MATH4995 Project1 Report
Machine Learning for Survival Prediction of Passengers on the Titanic

CAI Shizhan,
SONG Wenxin
Kaggle Team: math4995_Cai_Song

Abstract

The tragedy of the Titanic is a famous historical event. This project investigates what kinds of passengers are more likely to survive on the Titanic. To achieve this goal, we use a dataset containing information about passengers' age, gender, and some other information. Several statistical models are discussed.

1. Introduction

1.1 Objective

The main objective of this project is to fit a machine learning model to predict survival based on attributes like age, gender, and some other information.

To achieve this goal, we will perform model fitting using several popular models. The project will discuss the following models: classical logistic regression, Decision tree and Random Forest. The performance of these models is evaluated on R^2 statistics using a validation set approach.

The model should satisfy the following properties:

1) Satisfies the real-life situation
   This means that during the project we will make some process based on the common sense. Some variables may not be used, and some will be treated as categorical variables.
2) Those models which assume the underlying true model is linear should satisfy the assumptions of classical linear regression, namely:
   - Linearity: The true relation between the mean of response and independent variables is linear.
   - Independence: The error terms are independent, which can be assumed as the data values are time dependent.

All tests will be performed at the 5% level of significance using python 3.7.

1.2 Project methodology

The project will be carried out in the following steps:

Step 1: Data overview, processing, and exploratory analysis:
This step will give a brief view of the data. Exploratory analysis will be conducted to have a better understanding of variables and the relationship between them.

Step 2: Model diagnostics:
This step will deal with some substantial problems that may affect model fitting.

Step 3: Model selection:
This step we fit the data using different models and compare them.

Step 4: Result and Discussion:
At the end we will derive the final model and further discussion about the model.

2. Data

2.1 Data overview

The survival rate on the training set is 0.383838.

2.1.1 Check the missing data

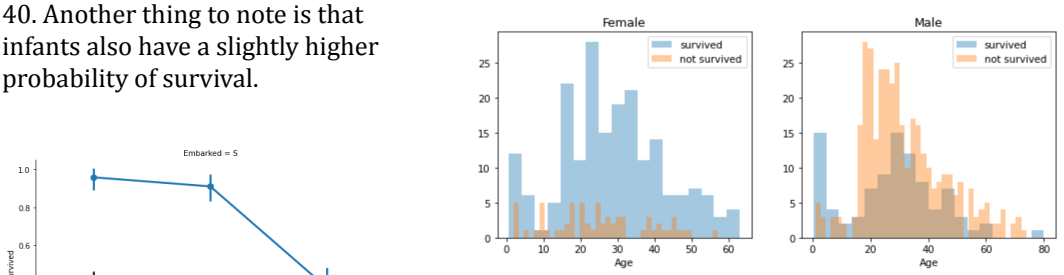
	Total	%
Cabin	687	77.1
Age	177	19.9
Embarked	2	0.2
Fare	0	0.0
Ticket	0	0.0
Parch	0	0.0
SibSp	0	0.0
Sex	0	0.0
Name	0	0.0
Pclass	0	0.0
Survived	0	0.0
PassengerId	0	0.0

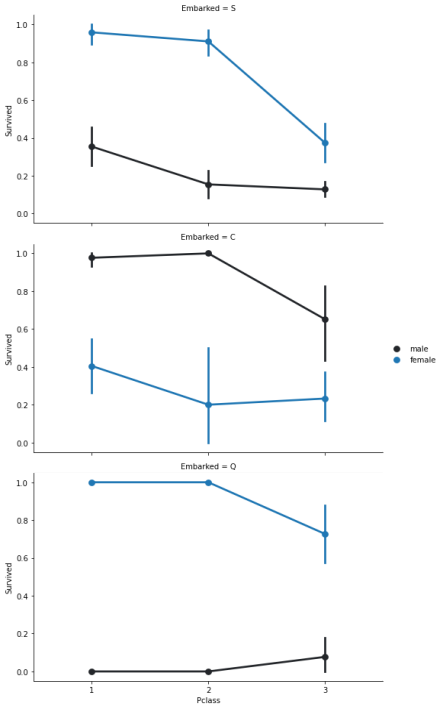
The *Embarked* feature has only 2 missing values. The 'Cabin' feature has 77 % of missing values.

2.1.2 Age and Sex

We can see that men have a high probability of survival when they are between

18 and 30 years old, which is also a little bit true for women but not fully. For women the survival chances are higher between 14 and 40. Another thing to note is that infants also have a slightly higher probability of survival.

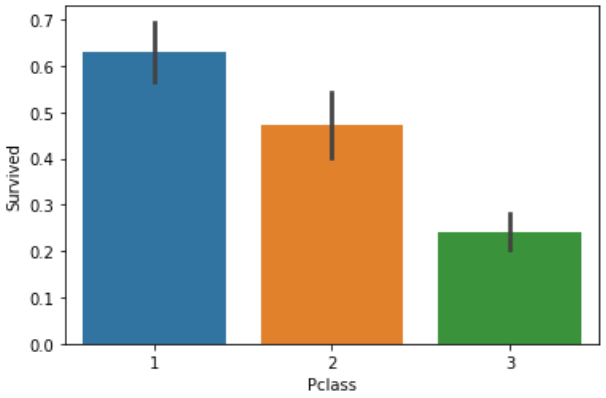




2.1.3 *Embarked*, *Pclass* and *Sex*

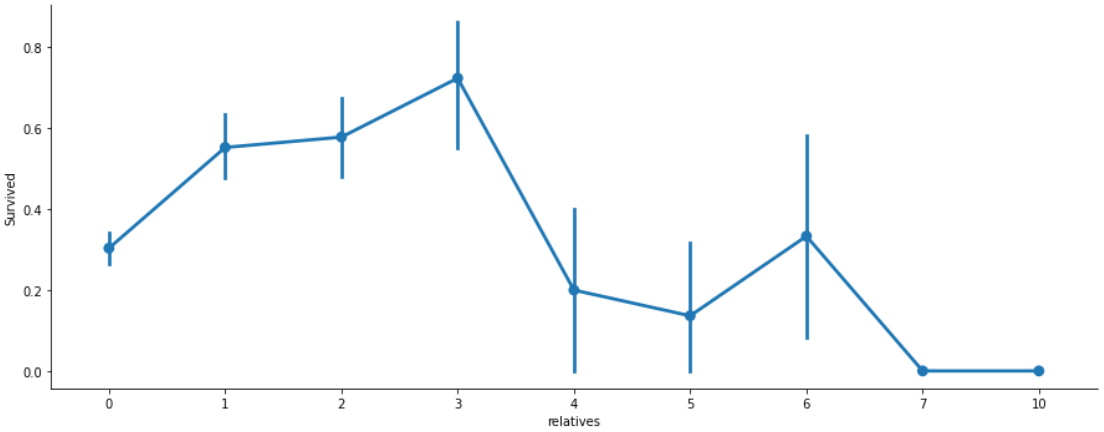
Here we could see clearly that *Pclass* is contributing to a person's chance of survival, especially if this person is in class 1. We will create another *Pclass* plot below.

Embarked seems to be correlated with survival, depending on the gender.Women on port Q and on port S have a higher chance of survival. The inverse is true, if they are at port C. Men have a high survival probability if they are on port C, but a low probability if they are on port Q or S.



2.1.4 *SibSp* and *Parch*

SibSp and *Parch* would make more sense as a combined feature that shows the total number of relatives a person has on the Titanic. I will create it below and also a feature that shows if someone is not alone.



2.2 Data Cleaning and Processing

(1) We would drop 'PassengerId' from the train set, because it does not contribute to a person's survival probability. Meanwhile, there are 681 unique tickets, so we also drop 'Ticket'. We remove 'Cabin' since they are less correlated with survival state.

(2) Deal with some NA values in the datasets

- For missing values in *Embarked* and *Fare*, we calculated the average fare for *Embarked* level C, Q and S. For missing values in *Fare,* we assign it with the average fare given the

passenger's embarked level. For missing values in *Embarked*, we assign it to the embarked level whose average fare is most close to the passenger's fare.

- For missing values in *Age*, we build a model based on other parameters to estimate it. To avoid multicollinearity, we used a non-linear model, random forest, due to its good performance on non-linear data.

(3) Converting 'Name' feature

- We will use the Name feature to extract the Titles from the Name, so that we can build a new feature out of that. However, after adding the 'Title' feature, the result in the best model has no improvement. We think the majority of the titles are 'Mr' and 'Mrs' which are included in the 'Sex' feature. It's basically useless for our model. Therefore, we decided to drop the 'Name' feature.

(4) Categorise features

- By checking the table from the data exploration part, we could find that survival rate may be correlated to 'Pclass' and 'Age'. Therefore we will create the new 'Age' and 'Fare' variable, by categorizing every age and fare into a group. However, these didn't help us to have a better performance. To keep the readability of the dataset, we just used the original 'Age' and 'Fare' as our input.

(5) Transform certain attributes

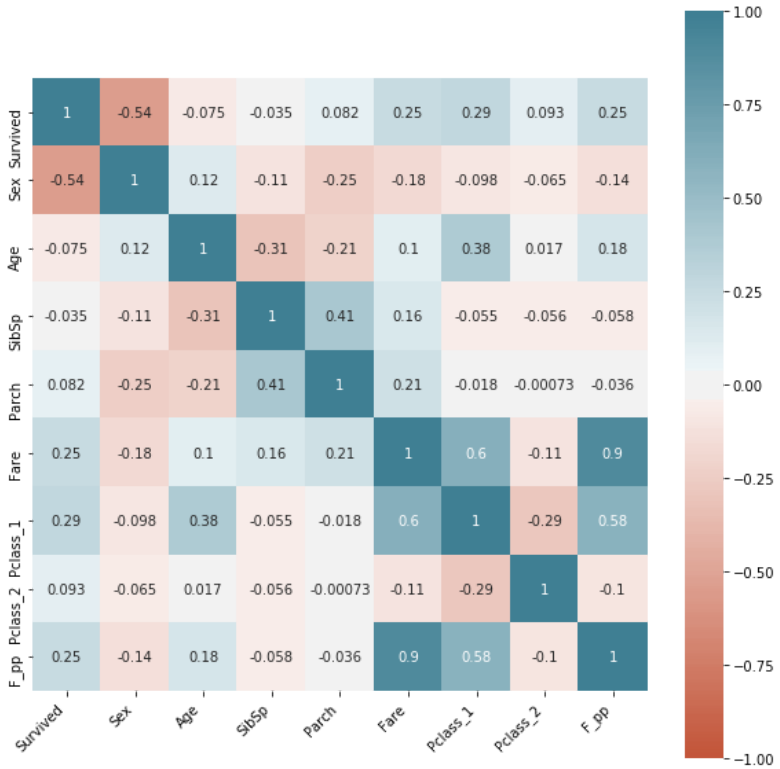
- *Fam_count*
  According to common sense, additional parents or children may have a similar effect as additional siblings do during disasters. So, we add an attribute called Fam_count to measure this effect.
- *isAlone*
  According to our commonsense, not being alone is significantly more advantageous for people when facing disasters. So, we add an attribute called isAlone to measure this effect.
- *F_pp*
  Based on our observation, we observed that some ticket fares are greatly higher than the others and the number of distinct values in tickets is smaller than the number of passengers. We suspect that some people might buy package tickets as a family or a couple. So, we use *F_pp* to calculate fare per person.

3. Model diagnosis

After processing, we decide to use the following processed variables:

Variables	Description	Remark
Sex	gender of the passenger	Binary
Age	age of the passenger	Numerical
SibSp	number of siblings/spouses aboard	Numerical
Parch	number of parents/children aboard	Numerical
Fare	passenger fare	Numerical
Family	number of family members aboard	Numerical
isAlone	Indicator of travelling alone	Binary
Pclass_1	dummy variables generated from *Pclass*, i.e. ticket class	Binary
Pclass_2		Binary
Embarked_S	dummy variables generated from *Embarked*, i.e. port of embarkation	Binary
Embarked_C		Binary

Table . Summary of variables used



Correlation Heatmap for some features

Notice that outliers and influential points mainly affect inference but not estimation, thus we do not discuss removal of outliers.

3.1 Multicollinearity check

Two explanatory variables being highly correlated will cause variance inflation and hence making it difficult to reject the null hypothesis that the related variables have coefficient zeros. We will delete the variable if the variable inflation is larger than 10.

There is no multicollinearity in the data.

4. Model selection

In this part, we will discuss 3 kinds of models: logistic regression with forward selection, logistic regression with backward selection, Decision Tree, and Random Forest. We will use a validation set approach: divide the training dataset into 712 training data and 179 validation data.

The model comparison will be evaluated on the prediction accuracy. We first use a cross validation method for feature selection. Then the model with the highest accuracy will be chosen. After tuning hyper parameters, we use it to make a prediction. To ensure randomness, we will repeat the procedure and calculate the average accuracy.

4.1 Model Comparison

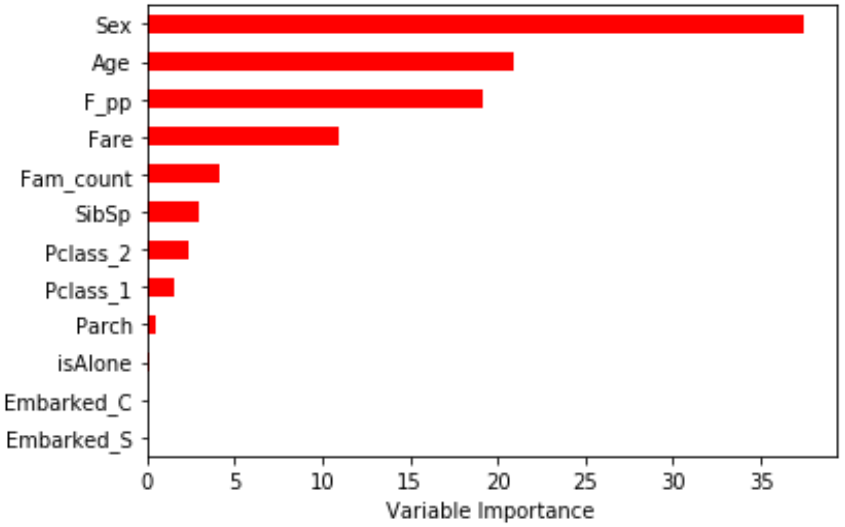
After feature selection and hyper parameter tuning, the result is shown in the following table.

Models	Accuracy
Logistic Regression with Forward Selection	0.818715
Logistic Regression with Backward Selection	0.809203
KNN	0.801387
SVM	0.807398
Decision Tree	0.811006
Random Forest	0.821792

Random forest has the best performance, and Decision Tree has a very close accuracy. This might be because the underlying pattern is not strictly linear.

4.2 Best Model

After hyper parameter tuning, the average accuracy on the validation set is 0.844021. Attached are some important features in the final model.



We could observe that sex, age, fare, number of families aboard and the ticket class play essential roles in the passenger's survival. Sex and age are of the greatest siginificance, this agrees with our common sense since these factors affect people's ability to deal with disaster. Ticket fare and class are also important, these features are highly correlated with each other, since higher class means higher ticket fare. The result meets with our data analysis on the survival rate in different classes. The first class has the highest survival rate. Number of families aboard is also an important factor. Among 3 features relevant to the number of families aboard, the number of siblings/spouses plays a more essential role than the number of parents/children. This might be because siblings/spouses are usually adults, while parents might be old people and children might be juveniles, which need to be taken care of.

5. Conclusion and Discussion

We can see that the performance of the methods is very close to each other. Random forest has the overall best performance. It does not assume linearity between response and independent variables and can handle high dimensional data. Logistic regression has a less satisfactory performance since it puts no constraints on the model and is likely to overfit.

There are some improvements we could make to improve the model. For example, we could improve the model performance by removing the outliers and influential points. Another way to improve logistic regression models is to introduce interaction terms, which could be further investigated.

To conclude, random forest has the best performance with an accuracy of 84% on the validation set. Female, young passengers who held first class tickets with families aboard were more likely to survive. Our final prediction obtained a 0.82 score on Kaggle.

Reference:

Sinan Hascelik. 2019. Logistic regression with forward/backward selection
https://github.com/talhahascelik/python_stepwiseSelection/blob/master/test.py

Contribution:

SONG, Wenxin

1. Data cleaning
2. Data processing
3. Feature selection
4. Model selection and hyper parameter tuning

CAI, Shizhan

1. Data exploration
2. Data preprocessing
3. Feature selection
4. Model comparison