What sorts of people were more likely to survive?

WONG, Wing Kin

Introduction

We mainly focus on feature selection process. With the right feature, the prediction result will be much better and indicate that those feature determine if that person will survive or not in the titanic.

We first try to understand the characteristics of the training dataset and try to plug in into different model to see whether it will affect the prediction result.

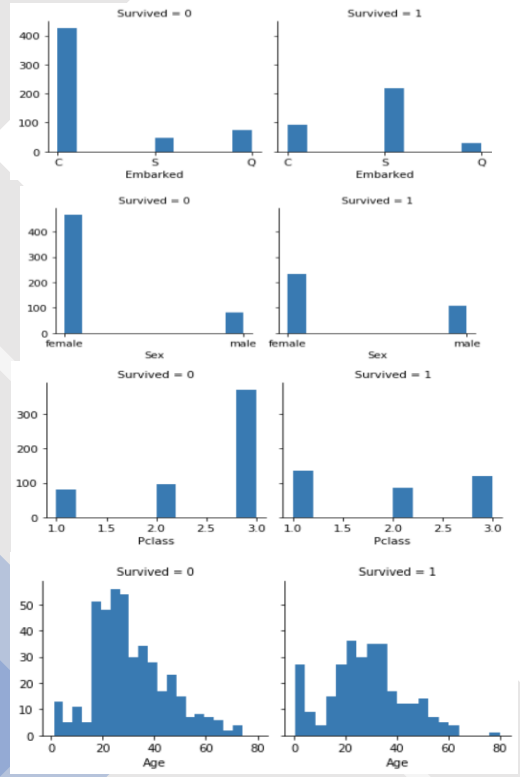
Attributes:

- Passengerid
- Pclass (1 / 2/ 3)
- Name
- Sex (M / F)
- Age
- SibSp
- Parch
- Ticket
- Fare
- Cabin
- Embarked (C / S / Q)

EDA

First, we plot some graph to have a basic understating of the data. From the graph, we can observe that with the following value in each attribute, they were more likely to survive.

- Sex = Male,
- Pclass = 3,
- Embarked = C
- Fare < 50



Linear Regression

To find out which feature is important (or not that important), we can consider the p-value in regression model.

We have fit all the attribute into the model and see the p-value of each attribute. The larger the p-value is, the feature may not be important.

From the table below, we can see that embarked = Q / C are not that related, however, embarked = C may be an indication to predict the survival.

Also, Fare and Parch are also not that related since the p-value are very high.

We have selected some attributes in the below table and let's fit it into a single perceptron for prediction.

	coef	std err	t	P>|t|
Intercept	1.4031	0.082	17.214	0.000
Sex[T.male]	-0.4854	0.032	-15.389	0.000
Embarked[T.Q]	-0.0987	0.082	-1.198	0.231
Embarked[T.S]	-0.0664	0.040	-1.674	0.095
Pclass	-0.1874	0.023	-8.183	0.000
Age	-0.0064	0.001	-5.661	0.000
SibSp	-0.0508	0.017	-2.912	0.004
Parch	-0.0107	0.019	-0.561	0.575
Fare	0.0002	0.000	0.565	0.572

Perceptron

If the attribute are highly correlated with the survival, then fitting it into the perceptron should be able to predict the survival accurately.

We have done some data conversion, such as convert the categorical data into one-hot vector so that the data can be used in Perceptron, For the null value, we simply remove it. We have split the data into training set and validation set, and only the validation set will be evaluated. The result are below:

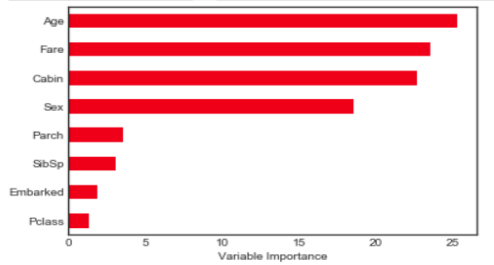
All attributes: 68.15%
Pclass, Sex, C, Q, S :75.14%
Pclass, Sex, Age, :72.62%
Pclass, Sex : 72.06%
Pclass, Sex, C, :75.41%
Sex, Age, :78.77%

We observed that with only using the sex and age can achieve the highest accuracy, meaning that these 2 attributes should be the most important. However, with the other attributes, the accuracy will become lower. This is because the other attribute may not contain the survival data (i.e. those attributes are mostly noise.)

Tree

We have also fit the data into the tree model, and the graph on the right shows the variable importance .

Cabin are shown to be important in this model.



Conclusion

Since the problem is not linearly separable, using the above model is not enough to achieve 100% accuracy. We find that sex and age are the most important factor among different model. (i.e. male and adult are more likely to survive) However, there are some inconsistency in the tree model which indicate that fare and cabin are also important factor to affect the survival.