Project Report-EE551A-2019Spring

Name: Jingxuan Liu CWID: 10441874

Introduction

‘Titanic: Machine Learning from Disaster’ is an interesting competition about a famous event sinking of Titanic of in history. We can get some information about people in this event. The data is not very large so it is suitable project for a beginner of machine learning like me.

This competition is from Kaggle and provide us with two data set, train.csv(training set) and test.csv(test set). Our propose is to use the training set to establish a suitable model and predict output of the test set. Therefore, we need to find which features influence the rate of surviving, and

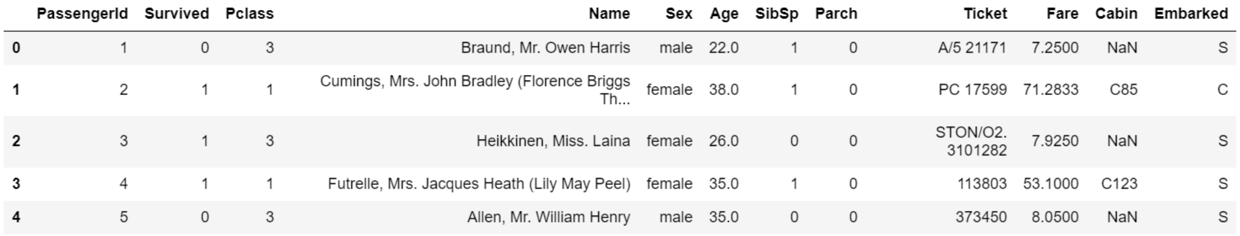
My code is divided into three part: data analysis, feature selection and machine learning.

Tools: Python, Jupyter notebook, Anaconda

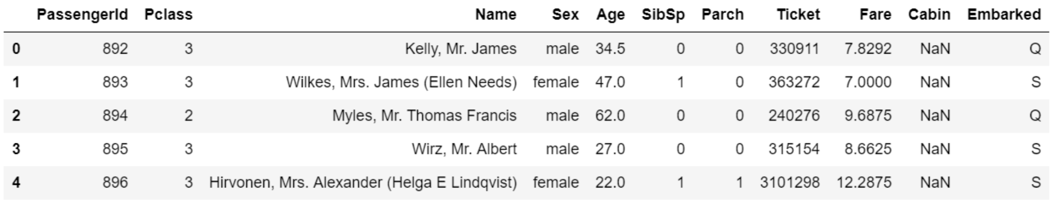
Data Analysis

First, download the training set and test set from Kaggle. Install the Anaconda in computer and install the Jupyter notebook in Anaconda. Jupyter notebook is a good tool to write python code and we can easily install necessary packages in python by environment in Anaconda such as numpy, pandas, matlabplot and so on.

We need to open the csv file, train.csv and test.csv in jupyter notebook by pandas function read_csv(). Display the first five line of the training set and test set.



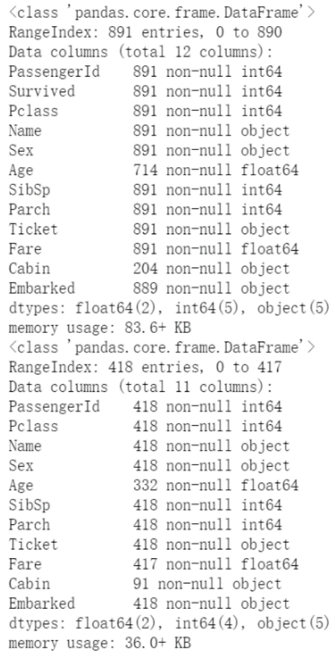
	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S



	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	Q
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	S
2	894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	Q
3	895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN	S
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	NaN	S

From the figure above, training set has one more label, ‘Survived’, than test set and the features ‘Name’, ‘Sex’, ‘Cabin’ and ‘Embarked’ are string while others are number.

Then, use the .info() function to show the overall information in two dataset.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
PassengerId    891 non-null int64
Survived       891 non-null int64
Pclass         891 non-null int64
Name           891 non-null object
Sex            891 non-null object
Age            714 non-null float64
SibSp          891 non-null int64
Parch          891 non-null int64
Ticket         891 non-null object
Fare           891 non-null float64
Cabin          204 non-null object
Embarked       889 non-null object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.6+ KB
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 418 entries, 0 to 417
Data columns (total 11 columns):
PassengerId    418 non-null int64
Pclass         418 non-null int64
Name           418 non-null object
Sex            418 non-null object
Age            332 non-null float64
SibSp          418 non-null int64
Parch          418 non-null int64
Ticket         418 non-null object
Fare           417 non-null float64
Cabin          91 non-null object
Embarked       418 non-null object
dtypes: float64(2), int64(4), object(5)
memory usage: 36.0+ KB
```

In training set, there are 891 objects but 3 features have missing value. In test set, there are 418 objects but 3 features have missing value.

Then, we need to fix missing value. In training set, I use mean of all values of ‘Age’ expect missing value missing value and use mode to fix missing value of ‘Embarked’ because it only loses two values. Because of too much missing value, I don’ t fix the missing value of ‘Cabin’ and I need to see the relationship between ‘Survived’ and whether value is missing later.

Feature Selection

1.PassengerId

PassengerId is just the label of each people. Therefore, I think this is a useless feature and I will delete it later.

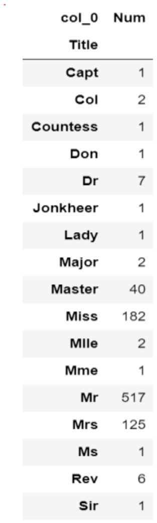
2.Plass

‘Plass’ can be 1, 2 and 3. Plot the line graph.

3.Name

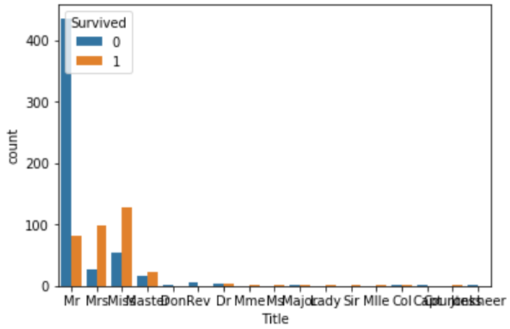
This feature is not clear to reflect some useful information. So, I establish 2 new features, ‘Title’ and ‘len_Name’ to represent title of people and the length of name.

According to the order ’A’ to ‘Z’, I list all the titles below:

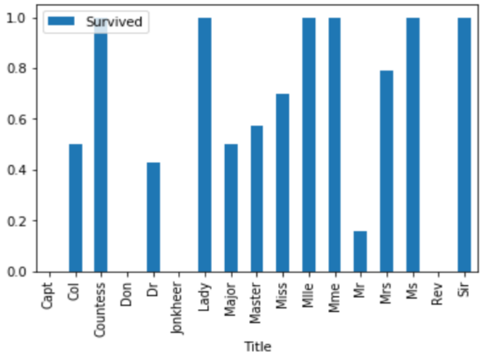


col_0	Num
Title	
Capt	1
Col	2
Countess	1
Don	1
Dr	7
Jonkheer	1
Lady	1
Major	2
Master	40
Miss	182
Mlle	2
Mme	1
Mr	517
Mrs	125
Ms	1
Rev	6
Sir	1

Plot as bar graph:



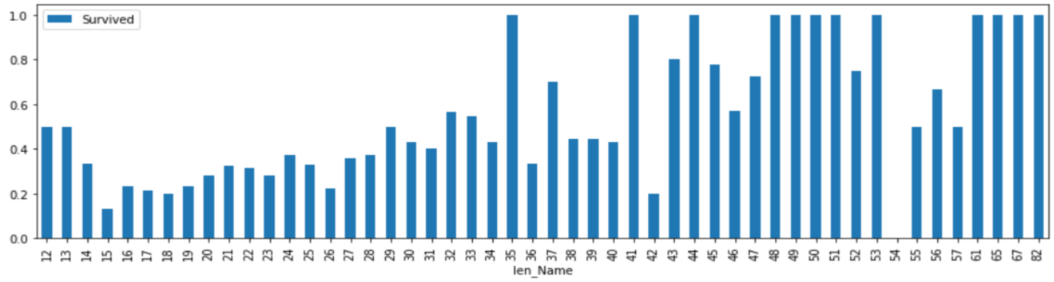
However, it is exactly not clear. So, I use the mean of the title which means the number of survived people divided by number of all the people of each title. This is actually equal to the rate of surviving of each title. Plot new graph:



Then, I met some problem here because I use every title as a feature after getting dummy value. But the title in training set is more than that in test set. Therefore, I need to do more process of the title.

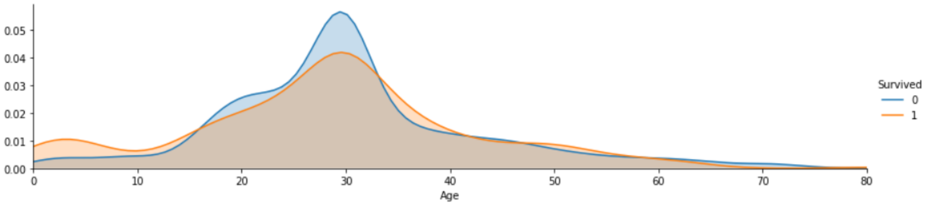
I consulted method(https://blog.csdn.net/Koala_Tree/article/details/78725881), and divided titles into 6 labels as new ‘Title’ feature.

Use same way above to plot graph of ‘len_Name’:



4. Age

First, I plot two violin graphs to show situation when ‘Survived’ = 0 or 1. However, the violin chart can’ t show the difference of the survived people in every age group. So, I plot new graph:



Then, I separate values in four age group with the boundary ’12’, ’18’ and ’65’: children, youth, adult and old as features.

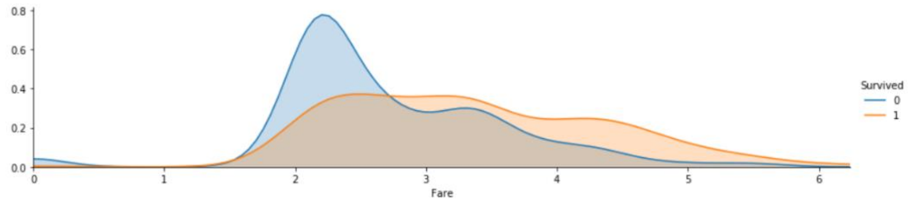
5. SibSp & Parch

These two features are similar and I do the same process for them: separate values in three group, ’large’, ‘middle’ and ‘small’.

6. Fare

Plot the violin graph. I find it is not well-distributed graph. So, I use the log(x+1) to normalize x and get new graph.

This feature is similar to ‘Age’ and I use the same method to show the data:



Then, I separate values in two group with the boundary ‘2.7’: rich and poor.

7. Cabin

Plot the bar graph of ‘Cabin’ and I find that whether value is missing influence the rate of survived. Therefore, separate values in two group: no and yes.

After showing all the data in training set, I get dummies from all the features and calculate the correlation between ‘Survived’ and other features. List the correlation. Then , delete output ‘Survived’ and features with lower correlation(‘Name’,’PassengerId’,’Ticket’).

As for the test set, first, I also need to fix the missing value ‘Age’ by mean and ‘Fare’ by mode. And do the same process of each features as training set do.

Machine Learning

First, import the packages, I decide to use the decision tree, random forest, SVM and K- nearest neighbors algorithm I learned from the class CPE695-Machine Learning and CPE646-Pattern Recognition.

In addition, I also use the two boost algorithms: Adaboost and Xgboost. Because boost algorithm is suitable for the classification and can improve accuracy to large extent.

Due to the small size of dataset, I use the grid search to train the model. In addition, set the range of parameters of each algorithm.

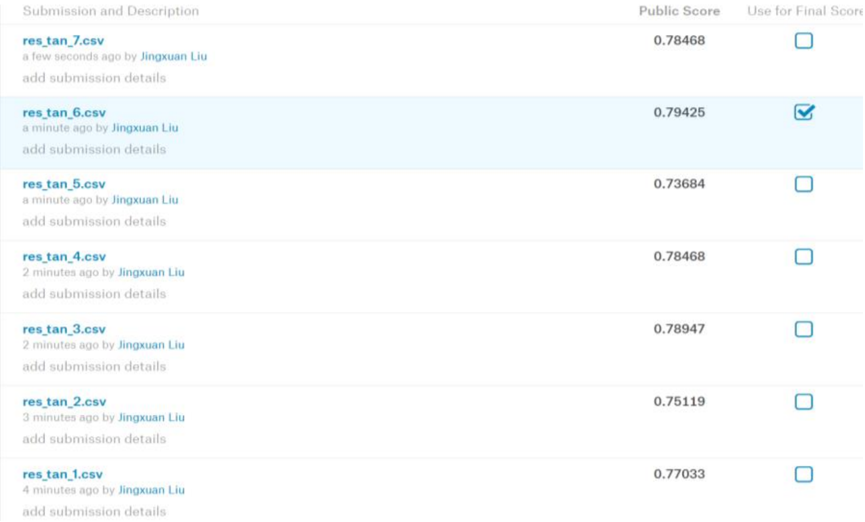
Then, when I set range of parameter ‘min_samples_split’ as (1, 10) in decision tree and random forest algorithm, I got error that this parameter can bot be 1. So, I change the range from (1,10) to (2,10) to solve this error.

After parameter adjustment, I get 6 trained model(clf1,cf2, clf3, clf4, clf5, clf6) from decision tree, SVM, random forest, adabost, KNN and xgboost.

Save these output to 6 files(res_tan_1.csv, res_tan_2. csv, res_tan_3. csv, res_tan_4. csv, res_tan_5. csv, res_tan_6. csv).

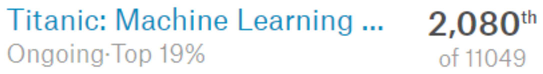
Then, I use the majority vote algorithm to get new output file ‘res_tan_7.csv’ from 6 outputs.

Then, upload the result files to Kaggle:



Submission and Description	Public Score	Use for Final Score
res_tan_7.csv a few seconds ago by Jingxuan Liu add submission details	0.78468	□
res_tan_6.csv a minute ago by Jingxuan Liu add submission details	0.79425	☑
res_tan_5.csv a minute ago by Jingxuan Liu add submission details	0.73684	□
res_tan_4.csv 2 minutes ago by Jingxuan Liu add submission details	0.78468	□
res_tan_3.csv 2 minutes ago by Jingxuan Liu add submission details	0.78947	□
res_tan_2.csv 3 minutes ago by Jingxuan Liu add submission details	0.75119	□
res_tan_1.csv 4 minutes ago by Jingxuan Liu add submission details	0.77033	□

We can see the res_tan_6.csv get the highest score which uses the xgboost algorithm.



Titanic: Machine Learning ...
Ongoing·Top 19%

2,080th
of 11049

I get 2080th of 11049.

From results, the xgboost algorithm is better than other algorithm in this competition. We can see different algorithm can influence a lot about the accuracy. In addition, I tried to delete the lower correlation feature ‘Age’, but I got the worse result than before. After reading karnels in Kaggle, I find I can improve my accuracy in the data process. For example, using mean to fix ‘Age’ feature can be replaced by use the random forest by other features to fix missing value.