Data Mining Term Project Report

20161190 leeminjae

Problem definition

Our purpose is to predict whether passengers are alive or not, by the various features of Titanic passengers. I will make a hypothesis by graphing the survival results according to various features. Then I'll remove unnecessary features and patch up missing values according to circumstances. Then, I will make a model by various ways, evaluate and compare it. In summery, Our goal is the creation and evaluation of a model that determines y variale(Survival status), by the features of people on the Titanic.

Hypothesis construction and Exploratory data analysis

First, I looked at the features of the data. Categorical types include Survival, Pclass, and Sex, while numerical types include Sibsp, Parch, Ticket, Fare, Cain, Embarked features.

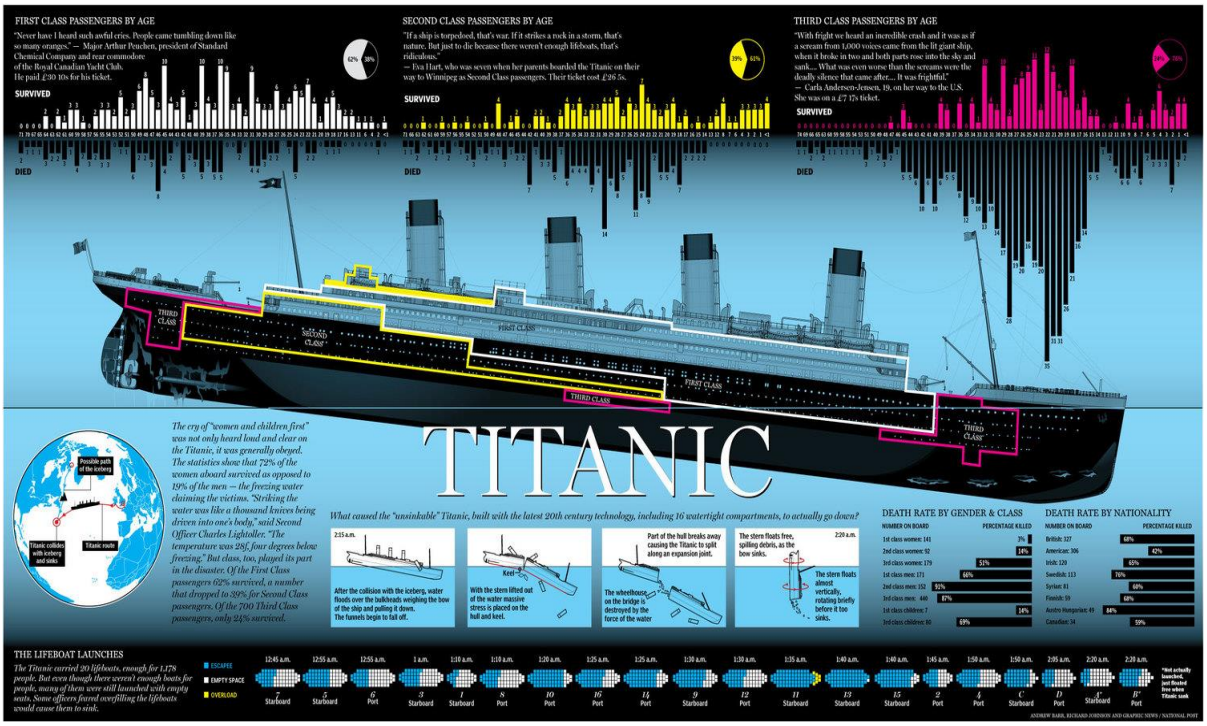

Figure 1. Titanic

Figure 1 shows that the third-class compartment, the pink border on the right, sank first. In other words, people in the third-class compartment would be relatively more likely to die, and first and two class would be less likely to die. Pclass will also be associated with fair, so the probability of death will vary depending on fair. Also, I will check the correlation graph later if there is anything related to Pclass. It is also said that the first-class compartment and the life-saving tube were close. Age and gender would also have affected survival probabilities, as there was also a cultural tendency to rescue women than men and children and the elderly than adults. It is also expected that more people who

come together than alone will have a higher chance of survival by saving or cooperating with each other.

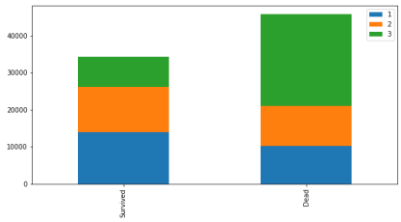

Figure 2-1. Pclass

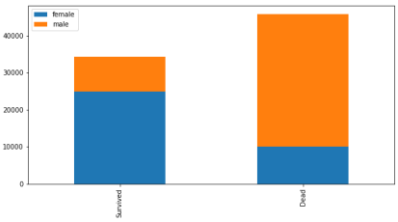

Figure 2-1. Sex

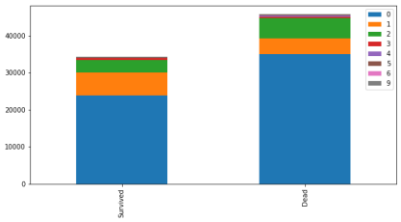

Figure 2-1. Sibsp

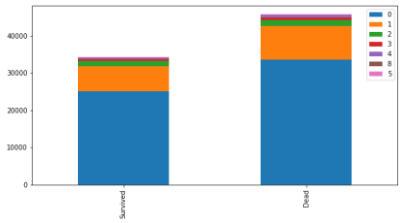

Figure 2-1. Parch

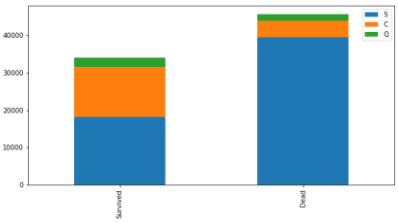

Figure 2-1. Embarked

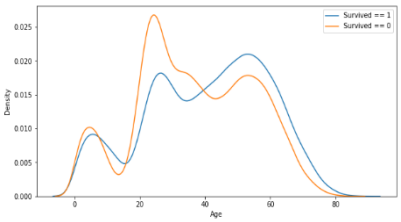

Figure 2-1. Age

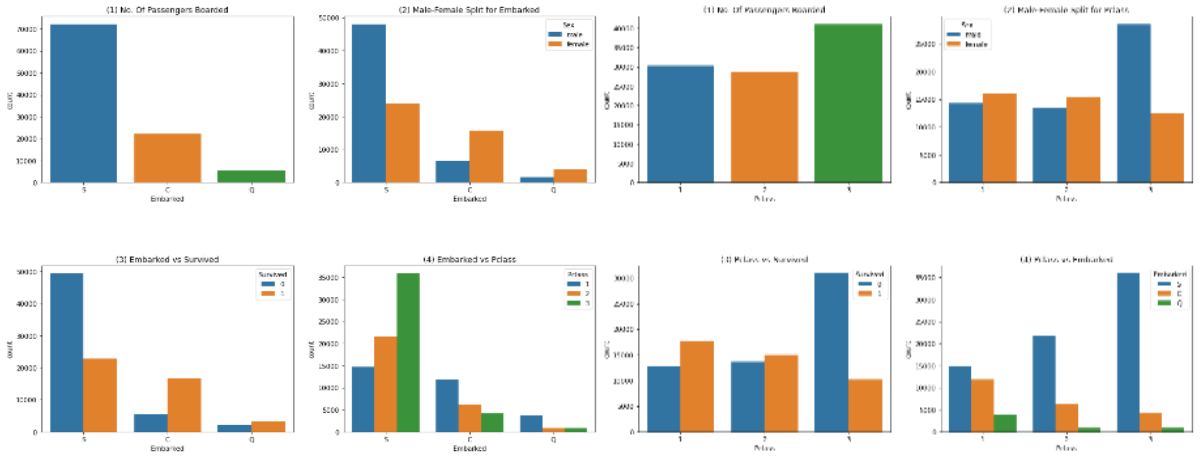

Figure 2-2. Pclass_detail

Figure 2-2. Embarked_detail

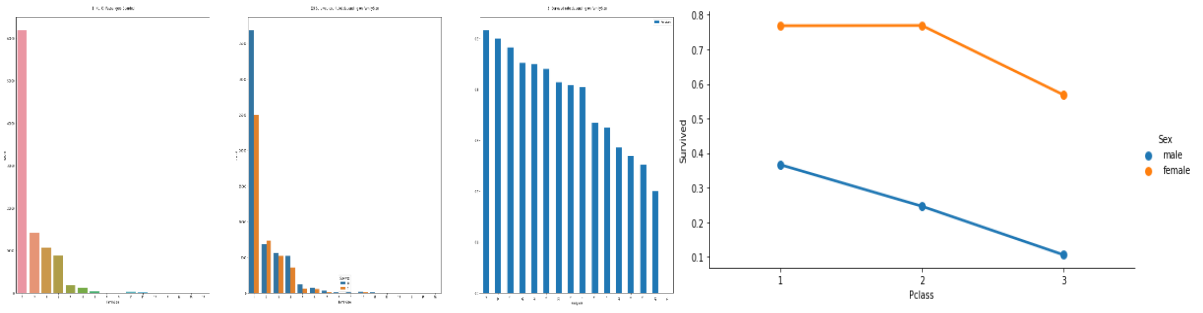

Figure 2-2. Family

Figure 2-2. Sex- Pclass

Pclass Figure : As expected in the first figure, many people died in the third class. As expected in the first picture, many people died in the third class. The first and second classes died at a similar rate. Unlike other classes where male and female rode similarly, male rode twice as much in third class as female. Most of them were embarked S in all class, but especially in the third class, Most were embarked S.

Sex Figure : Male have died more than female. At figure 2-2, male died more than female in all classes. Gender can also be seen to have a significant impact on survival probability. Also, third class male has very little chance of survival.

Sibsp and Parch : Both Sibsp and Parch were more likely to die alone. Both of them are similar features, so they are combined into Family. As expected, those who came alone had the lowest chance of survival in Family. The case with the highest chance of survival is when two people come together, probably because they are a couple so cooperate to save each other.

Embarked : Mostly Embarked S died a lot. C lived rather a lot, and Q had a similar survival rate. Male rode twice as much as a woman in S. The rest, female rode more than male. Also, third class was used a lot in S and first class in C and Q. In C and Q, relatively wealthy people would have ridden.

Age : Many people, mostly in their 20s and 30s, died. Surprisingly, young children had a half-and-half survival rate, while those in their 40s and older had a slightly higher chance of living. Perhaps people in their 20s and 30s are trying to save people of different ages, resulting in this result.

Data pre-processing

I first confirm the distribution of Survived through Figure 3. Since it is supposed to be Balanced, I did feature engineering without any other procedures. And checked copy data for preventing damage in raw training data

I changed feature Sex and Embarked to number in order to make it easier to learn model. In case of Sex, female was changed to 0, and male was changed to 1. Case of Embarked, changed C to 0, Q to 1, and S to 2.

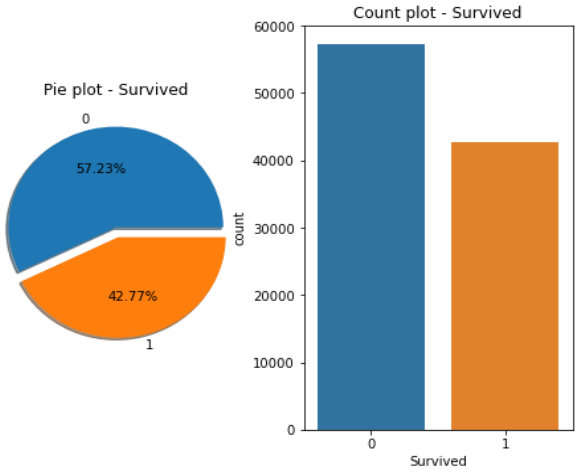

Figure 3. Survived distribution

Feature extraction : I checked missing value by using df.isnull().sum(). And I dropped four features that I thought were unnecessary for a few reasons.

Passenger id : Passenger id dropped because it is just index value.

Name : In Name, no useful information was found and dropped because it was difficult to parameterize.

Ticket : Ticket seemed to be able to guess the location of the rooms or the number of floors, but Fare and Pclass contain the information so i dropped ticket

Cabin : Cabin dropped because there were too many missing value. Also, I saw strange data such as a good room but low fair value. So I threw it away more boldly.

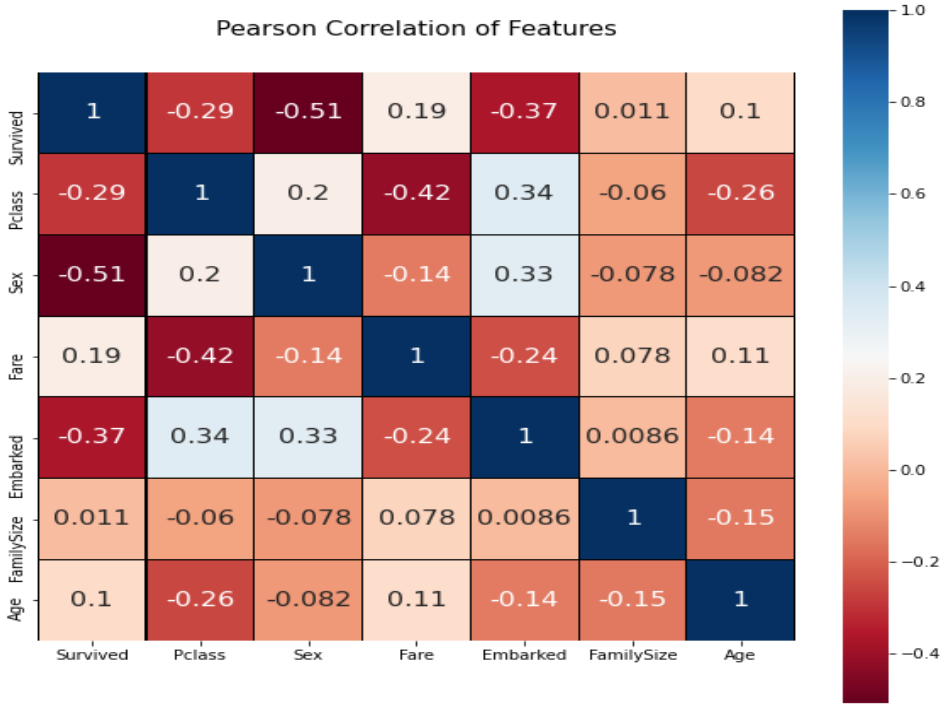

Figure 4. correlation of features

Fill missing values : I filled in the missing value considering correction.

Age : There was no feature directly related to Age. According to the P-class, people with a first-class rating are classified as older people. Or I was going to guess the age according to Family Size. But It was complicated and matched the missing value with the average value. For the speed of the model, the number was made according to the range of the age. It is defined smaller than 16 is 0, 16 ~ 32 is 1, 32 ~ 48 is 2, 48 ~64 is 3, larger than 64 is 4.

Fare : Fare is most relevant to Pclass in terms of the correction distribution. So I filled the price with the median of each Pclass. At first, I agonized between the Mean and the Median, but decide to put Median because it depends on the location of the floor and other services, so I thought it would be better to fill the most common and large values Median.

Embarked : Embarked, like Fare, fills the missing values according to Pclass because Correlations are most associated with Pclass. Earlier, Pclass changed to 0, 1 and 2 so I filled the missing value using Median.

Fill in the Missing value and then "df.isnull().Sum()" was used to verify that there were no missing values.

Modeling

I choose K-Nearest Neighbor(KNN), Random Forest(RF), Naïve Bayes(NB) and Logestic Regression(LR) for modeling. Especially, Random Forest(RF) is mainly used in classification and is

fast and has high accuracy. It also worked well on large datasets, so I use random forest. The Support Vector Machine(SVM) did not use, because it takes long time to learn.

Each accuracy was measured by 75.26 for KNN, 70.29 for random forest, 73.47 for naive bays, and 76.6 for LogisticRegression.

Among them, I will proceed with the evaluation using top three accurate model. Logistic Regression, K-nearest neighbor and Naïve Bayes.

Evaluation

Internal evaluation

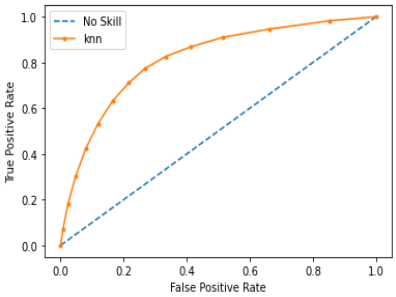

Figure 5. knn-roc

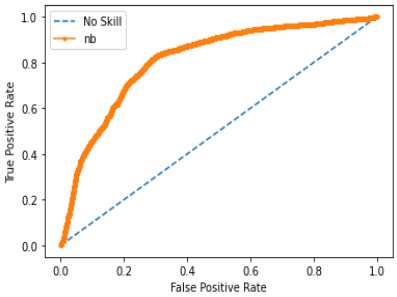

Figure 5. Nb-roc

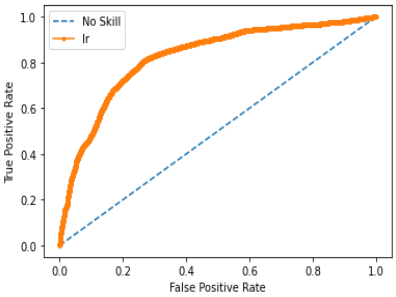

Figure 5. Lr-roc

Accuracy alone cannot accurately evaluate performance, so we use Roc curve that considers Sensitivity and Specificity together. I used Roc curve in my internal assessment. The values of auc, which obtained the under area of the roc-curve respectively, were 0.815(knn), 0.813(nb), and 0.825(lr), similar to the order of the accuracy. In other words, both logistic regression 's accruacy and rock curve sides showed better performance.

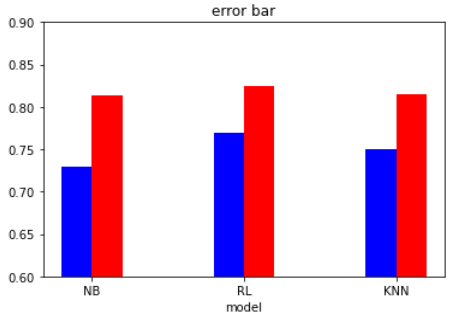

Figure 5. error bar

External evaluation

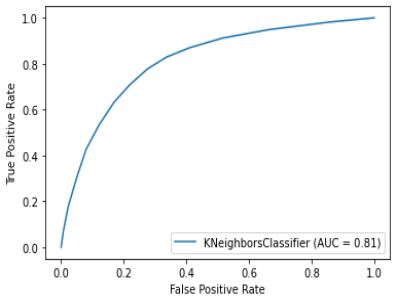

Figure 6. knn-roc

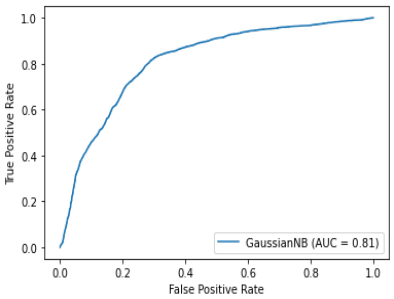

Figure 6. Nb-roc

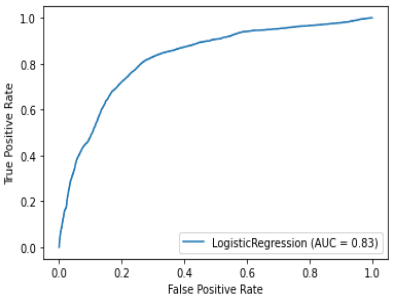

Figure 6. Lr-roc

Three models were also measured in external evaluation. As before, the highest score in logistic regression was achieved. Similarly, by 0.81(knn), 0.81(nb), 0.83(lr) the logistic regression showed the highest performance in external evaluation. Internal evaluation and External evaluation did not show much difference. So, overfitting and underfitting does not occur. It can be said that the model was well trained.

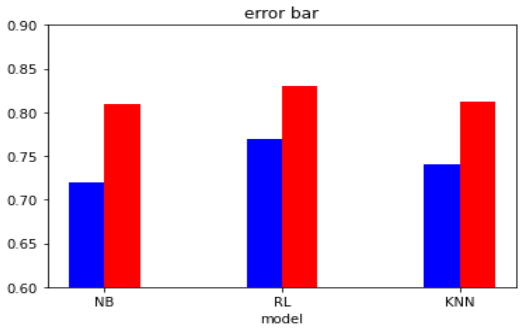

Figure 6. error bar

Conclusion

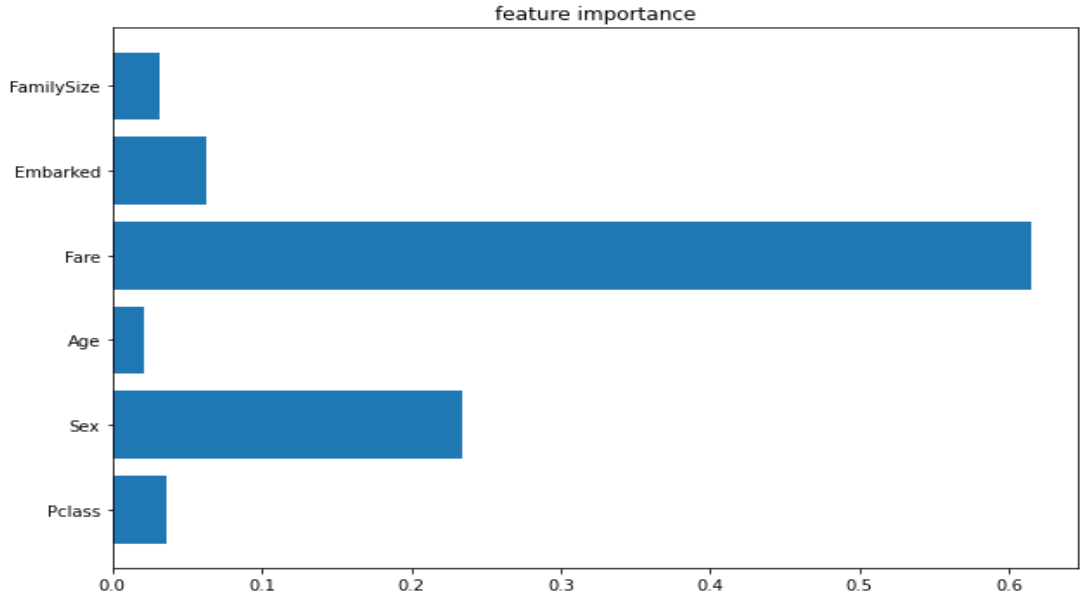

Figure 7. feature importance

A number of features were analyzed, and feature engineering was performed, such as filling the missing value and dropping unnecessary features by using correlation. I chose several models with high scores using several models(knn, logistic regression, random forest, SVM gaussianNB, and so on). Using the selected models, internal evaluation was carried out using the roc-curve. We also used the test set to draw the roc-curve and perform the external evaluation. As a result, logistic regression is the best model.

What was different from my expectation was that the random forest had a low score. This is because random forest works well on large datasets and has high accuracy. After considering the reason, I think if it was because underfitting occurred and there was a little noise in the prediction result. I also checked that the data set I received was not real data, Sometimes there was strange data. So I guess that's why the random forest score is low because there are wrong data in between.

As expected, Fare and Sex were important features of survival rates. A person with a high Fare would not be assigned to a third class downstairs with a high density of people. The more expensive Fare is, the higher floor and the less the density of people, the safer the room will be assigned. There

were also life jacket and lifeboats near the first class room. That's why Fare has important feature results. Sex is also an important feature. Just as the Titanic movie shows a scene in which women and children are put on a lifeboat first, there must have been a cultural aspect of saving women and children first in real situation.

The importance of Sex was high, while the importance of Age was lower than expected. This is because both must be of high importance according to the cultural aspect of rescuing the weak first. I guessed the reason is that children and the elderly are have the lack of the ability to escape. In addition to the features given, I looked at the Titanic's design and found that there were other factors related to survival rates, such as poor Titanic design.

Through this final term, I realized that feature engineering was important and need put a lot of effort and time, not just to import package and train models. I also realized that the accuracy is different each models we learned in theory. And each model's importance of features and result are different. Finally, there were some results that were slightly different from the hypotheses I set. So I felt that detailed analysis was important.

Reference

https://www.youtube.com/channel/UCxP77kNgVfiiG6CXZ5WMuAQ
https://www.kaggle.com/c/titanic/code?competitionId=3136&sortBy=voteCount
https://www.kaggle.com/startupsci/titanic-data-science-solutions
https://www.kaggle.com/daehungwak/guide-kor-dg