Prediction of Titanic Survival Rate

This report aims at making use of Machine Learning to identify essential features that will influence someone's survival rate in the Titanic Disaster, use cleaned data to train model, and then apply best model to make predictions. The report is organized by describing the process of cleaning and exploring data, showing the process of training model, and providing some suggestions.

Cleaning data

By checking training dataset's structure, the 'Cabin' variable has 687 null values out of 891 rows. If 'Cabin' is filled with some other values, it will increase too much inaccuracy because of too many null values. Hence, this variable cannot be kept. For 'PassengerId', 'Name' and 'Ticket', they just show passengers' basic information and the texts will be hard to operate for further analysis, which indicates they have to be removed.

After removing invalid variables, some null values of other variables need to be filled. On the one hand, 'Age' has 177 null values, which will be filled by some random values between minimum and maximum age. The reason why it is modified in this way is because random values will increase much more accuracy compared with mean or most frequent ones. On the other hand, 'Embarked' (the ports passengers have embarked) has too null values. And it is filled with highest frequency value – 'S'.

Part of the training dataset now should be like this:

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	0	3	male	22	1	0	7.2500	S
1	1	1	female	38	1	0	71.2833	C
2	1	3	female	26	0	0	7.9250	S
3	1	1	female	35	1	0	53.1000	S
4	0	3	male	35	0	0	8.0500	S

Exploratory data analysis

This part focuses on digging out the relationship between survival rate and other variables. And after determining the necessary features, some of them will be adjusted or transformed to other forms for final training.

In the beginning, a heatmap shows that actually every existing variable will affect passengers' survival rate. (The heatmap will be shown in the Kaggle link) And then, the barplots have been used to uncover the detailed connections between survival rate and other variables. For some categorical data, such as 'Sex', 'Pclass'(class of travel), 'SibSp'(number of sibling/spouse), 'Parch'(number of parent/child) and 'Embarked', the codes can plot them directly. The x-axis is

one categorical variable and y-axis stands for the respective survival rate. However, these variables have to be transformed to make sure that they can be put into models. Every variable needs to be changed into dummy variables, using zero and one to show all of the information. For example, 'Pclass' have three levels. It will be separated into two dummy variables. One/zero, zero/one and zero/zero mean different levels.

The most difficult part of dealing with these raw variables is how to modify 'Age' and 'Fare', which are both numerical ones. The more suitable way is to normalize them, putting their values in [0,1]. By doing this, 'Age' and 'Fare' variables are consistent with other dummy ones.

Part of the result after all the operations is like:

	Survived	High Class	Median Class	Sex	Age	SibSp	Parch	Fare	Embark C	Embark Q
0	0	0	0	0	0.27	1	0	0.0141	0	0
1	1	1	0	1	0.47	1	0	0.1391	1	0
2	1	0	0	1	0.32	0	0	0.0154	0	0
3	1	1	0	1	0.43	1	0	0.0103	0	0
4	0	0	0	0	0.43	0	0	0.0157	0	0

Train model and give prediction

When the training dataset is prepared, it will be manipulated by different machine learning models, including Gaussian Model, Perceptron Model, Logistic Model, RandomForest Model, DecisionTree Model, etc. The model with highest score will be chosen as best prediction model. When the model is determined, test dataset should be cleaned and modified as the training dataset. Then it will be applied into prediction model to get final forecast. The final accuracy of the prediction is 0.77.

Provide some suggestions

According to exploratory data analysis, women, child and older people tend to have higher survival rate. And passengers with higher fare and higher class also tend to survive with higher chance. Based on these, the administration should find some ways to make the survival rate of men, middle class and lower-fare-passengers higher. Therefore, they can increase many more lifeboats and life buoys to take many more passengers away and save more lives. Furthermore, they should arrange different life channels to the passengers with different class and fare levels. By doing this, they can avoid chaos and make the rescue operations orderly. With these measure, they can make sure that more people will survive in future possible disasters.

Kaggle address: https://www.kaggle.com/zhenyufan/machine-learning-of-titanic-new-beginner