Machine Learning Report
Titanic

SX1608026
马腾飞

December 18, 2016

Abstract

In this report, the author introduces the competition rules of Kaggle and puts forward solving method. Firstly, the author extracts some features form the data official given, simply, feature engineering. Secondly, the author tried some models to predict the Survived of test set, and find the alogrithm of random forest can get the best score. The author achieved the top four percent results in all the compertitors using the model of random forest. Finally, the author summarizes and analyzes the results.

Keywords: Kaggle; Titanic; Random forest; R Language

1 Introduction

1.1 Problem Description

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships.

One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.

In this challenge, we ask you to complete the analysis of what sorts of people were likely to survive. In particular, we ask you to apply the tools of machine learning to predict which passengers survived the tragedy.

1.2 About the data

The subject give us two data.one of the data is training set ,which mean that we have knew the personnal's information and he was death or not;and another data is testing set that we could know some information about the person,but we could not know he was death or not.

Every person's information include that name, sex, age, sibsp, parch, ticket, fare, cabin, embarked etc in the data set and survived included in training set.More information about variables if you want to get,you can refer to the website of https://www.kaggle.com/c/titanic/data.

1.3 Evaluation

The historical data has been split into two groups, a 'training set' and a 'test set'. For the training set, we provide the outcome ('ground truth') for each passenger. You will use this set to build your model to generate predictions for the test set.

For each passenger in the test set, you must predict whether or not they survived the sinking (0 for deceased, 1 for survived). Your score is the percentage of passengers you correctly predict.

1.4 Analysis of the subject

It is obvious that we need to make a binary classification model to predict that the people at the shipwreck was death or not. It has so many classification model, for instance, SVM(Support vector machine), Logistic regression, Naive Bayes, Decision tree, random forest etc, but which one should we choose? After several attempts, we found that the results of the model of random forest is excellent in these models we tested. So, finally, we chose the model of random forest to predict.

2 Feature Engineering

Data plays an important role in any model, because as a teacher said, the quality and quantity of data decide the upper bound of results, and every model we used just approch the upper bound. Therefore, we must spend a lot of time to clear our initial data and carefully select data feature.

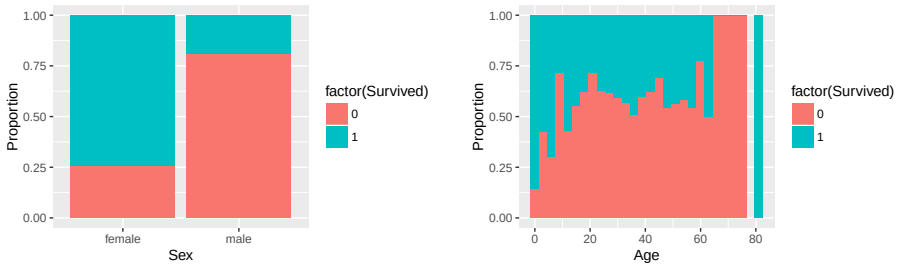
2.1 Data cleaning

We noticed that some variables have missing value, for instance,'Age','Cabin' and Fare. The variable of Fare's has just one missing value, so we would use the mean of the variable to instead of the missing value. It exits a number of missing value for both Age variable and Cabin variable, and we will use two methods to deal with the problem. For Age variable, we utilized the

package of mice to fill the missing values. For the variable of Cabin, firstly, all missing value was insteaded of the variable of 'N'.

2.2 Feature selected

We can easily guess that gender and age may be the critical factor that the person is survived, because we preferentially rescue women or kids in the event of an accident. Let's vertify our conjecture. As is shown in figure 1, we can see that the proportion of survivors in females is more than males. In the other words, gender is almost determined to Survived. Let's analyse another feature variable of age. we can notice that the trendency of the rate of survived is declining along with the increase of age. Therefore, our conjecture is correct!



(a) Survived and Gender (b) Survived and Age

Figure 1: Influence of Gender and Age on Survived

If one person has a higher status, he may has more chance to be survived. Look at the figure 2. The higher class, the greater the possibility of rescue. So, the Pclass is selected to be feature vriable in the next model.

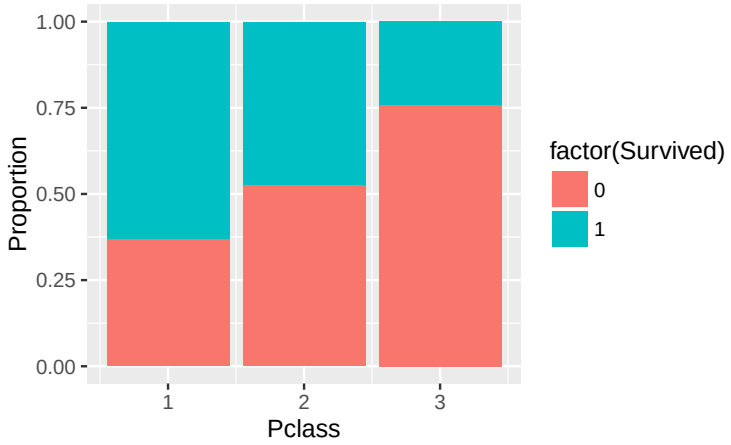


Figure 2: Influence of Pclass on Survived

Next, we observe how the Fare influnce the survived. Looking at the figure 3, we can clearly find that if you had a higher price ticket, you would have more chance to enter the lifeboat. So, the Fare is selected too.

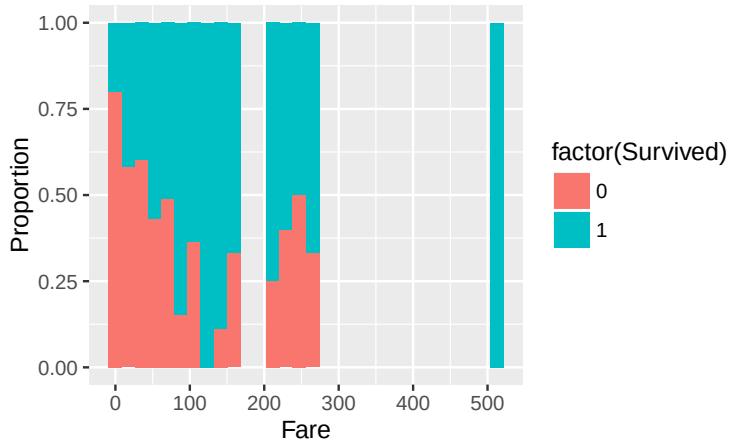


Figure 3: Influence of Fare on Survived

We drive a new feature variable which is used the two variable of Parch and Sibsp, and we name it FaminlySize. A person's title can response many information. Therefore, we extract the person's title from the variable of Name, and name it Title. If some persons have the same amount FaminlySize and the same firt name, we will say they are a family. A family may ride a lifeboat together, so we define the feature variable of Family by FamilySize and first name.

I will further deal with the feature variable of Cabin, because I think Cabin may play a key role when the ship hit the iceberg. The Cabin will be devided into three parts according to Survived. In train set, if all persons in a cabin are survived, we set it 'S'(survived), and if all persons in cabin are unsurvived, we set it 'US'(unsurvived), and if some persons are unsurvived and some persons are survived in a cabin, we set the cabin with 'UK'(unknow).

All above feature vriable and Embarked will be used in the next model.

3 Method of the project

3.1 Motivation of the method

Before did the competition, I have been familiar with some algorithms, for instance, SVM, LR, Naive Bayes, Decision tree, Random forest, and I have utilized these models to test the accuracy of prediction. I fuond that the model of Random Forest has the best performence in these models, followed

by SVM. So, I want to mainly introduce the model of Random Forest in this report. Firstly, I will introduce the algorithm of Descion Tree.

3.2 Decision tree

Decision tree is a commonly method of mearchin learning. A decision tree is a flowchart-like structure in which each internal node represents a 'test' on an attribute , each branch represents the outcome of the test and each leaf node represents a class label. The paths from root to leaf represents classification rules. Reference to algorithm 1.

Algorithm 1 Decision Tree

Input:

Train set $D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\}$

Attribute set $A = \{a_1, a_2, \cdots, a_d\}$.

Process:

1: generate a node;
2: **if** all samples in D belong to the same category C **then**
3: mark the node for C;
4: **end if**
5: **if** A $= \emptyset$ **or** all samples in D have same value **then**
6: mark the node for leaf node and its class is marked the majority class in D;
7: **end if**
8: select the optimal attribute $a_\star$ from A;
9: **for** all $a_\star^v \in a_\star$ **do**
10: generate a branch of node;
11: D_v denotes the subset of D which $a_\star = a_\star^v$;
12: **if** $D_v = \emptyset$ **then**
13: make the branch node to be leaf node and its class is maked the majority class in D;
14: **return**
15: **else**
16: make the TreeGenerate(D_v,A,$a_\star$) to be brach node;
17: **end if**
18: **end for**

Output: a descion tree.

3.3 Random Forest

In recent years, ensemble learning has attracted the attention of many researchers. In statistics and machine learning, ensemble methods use multiple

learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone.

Random forests is one of the kind of ensemble learning. Random forests is a substantial modification of bagging that builds a large collection of de-correlated trees, and then averages them. On many problems the performance of random forests is very similar to boosting, and they are simpler to train and tune. More concretely, for all nodes of decision tree, firstly, Random forests randomly select a attributes subset including with k attributes from the node of attributes set, and then a optimal attribute selected will be used partition.

4 Results and analysis

4.1 Experimental environment

All analysis and experiment in this report was used R and the R version is 3.3.1. In order to analysis the subject and utilize some mature models, we used some package,including 'dplyr', 'random forest', 'party' 'DMwr', 'ggplot2', 'e1071', 'mice' etc. The result of this report is generated by the package of 'party', which include a random forest process 'cforest'.

4.2 Results

We make a lot of experiments in my PC using the R package and use the combination of different feature variables. The best score is 0.818(accuracy rate). The rank of my best is 234th/6096(top 4%). But in this model of best score, feature variable are just including with Pclass, Sex, Fare, title, Embarked, Family, FamilySize. When we use the feature variable of Age or Cabin, the result indicate correct rate is decreased. The reason why the correct is decreased I guess may be the two variable have too many missing value. Look at the figure 4, we rank feature variable according to its importance. As is shown in the figure, the most important variable is title, followed by Sex, and then Pclass. This fits our daily logic.

4.3 Thoughts and ideas

I have tried many ways including combination of some models or combination of kinds of variable. But the corecct rate is not very good compared to other competitor, for instance, someone's rate more than 85%. I think there are two reasons why the corecct rate is not so high. For one thing, I can't find the right model to train the data due to the alogrithm I konw too little. I will continue to study the alogrithm of mearchin learning, although I am just a student of mathmatic. For another thing, there are too many thing

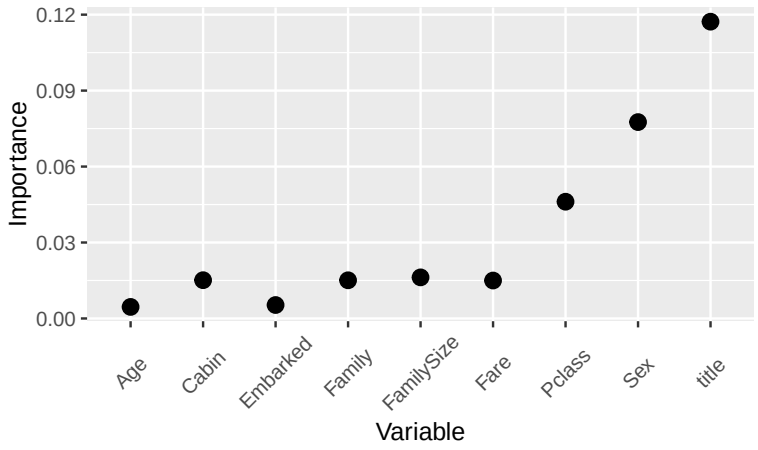


Figure 4: The importance of feature variable

to do in feature engineering, and too many unkonwn informations waitting us to mine.

Finally, though this practice, I learned many kownledges and techniques, such as coding, data plotting, because for a math student, there are too little chances to do practice.

References

[1] 机器学习[M]. 清华大学出版社, 2016.

[2] Trevor Stephens. "Titanic: Getting Started With R" http://trevorstephens.com/ kaggle-titanic-tutorial/ getting-started-with-r/

[3] Stef van Buuren etc, November 9, 2015. "Package 'mice' ". https://cran.r-project.org/web/packages/mice/mice.pdf.

[4] Torsten Hothorn etc. November 28, 2016. "Package 'party' ". https://cran.r-project.org/web/packages/party/party.pdf.