Predict Survival on the Titanic Using R

Peng Xi
pex7@pitt.edu
School of Information Sciences
University of Pittsburgh

Yu Zhang
yuz86@pitt.edu
School of Information Sciences
University of Pittsburgh

BACKGROUND

Our work is based on a historical tragedy Titanic, which happened on the early morning of April 15th, 1912. RMS Titanic was a British passenger liner that sank in the North Atlantic Ocean during her maiden voyage from Southampton, UK. Titanic sank after colliding with an iceberg and resulted in the loss of more than 1500 passengers and crew. The tragedy was one of the deadliest commercial peacetime maritime disasters in the history.

One of the reasons that the tragedy led to such loss of life is that there were not enough lifeboats on Titanic and it happened in the early morning, so most of passengers were sleeping. There are some factors for surviving involved, and one main point is that women and children had the priority to get on the lifeboats, which was reported on news. And the passengers who had the tickets of upper-class decks and shorter journey to lifeboats, were more likely to survive.

The goal of the work is to analyze data related to passengers and predict what sorts of people had higher possibility to survive in the disaster. The datasets are sourced from Kaggle, including two historical datasets: training set and testing set, which are structured in csv files. Both sets provide passengers information: name, class, sex, age, ticket number, fare, cabin, port of embarkation, number of siblings / spouses aboard, number of parents / children aboard, while the training set has one more feature: survived or not as the response variable. So we would learn multiple logistic models and predict survivals on the testing set.

DATASET

There are 12 variables of datasets shown in Table 1. *Survived* is as the response variables and others as predictors. And the training set has 891 rows and testing set 418 rows. Here we initialize *Survived* for the testing set with value "NA" and combine the two sets for data processing, since there are some latent information could be retrieved from them, e.g., those passengers who had the same surnames. Another consideration is that there is a large proportion of missing values on *Age* and *Cabin*, which can be predicted using other features and it would benefit the accuracy of prediction on *Survived* more if two datasets are involved instead of the training set alone.

Variable	Description	Type
PassengerId	Passenger ID Number	Numeric
Survived	Survival (0 = No; 1 = Yes)	Categorical
Pclass	Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)	Categorical
Name	Passenger Name	Text
Sex	Sex (male, female)	Categorical
Age	Age	Numeric
Sibsp	Number of Siblings/Spouses Aboard	Integer
Parch	Number of Parents/Children Aboard	Integer
Ticket	Ticket Number	Text
Fare	Passenger Fare	Numeric
Cabin	Cabin (e.g., C65)	Text
Embarked	Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)	Categorical

Table 1: Variables of training and testing sets

Missing Values

Before processing the data, we need to check the missing values that should be omitted or substituted by some reasonable values. Figure 1 shows that *Age* and *Cabin* have about 20% and 70% missing values respectively, and there are only few missing *Fare* and *Embarked* values. The combinations of variables for some passengers are missing so that the model performance would gain not much contribution from them. The next chapter will introduce how to fill the missing values.

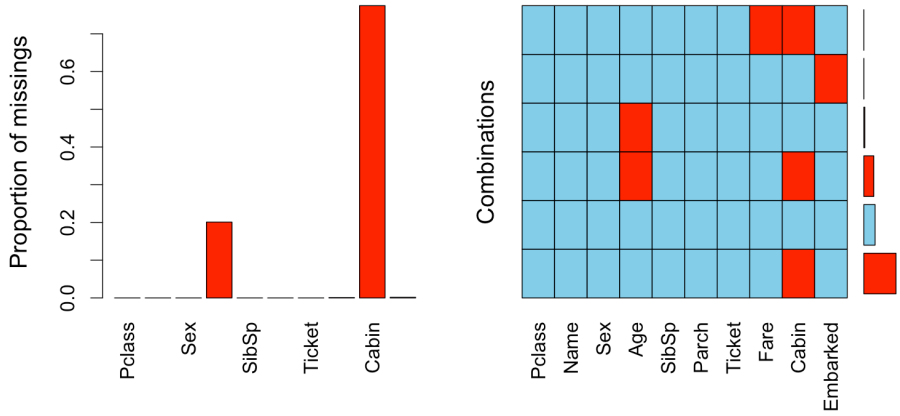

Figure 1: Missing values

Predictors

The various combinations of predictors contribute differently to the model performance. We examine each predictor to check the correlations with the response variable, or even extract new features, and then determine what combination of predictors have the best model performance. And according to the principle "women and children first", we assume that women, children

and the upper-class passengers had the priority to get on the lifeboats and more likely to survive. Namely, *Sex*, *Age* and *Pclass* would be important to predict *Survived*.

Pclass, Sex

The histogram in Figure 2 shows that correlations of *Survived ~ Pclass + Sex*. In the first and second classes, up to 80% - 90% of female passengers survived, while only half of women in the third class survived. And most of male passengers in all classes did not survive. The density plots that ages of adult passengers no matter survived or not are distributed normally for all the three classes, but the children in the second and third classes had larger possibility to survive. These findings support the assumption aforementioned that women and children were more likely to survive.

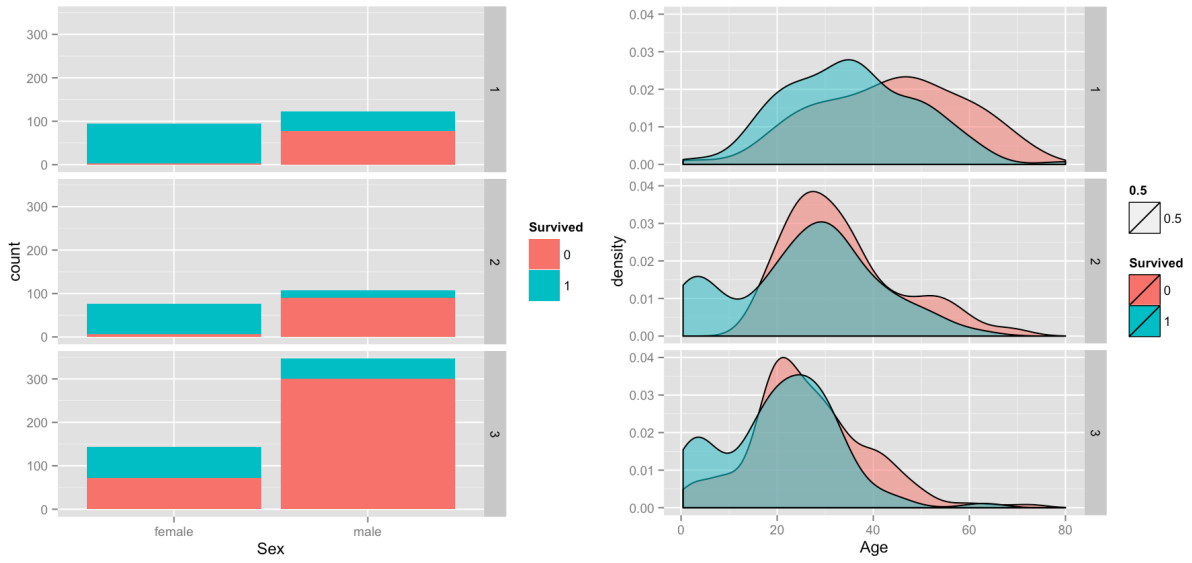

Figure 2: (a) *Survived ~ Pclass + Sex*; (b) *Survived ~ Pclass + Age*

Name

The names of passengers give not too much information intuitively, while the name is formatted in "Surname, Title. First Name", so we can extract the title and surname by splitting *Name*. There are only 875 unique surnames from 1309 names, namely, at least 434 passengers made families with some other passengers. And actually, some large families had up to 7 or 8 members, e.g., the Andersson and the Goodwin.

There are 18 unique titles retrieved from the names, but some of them can be grouped as the same one. For example, we rename "Mme", "Mlle" and "Ms" as "Miss"; "Capt", "Don", "Jonkheer", "Major" and "Rev" as "Mr"; "Lady", "the Countess" as "Mrs". After grouping, we obtain the results in Table 2. It shows that most of male passengers were named after title "Mr", and most of females were called after "Mrs" and "Miss".

Title	Col	Dona	Dr	Master	Miss	Mr	Mrs	Sir
Number of passengers	4	1	8	61	265	770	199	1

Table 2: Titles from passengers' names

Sibsp, Parch

These two variables stand for number of siblings / spouses aboard and number of parents / children aboard. Here we combine them to make a new feature named *FamilySize* which equals to (*Sibsp* + *Parch* + *1*), since the passengers probably stayed with their families together if they had ones. From Figure 3 (a), it can be concluded that large family were not positive for surviving. And also we add two more new features: *FamilyID* includes *FamilySize* (> 1) and surname if the passenger had more than one family member; *FamilyID2* includes *FamilySize* (> 2) and surname. The motivation of doing this is to feature the large families.

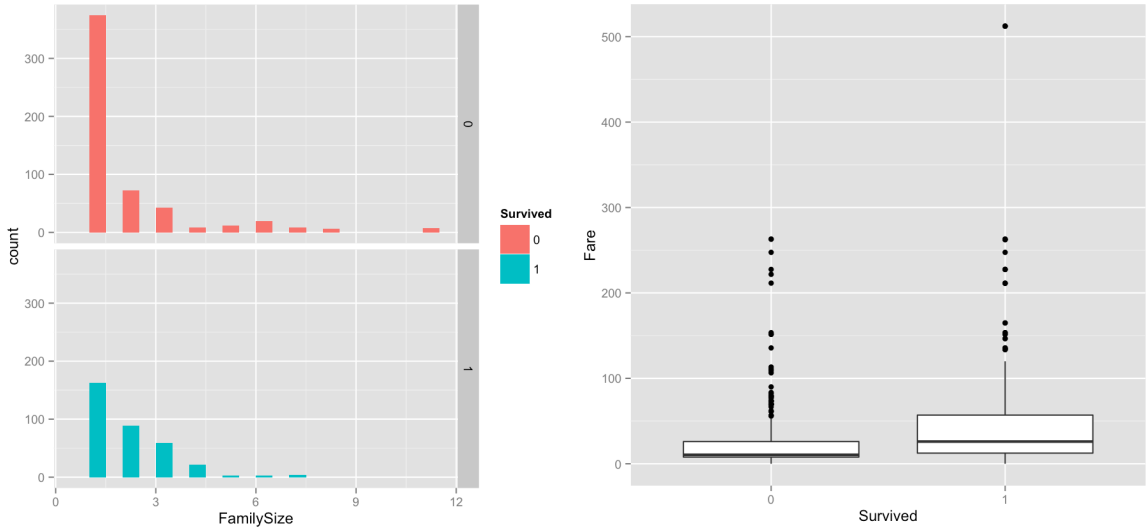

Figure 3: (a) Family size; (b) Fare

Ticket

Ticket number is a string of digits and / or letters, which could possibly mean latent information. Here we extract the prefix of the ticket number as *TicketID*, such as "A", "PC", "STON". If it contains only digits, we just count the number of digits as the *TicketID*. And there are 929 unique tickets, which means the duplicated tickets were bought for group people. It can give some latent information, e.g., ticket number "1601" appearing for 8 passengers who had no families, so they probably were friends which could not be inferred from the family information. Similar to *FamilyID*, we add a new feature *SameTicketID* which has value 1 if the ticket appears once or value contains count of ticket occurrences and ticket number. It is to feature the passengers who had the group tickets.

Fare

Passenger Fare is created a box-and-whisker plot in Figure 3 (b). It obviously describes that those passengers who purchased tickets with lower prices were less likely to survive. Overall, the passengers who did not survive had more intensive fares but more outliers than survivors, and the fares of about half of the victims were around $10. While for those group tickets, the fares need to be adjusted through being divided by the count of the same ticket occurrences. And we name the new feature as *Fare2*.

Age

Since there are about 20% missing values for Age, we train decision tree model using *Pclass* + *Sex* + *SibSp* + *Parch* + *Fare2* + *Embarked* + *Title* + *FamilySize*, then predict the missing data and fill it. And method "ANOVA" is used for the model.

As we mentioned "children first" before, but one child of 16-year-old in the first class overall had high priority to survive than one of the same age in the third class, so we differ the ages according to the passenger class by adding a new feature *Child*. If the passenger in the first class was younger than 19-year-old, set *Child* to 1; for the second class, it is 16-year-old; and for the third class, it is 14-year-old. Then it gets the plot in Figure 4 (a). Almost 100% of children in the first and second classes were survivals.

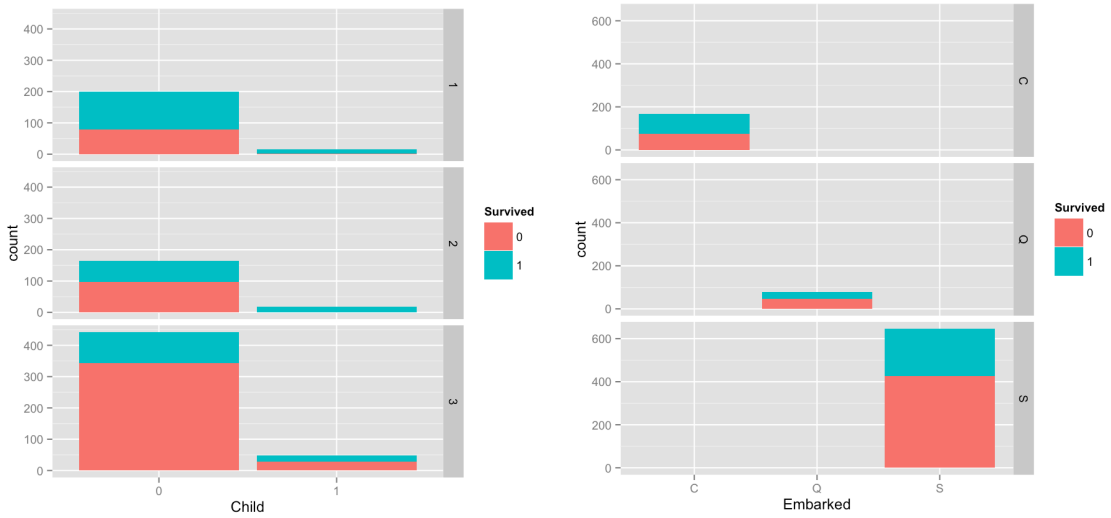

Figure 4: (a) Children; (b) Embarkation

Cabin

The data only provides 295 cabins and about 70% of cabins are missing. We extract the prefix of the cabin as *CabinID* which includes A, B, C, D, E, F, G, T as summarized in Table 3. And we impute the missing values using k-Nearest Neighbor Imputation (R method: *kNN()* in package VIM) based on a variation of the Gower Distance. But the Kaggle score for testing data was 0.79

and proved that it performed better if fill the missing data with decision tree on variables on variables *Pclass* + *Sex* + *Age* + *SibSp* + *Parch* + *Fare2* + *Embarked* + *Title* + *FamilySize* + *TicketID*.

Cabin	A	B	C	D	E	F	G	T
Number of passengers	22	65	94	46	41	21	5	1

Table 3: Cabins and the occurrences

Embarked

There are also two missing values which we just impute them with the mode "S". Figure 4 (b) shows that those passengers who boarded in port Cherbourg were more likely to survive.

Correlation

The previous work obtains the processed variables and some new features, so now we can create the correlation between variables as illustrated in Figure 5. Obviously, some variables are strongly positively or negatively correlated, e.g., *Pclass* and *CabinID*, *Fare2* and *Pclass*. In this work, we do not consider the interactions among variables. For the response variable *Survived*, the intuitively correlated variables are Sex, Pclass, Fare, SameTicketID, etc.

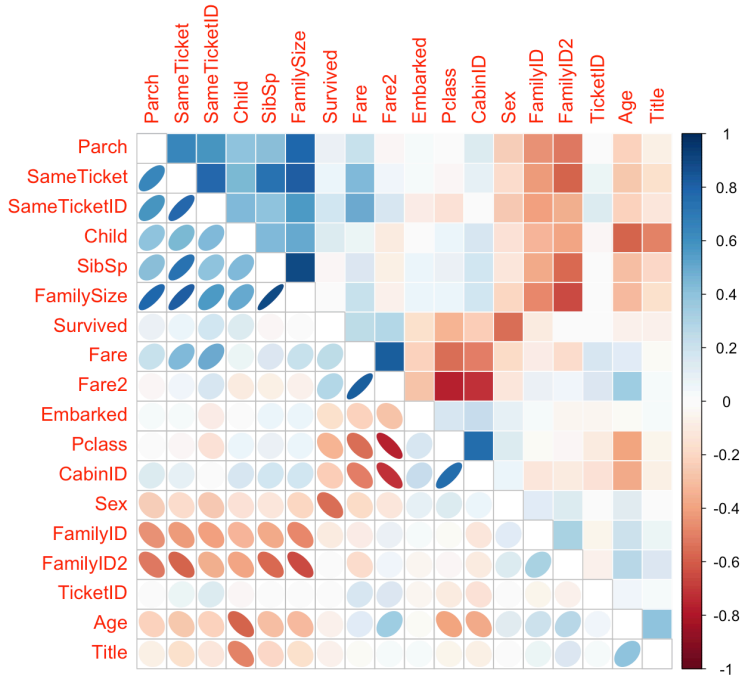

Figure 5: Correlation

METHODS

Refer to Trevor Stephens' work Titanic: Getting Started With R[1], Random Forest is an ideal method for the prediction. So we pick up this method and choose different combinations of variables to learn the model. Firstly, we try S3 method *randomForest()* from package randomForest in R to train the model, and the Kaggle score for the testing set is 0.78469. Then we turn to method *cforest()* from package party, which is an implementation of the random forest and bagging ensemble algorithms utilizing conditional inference trees as base learners. And it proves that the performance is better of Kaggle score 0.81340. Now we use this method to train models with variables and get the best performed one.

Models

The model is trained in R based on Random Forest method *cforest()*.Based on the preprocessing data, there are some important and insignificant features. Various combinations of variables could produce different contribution to the model performance. Vary the variables in the model fit formula, we can use 5-fold cross validation and obtain different mean performance measures: accuracy, precision, recall and F1-score.

Model	Formula	Accuracy	Precision	Recall	F1_Score
1	*Survived ~ Pclass + Sex + Age + Fare + Embarked + Title + FamilySize*	0.829383	0.8363485	0.8997149	0.8668062
2	*Survived ~ Pclass + Sex + Age + Parch + Fare2 + Embarked + Title + FamilySize + FamilyID + CabinID + Child*	0.8327412	0.8304258	0.9159519	0.8710278
3	*Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare2 + Embarked + Title + FamilySize + FamilyID + SameTicketID + CabinID + Child*	0.8327475	0.8294244	0.91784	0.8713101
4	*Survived ~ Pclass + Sex + Age + Parch + Fare2 + Embarked + Title + FamilySize + FamilyID + SameTicketID + CabinID + Child*	0.8349884	0.8309863	0.9197389	0.8730406
5	*Survived ~ Pclass + Sex + Age + Parch + Fare + Embarked + Title + FamilySize + FamilyID*	0.8349884	0.8316742	0.9176425	0.8724918
6	*Survived ~ Pclass + Sex + Age + Parch + Fare2 + Embarked + Title + FamilySize + FamilyID + TicketID + SameTicket + SameTicketID + CabinID + Child*	0.8361183	0.827677	0.9272111	0.874395
7	*Survived ~ Pclass + Sex + Age + Parch + Fare + Embarked + Title + FamilySize + FamilyID + TicketID + SameTicket + SameTicketID*	0.8394828	0.8319997	0.9270017	0.8767752

Table 4: Performance measures

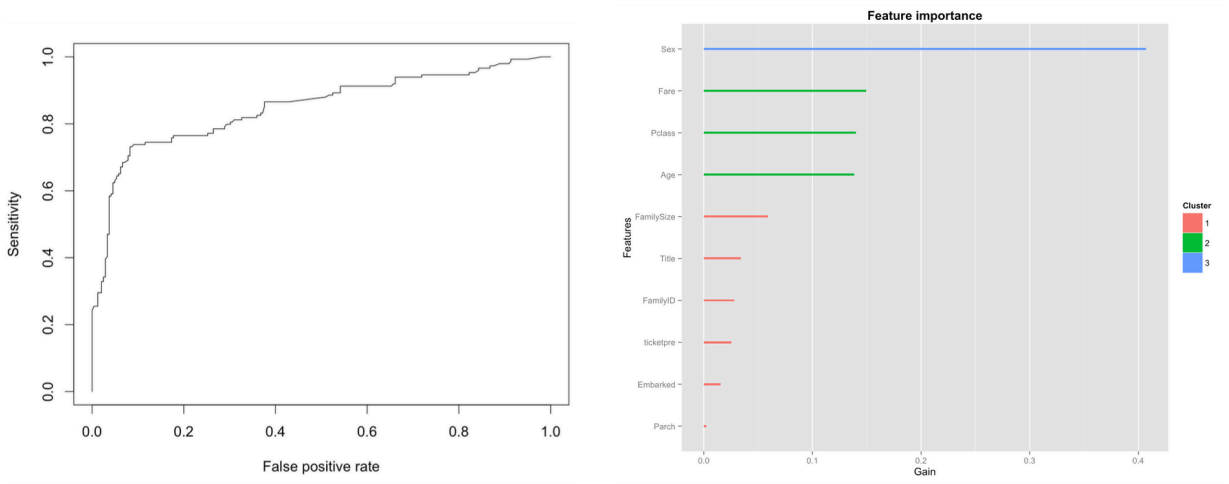

Figure 6: (a) ROC; (b) Feature importance

RESULTS

Table 4 shows that Model 7 outperforms others. And it renders Kaggle score 0.82297 for testing set while Model 6 renders 0.81818. In Figure 6, ROC curve shows that the top model achieves ideal sensitivity and AUC.

Figure 7 shows that top important predictors for model performance. The contribution to prediction can be gained from *Sex* at about 0.41. *Fare*, *Pclass*, and *Age* are secondary important to survivals. And the predictors are clustered into groups, so we could do more adjustments to the dataset of important features for improving the model accuracy or simplicity.

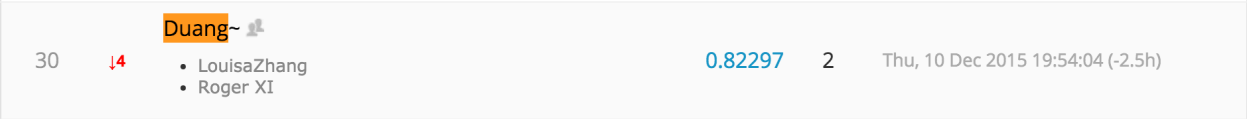

Figure 7: Kaggle score

DISCUSSION

Troubleshooting

One issue we met in the work is how to retrieve the latent information from some intuitively meaningless variables. For example, *Ticket* could be possibly related to the passenger class, the cabin, the fare, and even the group people as friends other than family information. And the model demonstrates that *Title* and *TicketID* we extract from *Name* and *Ticket* to be useful for model prediction.

Cabin is troublesome since it has a bunch of missing values. While it should help the model, because the disaster happened in the early morning, the more close to the lifeboats, the more likely the passengers will survive. Therefore, we extract the prefix of *Cabin* as *CabinID* that simplifies the model and impute the missing values with supervised learning algorithm Decision Tree. Unfortunately, the prediction accuracy improves little.

Another issue is that the method *cforest()* is time-consuming. And there are various combinations of predictors, so examine each predictor by plotting the correlations with *Survived* and pick up the most important features by 4-fold cross validation, instead of enumerating all possible fitting models.

Futrue Work

Although we try multiple solutions to improve the Kaggle score, there are still some alternative ways to be considered, e.g., the interaction between variables. The correlations show that the older the passengers, more likely they have large families.

The fare could be adjusted with the port of embarkation. As it is known, the route of Titanic went through Southampton, Cherbourg and Queenstown, so the fare of Queenstown should be less than the one of Southampton for the same cabin. Possibly some slightly adjustments could benefit the model performance.

REFERENCE

[1] T., Stephens. Titanic: Getting Started With R.
http://trevorstephens.com/post/72916401642/titanic-getting-started-with-r