteristics. 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 reproduc-

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

tive age (Frey, Savage, and Torgler 2009).[2] 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.

This case study has been greatly inspired by Dr. Frank Harrell's similar example in his *Regression Modeling Strategies* (2015, Chapter 12) book, here I attempt to propose my understanding 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) with code and text made public in this repo. A brief summary of each section is listed below

- Exploration. Use descriptive statistics to examine data distribution characteristics, data missing patterns and relative effects, followed by redundancy analysis to study dependencies among predictors. Finish with nonparametric loess regression exploring nonlinear trends.
- Model development. The key section in specifying, developing, validating and describing a binary logistic model, split into
    - Specification Prespecification of predictor complexity with a saturated model. Guide later development of the final model with importance ranking based on bootstrap resampling.
    - Multiple imputation: Use predictive mean matching to impute subject's age, resulting in 30 complete dataset.
    - Model fitting, validation and calibration. Obtain pooled parameter estimates based on prespecified complexity and imputation results. Use bootstrap validation and calibration curve (the ".632" method) to study model performance and optimism.
    - Interpratation. Summarize the model with estimation and hypothesis testing, combined with graphical methods like partial effect plots and nomogram.
- Discussion. Model-based explanation to address former questions.
- Conclusion. Conclusion and further study.

2. Exploration

2.1. *Descriptive statistics and data processing*

A graphical summary of of the data is given by the `Hmisc::describle` 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.

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

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

11 Variables 2208 Observations

---

survived

n	missing	distinct	Info	Sum	Mean	Gmd
2208	0	2	0.655	712	0.3225	0.4372

---

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

---

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

---

class

n	missing	distinct
2208	0	4

Value	1st	2nd	3rd	crew
Frequency	321	270	709	908
Proportion	0.145	0.122	0.321	0.411

---

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
2208	0	2	0.087	66	0.02989	0.05802

---

sibling

n	missing	distinct	Info	Mean	Gmd
2208	0	4	0.138	0.05752	0.1103

Value	0	1	2	3
Frequency	2101	91	12	4
Proportion	0.952	0.041	0.005	0.002

---

parent

n	missing	distinct	Info	Mean	Gmd
2208	0	3	0.079	0.03804	0.07441

Value	0	1	2
Frequency	2148	36	24
Proportion	0.973	0.016	0.011

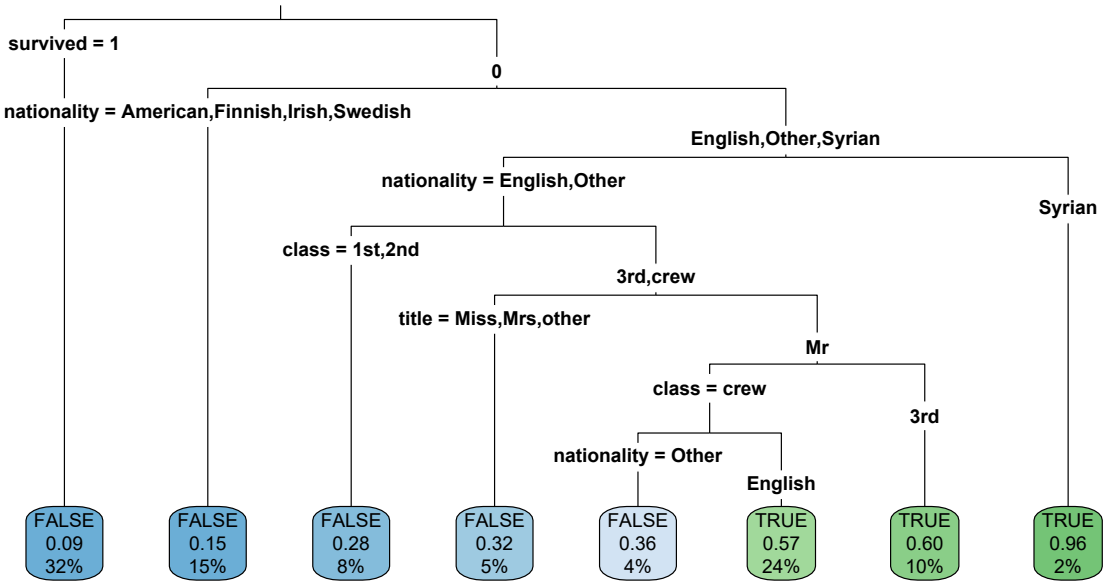

Figure 1. The decision tree for predicting `is.na(age)`, which finds strong patterns of missing related to class/department and gender (the Syrian node has very limited samples). Each node shows (top to bottom) the predicted class, the predicted probability of age being missing, the percentage of observations in the node.

children

n	missing	distinct	Info	Mean	Gmd
2208	0	5	0.077	0.03895	0.07636

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

Value          0     1     2     3     4
Frequency   2150    37    16     3     2
Proportion 0.974 0.017 0.007 0.001 0.001
```

There are several noteworthy patterns.[3]

Of special importance is the `age` variable, which has roughly 30% missingness. On the other hand, it has a nearly symmetric distribution with 80% known observations falling between 14 and 50. For further examination of patterns of missing data, we could fit a decision tree to predict which type of subject tend to have missing ages. Generally, for some third class male passenger or crew, age is mostly to miss.

```
na_tree <- rpart(factor(is.na(age)) ~ .,
                 data = t %>% mutate(survived = as.factor(survived)) ,
                 minbucket = 50)
# figure 1
rpart.plot::rpart.plot(na_tree, type = 3, cex = 0.6)
```

We see in figure 1 that survival status, gender and class are essential in determining age missingness. For a 3rd class male passenger who did not survive, age is missing with a probability of 60%. Interestingly, English male crew members are much more

[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 companion on the ship.

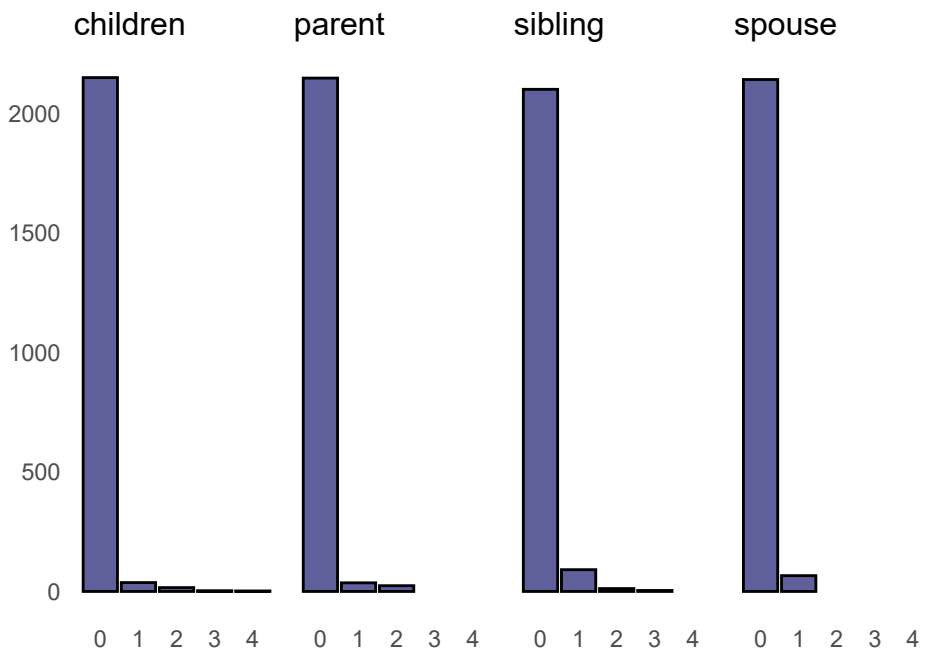

Figure 2. Few subjects have more than one companion in any of the 4 relations.

likely to have missing age than subjects of other nationality

Back to other variables in descriptive statistics. Distributions of subject's companion on Titanic are all too narrow, as illustrated in figure 2. There is too little variation to model them continuously. 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 presented in figure 3. For each column, we can build 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 due to a truly flat relationship, or to nonlinearity and predictors among variables that univariate method cannot detect.

The plot reveals appreciably strong effects of gender and cabin class on survival status. The effect of age seems trivial except for the missing subjects, but again, this figure exposes only linear relationship, and only after categorization. As we will see in the next section, age effect are much nonlinear and concentrated in the young subjects. The downside of this kind of univariate relationship is also exemplified in `title`, where "Miss". For the same reason effects of other variables cannot be determined.

We will finish with a redundancy analysis to study if any predictor can be readily explained by the rest of predictors, therefore does not much bring new information and may not enter the model. The checking algorithm involves

```
Redundancy Analysis
```

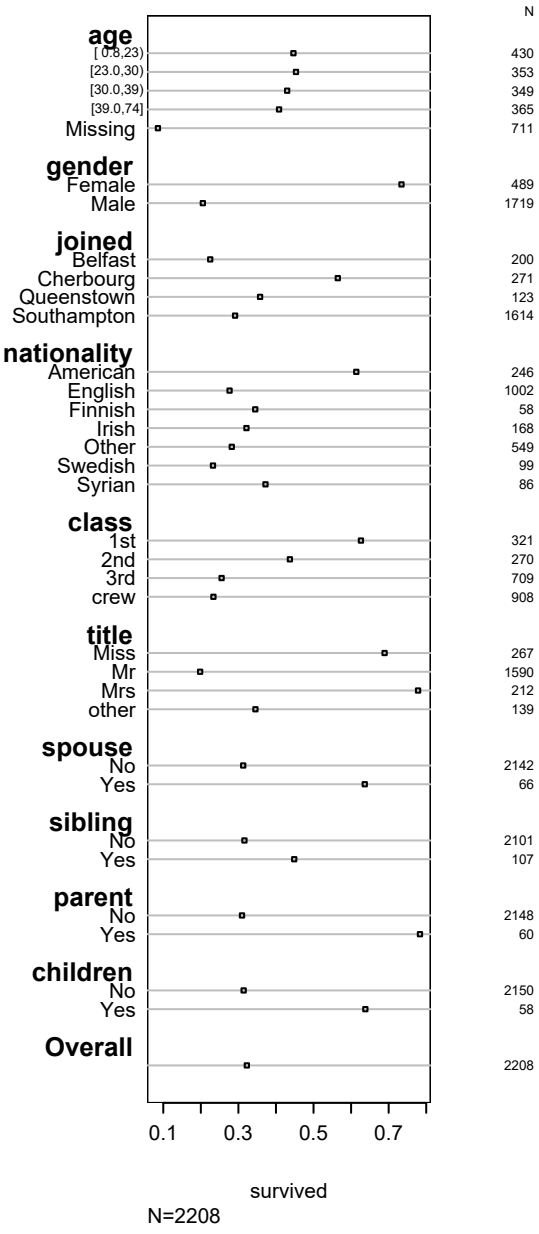

Figure 3. Univariate summary of relationship between survival and each predictor

```
redun(formula = ~age + gender + class + nationality + title +
    spouse + sibling + parent + children, data = t)

n: 1497     p: 9    nk: 3

Number of NAs:   711
Frequencies of Missing Values Due to Each Variable
        age      gender       class nationality       title      spouse
        711           0           0           0           0           0
    sibling      parent    children
          0           0           0


Transformation of target variables forced to be linear

R-squared cutoff: 0.9   Type: ordinary

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

        age      gender       class nationality       title      spouse
      0.340       0.977       0.551       0.492       0.977       0.341
    sibling      parent    children
      0.167       0.289       0.339

Rendundant variables:

title

Predicted from variables:

age gender class nationality spouse sibling parent children

  Variable Deleted   R^2 R^2 after later deletions
1            title 0.977
```

The redundancy analysis has reported , with more than 97% of uts variation explained by the rest of the predictors.

2.2. *Loess regression for nonlinear pattern*

The loess method is a common nonparametric regression model to study nonlinear relationship. In the case of binary response, the fitted value at $x = x_0$ is the weighted proportion of positive cases near the neighborhood of x_0. If the trend of a loess curve exhibits nonmonotoncity, it is reasonable to include that nonlinearity relationship in the model, e.g., modeling the predictor with polynomial transformation or with splines.

It was widely documented that gender interacts directly with age effects. Another important interaction, according to many follow up studies, happened with cabin class. Figure 4 displays loess estimates of survival probability given age under stratification. Not only in a powerful nonlinear fashion does age affect survival status (top left panel),

we also observe it interact with other two factors in a nonlinear way.

3. Model development

A typical modeling workflow begins with an choice of a statistical model or a machine learning model. A statistical model often stems from a hypothesized probabilistic data generating mechanism and assumes additivity, whereas machine learning models is algorithmatic in nature, optimized with parameter tuning. We choose to develop a statistical model, a "simple" binary logistic regression, for the following reasons.

We prefer probabilistic predictions to classification with output label 0 and 1, since we are placing emphasis upon the *tendency* of survival. And the value of the model consist not in a dichotomous prediction, but in what characteristics would increase or decrease the possibility of survival. The notion has ruled out most of the machine learning models for classification, say, random forest, support vector machines and neural network, which are not intrinsically probability oriented. Such classifiers can often only yield a forced choice.

Interpretability and inference matters. Many top data science competitions has reported moderately high signal to noise ratio (e.g., 90% prediction accuracy) that might tip the balance towards machine learning models, while interpretability is harmed. Specifically, statistical models favours additivity and explicit specification. It follows that there are natural distinctions between main effects and interactions, linearity and nonlinearity. And the inference procedure is well defined provided that the model is correctly specified. While in a multi-layer neural network, everything can interact with one another and it could be daunting to isolate effects and conduct former inference.

Machine learning models are data hungry and sometimes create the need for big data (van der Ploeg, Austin, and Steyerberg 2014). To guard against overfitting, the analyst has to have a sample size that is 10 times larger at least if he chooses a tree model instead of regression. While this case study uses a Titanic dataset that is about 1/3 larger than those only concerned with passengers, it is far less sufficient for a typical data-hungry machine learning model to validate well. The rationale is that a statistical model is a safer approach as Dr. Harrell commented

> If n is too small to do something simple, it is too small to do something complex

3.1. *Specification*

We start by fitting a relatively large model, to decide how model complexity should be properly represented. This includes deciding the number of knots for continuous predictors and the number of categories of categorical predictors, could we remove some term, where should we place interaction, etc. The large model also gives an overall sense of the predictive ability of each subject characteristics on survival status. This strategy as a starting point is also called prespecification of predictor complexity. It avoids creating phantom degrees of freedom when one has subjective judgment according to scatter diagrams or descriptive statistics on how to represent variables in a model. Commonly done, for example, is excluding a quadratic term simply because it is "non-significant", with p-value on the edge of 0.05. This approach is known to distort coefficient estimates, confidence intervals, p-value and calibration (too optimistic) of the final model (Grambsch and O'Brien 1991).[4] Because it fails to accounts for sampling

[4] confidence interval too narrow, p-value and standard errors too small and calibration too optimistic

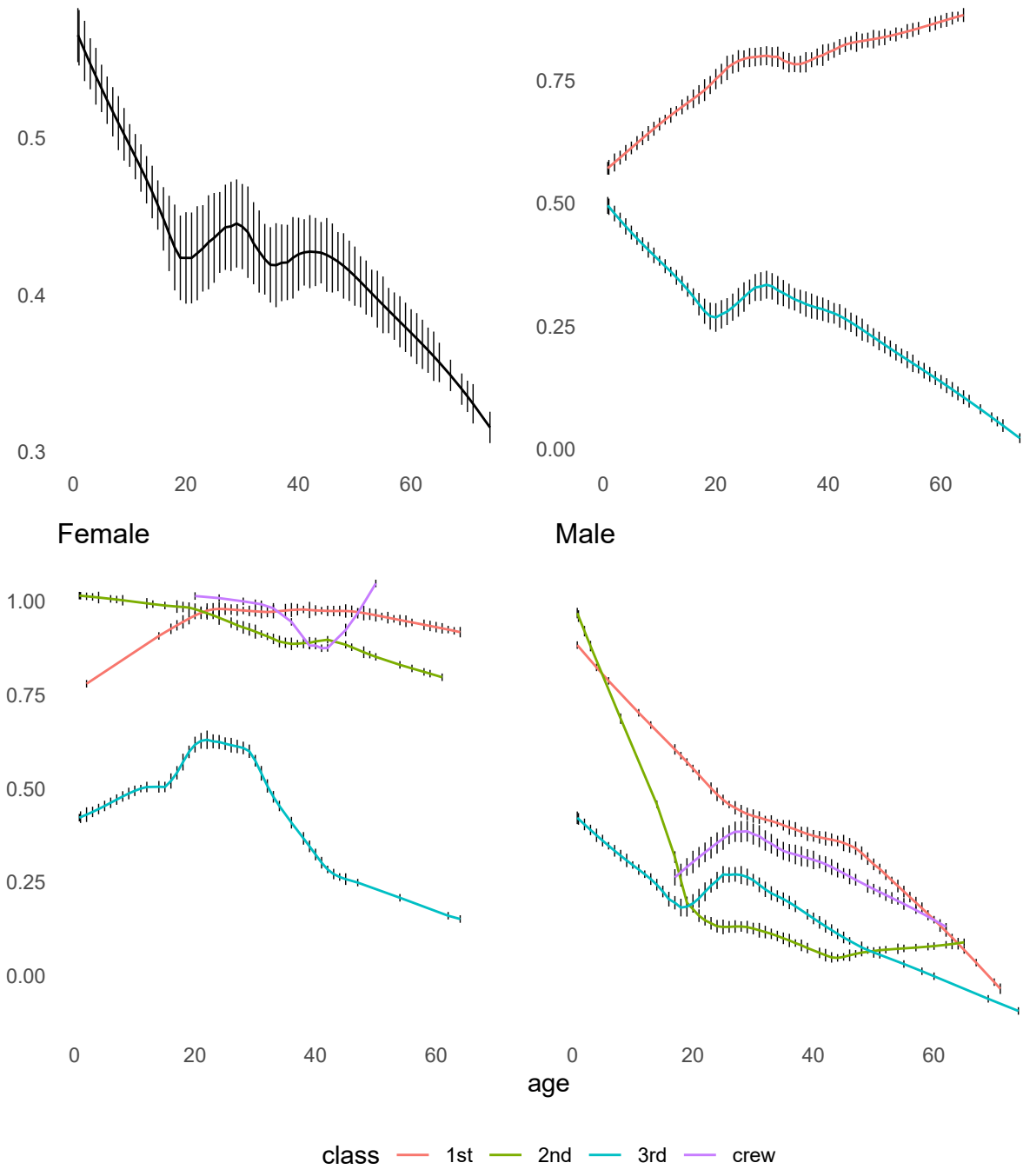

Figure 4. `loess` estimates of P(survived), with tick marks representing frequency counts within equal-width bins. Top left panel indicates nonlinear age effect without controlling other factors. Other plots give estimates under stratification by sex and class. The top right plot reflects women have a higher survival probability in general, and older females (red) and younger males (blue) are most likely to survive in their respective gender group. Two bottom plots depict survival pattern of subjects from different cabin class condition on sex.

variability and suffers selection bias. Therefore, it is essential to have extra caution, as demonstrated below using resampling, to do model simplification.

Prespecification of predictor complexity is done first by developing a saturated logistic model and then making necessary adjustments and improvements. In this model, we granted age effect maximal flexibility represented as natural splines with 5 knots, and all categorical predictors retain their original categories without pooling. Two way interactions have been specified between age and gender, age and class, and age and parent. Since this is an initial model, observations with missing age are not used. The model equation is

```
survived ~ (rcs(age, 5) + gender + class)^2 + (rcs(age, 5) * parent) +
           joined + spouse + sibling + children + nationality
```

Table 2 sees dominant main effects of gender, age and cabin class, be it linear or nonlinear ($p < 0.0001$). More notably are the strong nonlinear interaction terms between the 3 predictors. Of all 4 companion variables only parent manifests strong influence. The impact of port of embarkation is somewhat ambiguous ($p = 0.14.$). As a graphical illustration, figure 5 plots "adjusted" partial χ^2 statistic of each predictor in the saturated model, with correction for degrees of freedom allocated to them.[5] This adjustment levels the playing field for comparison of predictive ability. The larger the corresponding adjusted χ^2, the more likely a variable would have a non-flat impact on survival status.

As mentioned before, the goal of the saturated model is guiding model complexity. More specifically, should we allocate more degrees of freedom to a certain term because some complex effects has been underrepresented? Or is there a term that is highly irrelevant and could be deleted? The 5-knots natural spline on age and the resulting nonlinear interaction are promising, and further increasing knots or creating high-order interactions causes numerical problems. Therefore it is positive advantage to us to keep them as is. There are also not sufficient reasons to collapse levels for nationality and port. Binary variables like spouse and sibling have extremely large p-values ($p > 0.5$), indicating relatively small predictive power. Still, great care should be taken when one attempts to conduct aggressive model simplification based on hypothesis testing and p-values. A reliable way is using bootstrap resampling. Figure 6 studies the importance of all terms including main effects and interaction over 500 bootstrap resamples. In each resample, we fit the saturated model, rank all 13 terms by the adjusted statistic $\chi^2 - \text{d.f.}$ in ascending order so that 13 is most important and 1 is least important. The height of a bin indicates the number of times a term is ranked at that position.

The importance ranking echoes previous findings that gender, age and cabin class are predominant factors. It also reveals great variability in terms of assessing predictive power. For example, we are only confident that `joined` is not one of the 5 most influential predictors. Nonetheless, rankings of the aforementioned "weak" variables, sibling and spouse, are highly concentrated at 1 to 3. In fact, if we perform backward selection in nearly 500 bootstrap resamples with AIC as stopping rule, none of the 3 binary variables entered the selected model more than 20 times. This results in the final decision to remove them in the final model.

```
t$sibling <- NULL
t$spouse <- NULL
```

[5] The correction is done by subtracting the d.f. from the partial χ^2 statistic, its expected value under the null hypothesis.

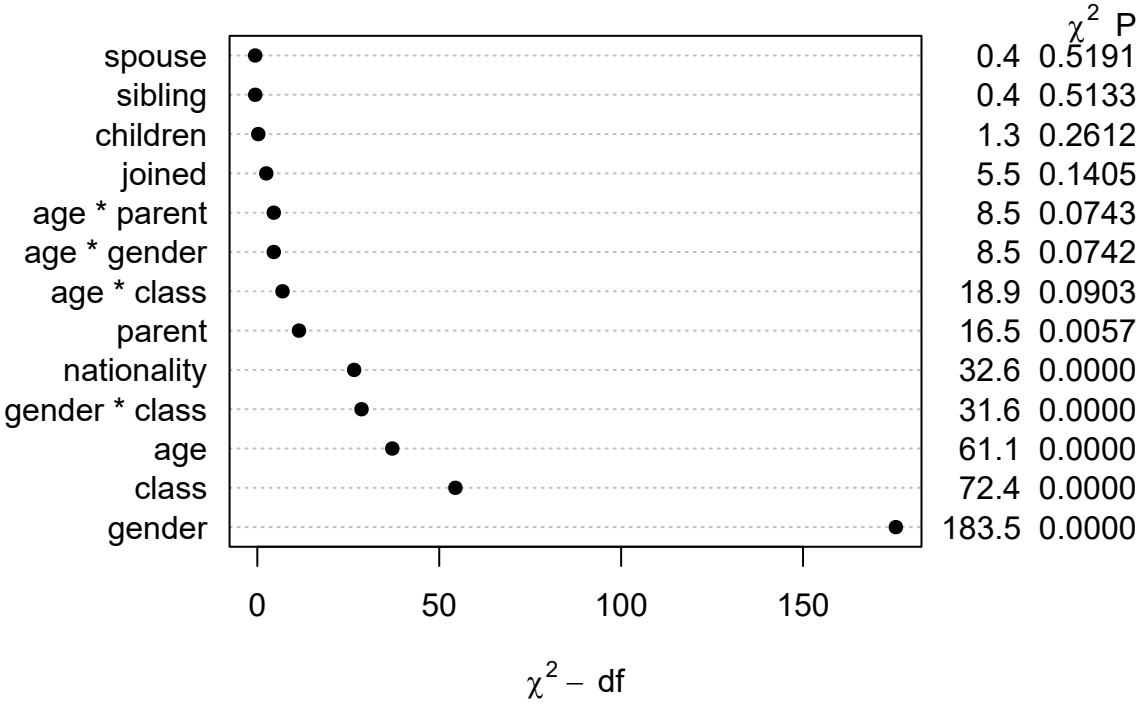

Figure 5. Ranking of predictive power in the saturated model based on adjusted χ^2

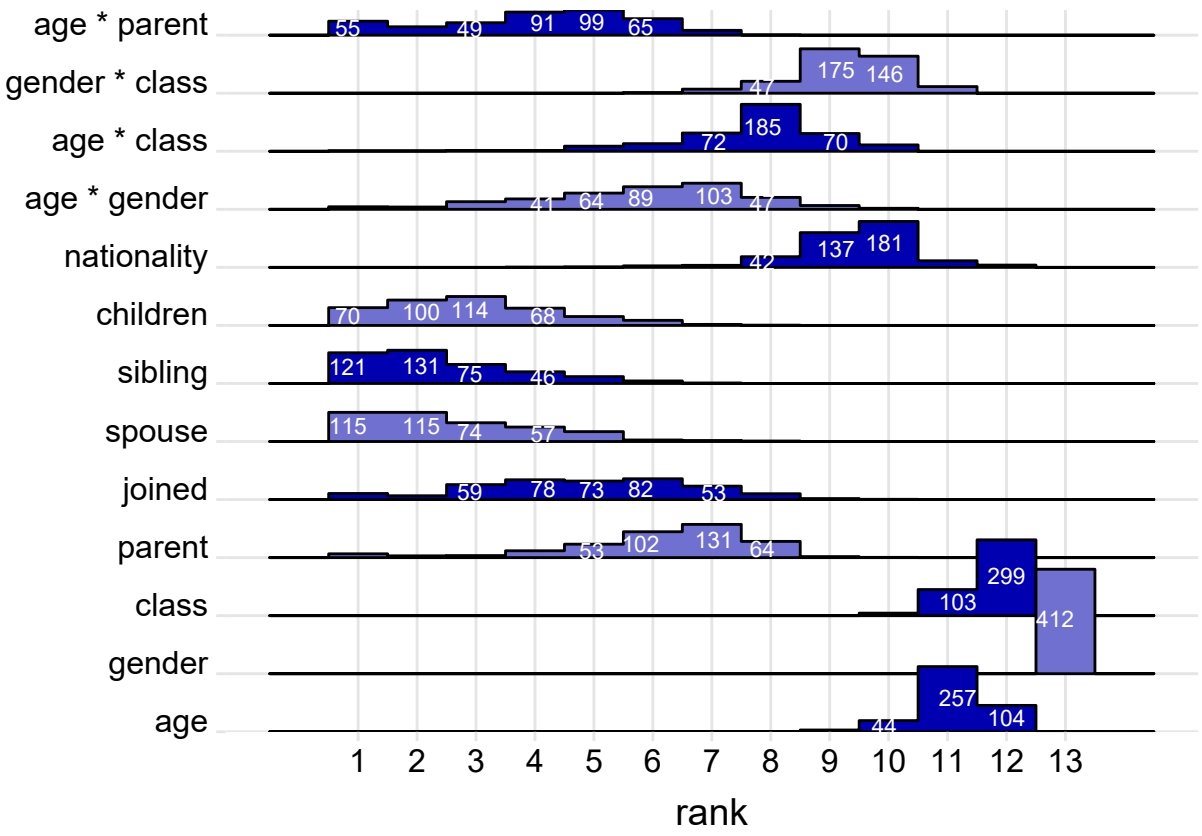

Figure 6. Distribution of importance ranking over 500 bootstrap resamples, 412 of which are actually fitted without numerical problems. Text indicates the number of times a term has a specific ranking, when the term ranked more than 40 times at that position. For example, gender ranks 13 (the most important) in all valid resamples.

Table 2. Hypothesis testing for the saturated model

	χ^2	d.f.	P
age (Factor+Higher Order Factors)	61.09	24	<0.0001
All Interactions	35.33	20	0.0184
Nonlinear (Factor+Higher Order Factors)	45.96	18	0.0003
gender (Factor+Higher Order Factors)	183.51	8	<0.0001
All Interactions	50.33	7	<0.0001
class (Factor+Higher Order Factors)	72.44	18	<0.0001
All Interactions	55.11	15	<0.0001
parent (Factor+Higher Order Factors)	16.45	5	0.0057
All Interactions	8.52	4	0.0743
spouse	0.42	1	0.5191
sibling	0.43	1	0.5133
children	1.26	1	0.2612
joined	5.47	3	0.1405
nationality	32.61	6	<0.0001
age × gender (Factor+Higher Order Factors)	8.52	4	0.0742
Nonlinear	8.09	3	0.0441
Nonlinear Interaction : f(A,B) vs. AB	8.09	3	0.0441
age × class (Factor+Higher Order Factors)	18.93	12	0.0903
Nonlinear	17.77	9	0.0380
Nonlinear Interaction : f(A,B) vs. AB	17.77	9	0.0380
gender × class (Factor+Higher Order Factors)	31.64	3	<0.0001
age × parent (Factor+Higher Order Factors)	8.52	4	0.0743
Nonlinear	1.52	3	0.6768
Nonlinear Interaction : f(A,B) vs. AB	1.52	3	0.6768
TOTAL NONLINEAR	45.96	18	0.0003
TOTAL INTERACTION	74.90	23	<0.0001
TOTAL NONLINEAR + INTERACTION	93.56	26	<0.0001
TOTAL	265.18	44	<0.0001

3.2. *Multiple imputation*

The last step before fitting the final model is imputing missing values for age. The goal of multiple imputation, in contrast to simple alternatives such as filling in conditional mean, is to provide an accurate estimate of the variance-covariance matrix that not only accounts for sampling variability, but also for the extra variance caused by missing values and finite number of imputations (Van Buuren 2018). Thus tests on individual parameters gain power and bias are reduced. The general idea is to generate multiple complete dataset, fit the model in parallel, and then obtain a pooled final estimate by averaging over all fitted models.

We use predictive mean matching with $m = 30$, since approximately 30% age are missing. The method selects a group of Titanic subjects from all complete cases that have predicted values closest to the predicted value for the subject with missing age. [6] One donor is randomly drawn from the candidates, and the observed age of the donor is taken to replace the missing value. We use the default "type 1 matching" and 5 donors (Van Buuren (2018), Section 3.4.2). Advantages of predictive mean matching in the Titanic age setting are manifold. Since imputations are based on values observed elsewhere, they are realistic (e.g., no negative age). For another, it is compatible with non-normaility which allows us to have fewer assumptions.

[6] The predicted value is generated by fitting a linear main effect model conditional on all other variables.

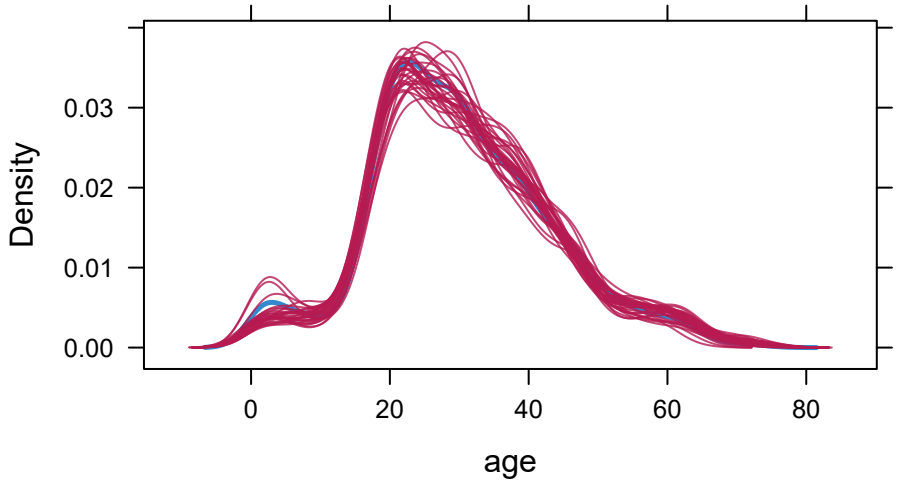

Figure 7. Density plot of observed and imputed data. In general, the imputed dataset mimic the age distribution seen in the observed data.

```
# multiple imputation with predictive mean matching to generate 30 complete dataset
imp <- mice(t, method = "pmm", m = 30, printFlag = FALSE)
```

3.3. *Model fitting, validation and calibration*

We fit the final logistic model for 30 complete dataset. Parameter estimates are obtained by averaging over all dataset. We also get an imputation-corrected variance–covariance matrix based on within– and between–imputation variances.

Table 3 again lists meaningful hypothesis testing for the final model. The χ^2 statistic of age decreased by a minor amount, resulting from using patterns of association with survival status to impute missing age. Remaining predictors generally have larger χ^2 statistic and smaller p-value compared to the saturated model in table 2, due to larger sample size in model development. Table 4 prints model assessment metrics and details individual parameters. Without validation, the model shows moderately high discrimination power, e.g. the ability to separate perished and survived subjects (concordance probability = area under the ROC curve ≈ 0.81). Brier score, as a proper quadratic scoring rule, is a promising 0.144.

	$\hat{\beta}$	S.E.	Wald Z	Pr($>\|Z\|$)
Intercept	-0.1475	1.9879	-0.07	0.9409
age	0.2225	0.1244	1.79	0.0737
age'	-0.9850	0.6877	-1.43	0.1520
age"	4.4844	3.6008	1.25	0.2130
age"'	-5.5980	4.9656	-1.13	0.2596
gender=Male	-1.8842	1.0197	-1.85	0.0646
class=2nd	5.4575	2.8123	1.94	0.0523
class=3rd	-1.0535	1.8583	-0.57	0.5708
class=crew	-2.4149	3.2651	-0.74	0.4595
parent=1	3.4640	1.2035	2.88	0.0040
nationality=English	0.0730	0.2757	0.26	0.7912

	$\hat{\beta}$	S.E.	Wald Z	Pr($>\|Z\|$)
nationality=Finnish	0.1843	0.4289	0.43	0.6674
nationality=Irish	0.0116	0.3901	0.03	0.9763
nationality=Other	-0.2328	0.2669	-0.87	0.3832
nationality=Swedish	-0.2673	0.3820	-0.70	0.4840
nationality=Syrian	-0.0582	0.4429	-0.13	0.8955
joined=Cherbourg	0.6565	0.3321	1.98	0.0480
joined=Queenstown	0.0538	0.4133	0.13	0.8965
joined=Southampton	-0.0632	0.2003	-0.32	0.7523
age × gender=Male	-0.1498	0.0573	-2.61	0.0090
age' × gender=Male	0.8213	0.3849	2.13	0.0329
age" × gender=Male	-3.9560	2.3149	-1.71	0.0875
age"' × gender=Male	4.8271	3.5665	1.35	0.1759
age × class=2nd	-0.3846	0.1754	-2.19	0.0283
age' × class=2nd	1.4771	0.8999	1.64	0.1007
age" × class=2nd	-6.7250	4.5517	-1.48	0.1395
age"' × class=2nd	8.5082	6.1731	1.38	0.1681
age × class=3rd	-0.1403	0.1159	-1.21	0.2262
age' × class=3rd	0.7699	0.6331	1.22	0.2240
age" × class=3rd	-4.9212	3.2961	-1.49	0.1354
age"' × class=3rd	7.7752	4.5292	1.72	0.0860
age × class=crew	0.0589	0.1890	0.31	0.7554
age' × class=crew	-0.0468	0.8023	-0.06	0.9535
age" × class=crew	-0.5748	3.6057	-0.16	0.8733
age"' × class=crew	1.5477	4.4606	0.35	0.7286
gender=Male × class=2nd	-0.4060	0.7087	-0.57	0.5667
gender=Male × class=3rd	2.1475	0.6288	3.41	0.0006
gender=Male × class=crew	0.5495	0.8726	0.63	0.5289
age × parent=1	-0.1190	0.1118	-1.06	0.2871
age' × parent=1	-0.6885	1.1693	-0.59	0.5560
age" × parent=1	10.4417	11.6224	0.90	0.3690
age"' × parent=1	-29.7852	35.4739	-0.84	0.4011

Although there will not be a second Titanic, making prediction a lesser problem, validation can still be used for good purposes. It quantifies the degree of overfitting by presented unbiased, optimism-corrected measures. More accurately, we will be using bootstrap internal validation to study the "future" performance of the model. In an award-winning solution to this legendary dataset submitted by IBM Watson, a holdout test set was used to validate their model. The data-splitting approach is known to require a significantly larger sample size (> 20000) than resampling methods on average to work acceptably well (Harrell, Jr. 2020b). Moreover, when the model developed on training sample is validated, the researcher would recombine training and testing set to fit a full model. This model, however, is never validated.

As a improved alternative, we choose Efron's "0.632" method for bootstrap internal validation. In each of the 494 bootstrap resamples, a model is developed and evaluated on observations omitted from bootstrap samples. Per-bootstrap optimism is then the apparent index of accuracy subtracting that in the test sample formed by omitted observations. An weighted average $\hat{\varepsilon}_0$ over all 494 bootstrap resamples is computed to estimate the true optimism, while the bias-corrected estimate of predictive accuracy is calculated as $0.632(\text{apparent accuracy} - \hat{\varepsilon}_0)$. Table 5 displays the results. It validates two general aspects of model accuracy, discrimination and calibration. Calibration is the ability to make unbiased estimates of survival status, while discrimination is the a measure in how separated predictions are for survivors and victims.

Table 3. Hypothesis testing for the final model

	χ^2	d.f.	P
age (Factor+Higher Order Factors)	57.01	24	0.0002
All Interactions	34.32	20	0.0240
Nonlinear (Factor+Higher Order Factors)	36.45	18	0.0062
gender (Factor+Higher Order Factors)	252.05	8	<0.0001
All Interactions	50.72	7	<0.0001
class (Factor+Higher Order Factors)	94.17	18	<0.0001
All Interactions	53.61	15	<0.0001
parent (Factor+Higher Order Factors)	13.82	5	0.0168
All Interactions	7.58	4	0.1082
children	5.26	1	0.0218
nationality	4.71	6	0.5815
joined	7.31	3	0.0626
age × gender (Factor+Higher Order Factors)	6.73	4	0.1506
Nonlinear	5.97	3	0.1129
Nonlinear Interaction : f(A,B) vs. AB	5.97	3	0.1129
age × class (Factor+Higher Order Factors)	16.18	12	0.1831
Nonlinear	13.64	9	0.1359
Nonlinear Interaction : f(A,B) vs. AB	13.64	9	0.1359
gender × class (Factor+Higher Order Factors)	30.97	3	<0.0001
age × parent (Factor+Higher Order Factors)	7.58	4	0.1082
Nonlinear	1.48	3	0.6870
Nonlinear Interaction : f(A,B) vs. AB	1.48	3	0.6870
TOTAL NONLINEAR	36.45	18	0.0062
TOTAL INTERACTION	74.55	23	<0.0001
TOTAL NONLINEAR + INTERACTION	86.92	26	<0.0001
TOTAL	356.20	42	<0.0001

The area under the ROC curve as well as the concordance probability is 0.8126587

$$D_{xy} = 2(c - 0.5)$$

As a integral component of validation, the calibration curve in figure 8 aims to gauge the concordance between predicted values and observed data. The 45 degree line indicates the ideal scenario in which prediction perfectly matches observation.

3.4. *Interpretation*

influence

`which.influence`

https://www.encyclopedia-titanica.org/community/threads/passengers-who-spoke-other-languages.20103/

Since the crew's instructions (in English) tended to be along the lines of "Wait down here for further orders" a lack of understanding might well have saved many lives. Also many of the immigrants in 3rd class were traveling in family or neighborhood groups which included at least one English-speaker (often an established immigrant returning to the US from a visit back home) who could act as their spokesperson.

Table 4. Model index and estimation

		Model Likelihood Ratio Test		Discrimination Indexes		Rank Discrim. Indexes	
Obs	2208	LR χ^2	783.83	R^2	0.418	C	0.813
0	1496	d.f.	41	g	1.799	D_{xy}	0.626
1	712	$\Pr(> \chi^2)$	<0.0001	g_r	6.053	γ	0.628
$\max \lvert \frac{\partial \log L}{\partial \beta} \rvert$	0.007			g_p	0.274	τ_a	0.274
				Brier	0.144		

Table 5. Optimism-corrected metrics

Index	Original Sample	Training Sample	Test Sample	Optimism	Corrected Index	n
D_{xy}	0.6253	0.6411	0.5642	0.0386	0.5867	492
R^2	0.4182	0.4364	0.3139	0.0659	0.3523	492
Intercept	0.0000	0.0000	−0.2479	0.1567	−0.1567	492
Slope	1.0000	1.0000	0.6921	0.1946	0.8054	492
$E_{\max}$	0.0000	0.0000	0.0769	0.0769	0.0769	492
D	0.3552	0.3743	0.2581	0.0614	0.2938	492
U	−0.0009	−0.0009	Inf	$-Inf$	Inf	492
Q	0.3561	0.3752	$-Inf$	Inf	$-Inf$	492
B	0.1437	0.1409	0.1506	−0.0044	0.1480	492
g	1.7827	1.9959	1.3862	0.2506	1.5321	492
g_p	0.2744	0.2815	0.2209	0.0339	0.2406	492

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.

Women and children first only for higher class passengers. If you are a third class female

5. Conclusion

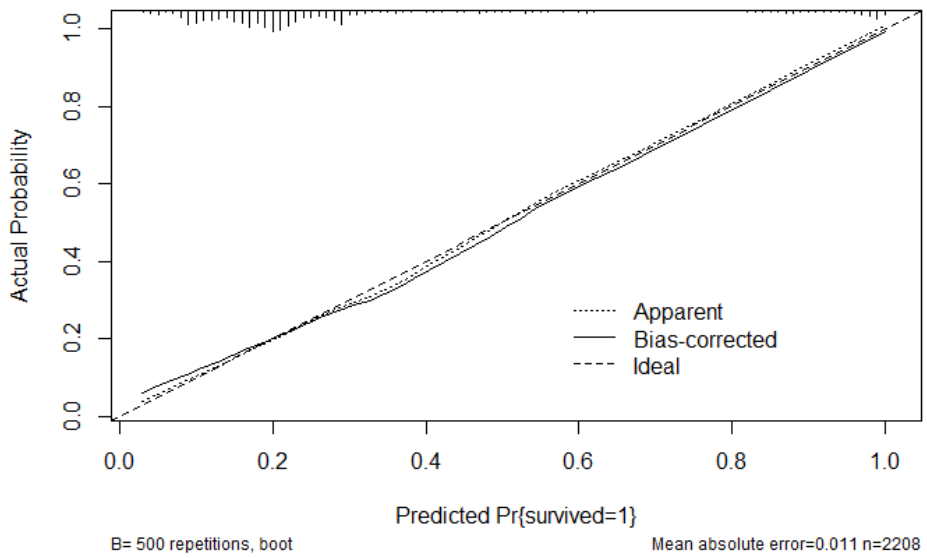

Figure 8. Calibration curve of the model output probabilites on resampled data

Appendix A. Data

A variety of other versions and forms of Titanic data sources have been collected due to public's constant interests in the tragedy as well as modern efforts trying to unveil the mystery. A comprehensive overview of several data variants is given by Symanzik, Friendly, and Onder (2018). Data in this case study 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 distinct 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. During the recent two decades the data has been constantly revised, many errors corrected, many missing ages filled in, and new variables created. Now it reflects our our most up-to-date understanding of the event, in the digital form, 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.[7] This is a predictor that has been widely reported to have good predictive ability in many submissions. However, as we see in the redundancy analysis at the end of Section 2.1, it should not even be accepted in the tentative, saturated model.

Passengers are classified according to their cabin class. Others on the vessel fall into one of crew and staff members. Crew includes victualling crew[8], 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.

Age information is presented as non-missing on the surface yet there is an indicator column representing when a person's age is only approximate and cannot be fully determined from current facts. These inaccurate age have been assigned NA. There were also ten subjects whose four companion variables were all explicitly missing. For simplicity, the mode `0` is filled in. Therefore, the problem of missing data is reduced to univariate missing of `age`.

Variables we do not utilize in this project includes name, date of birth and death, lifeboat number[9], fare, and cabin number.[10]

Appendix B. Model formula

The formula for our binary logistic model

Appendix C. Criterion used in model validation

Somer's D_{xy} index is a calibration measure, which is the rank correlation between predicted and actual response. It has a close relationship with the C index

Appendix D. Computing environment

```
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:
```

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

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

[9] 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.

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

```
[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] ggridges_0.5.2  rpart_4.1-15    patchwork_1.0.1 mice_3.11.0
 [5] rms_6.0-1       SparseM_1.78    Hmisc_4.4-1     Formula_1.2-4
 [9] survival_3.1-12 lattice_0.20-41 ggplot2_3.3.2   dplyr_1.0.2

loaded via a namespace (and not attached):
 [1] tidyr_1.1.2         splines_4.0.2       assertthat_0.2.1
 [4] latticeExtra_0.6-29 ymisc_0.0.0.9000    yaml_2.2.1
 [7] pillar_1.4.6        backports_1.2.0     quantreg_5.75
[10] glue_1.4.2          digest_0.6.27       RColorBrewer_1.1-2
[13] checkmate_2.0.0     colorspace_1.4-1    sandwich_3.0-0
[16] plyr_1.8.6          htmltools_0.5.0     Matrix_1.2-18
[19] conquer_1.0.2       pkgconfig_2.0.3     broom_0.7.2
[22] bookdown_0.21       purrr_0.3.4         mvtnorm_1.1-1
[25] scales_1.1.1        jpeg_0.1-8.1        MatrixModels_0.4-1
[28] htmlTable_2.1.0     tibble_3.0.4        rticles_0.17
[31] farver_2.0.3        generics_0.1.0      ellipsis_0.3.1
[34] TH.data_1.0-10      withr_2.3.0         nnet_7.3-14
[37] cli_2.1.0           magrittr_1.5        crayon_1.3.4
[40] polspline_1.1.19    evaluate_0.14       fansi_0.4.1
[43] nlme_3.1-148        MASS_7.3-51.6       foreign_0.8-80
[46] tools_4.0.2         data.table_1.13.2   hms_0.5.3
[49] lifecycle_0.2.0     matrixStats_0.57.0  multcomp_1.4-14
[52] stringr_1.4.0       rpart.plot_3.0.9    munsell_0.5.0
[55] cluster_2.1.0       compiler_4.0.2      rlang_0.4.8
[58] grid_4.0.2          rstudioapi_0.11     htmlwidgets_1.5.2
[61] labeling_0.4.2      base64enc_0.1-3     rmarkdown_2.5
[64] gtable_0.3.0        codetools_0.2-16    R6_2.5.0
[67] gridExtra_2.3       zoo_1.8-8           knitr_1.30
[70] readr_1.4.0         stringi_1.5.3       Rcpp_1.0.5
[73] vctrs_0.3.4         png_0.1-7           tidyselect_1.1.0
[76] xfun_0.19
```

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, Ram-

nath 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." .

Grambsch, Patricia M, and Peter C O'Brien. 1991. "The effects of transformations and preliminary tests for non-linearity in regression." *Statistics in Medicine* 10 (5): 697–709.

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

Harrell, Jr., Frank E. 2020b. "Split-Sample Model Validation." Https://www.fharrell.com/post/split-val/.

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.

Milborrow, Stephen. 2020. *rpart.plot: Plot rpart Models: An Enhanced Version of plot.rpart*. R package version 3.0.9, http://www.milbo.org/rpart-plot/index.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/.

Symanzik, Jürgen, Michael Friendly, and Ortac Onder. 2018. "The Unsinkable Titanic Data." .

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

Therneau, Terry, and Beth Atkinson. 2019. *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-15, https://CRAN.R-project.org/package=rpart.

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.

van der Ploeg, Tjeerd, Peter C Austin, and Ewout W Steyerberg. 2014. "Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints." *BMC medical research methodology* 14 (1): 137.

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.