Group 7 report about Titanic incident analyzing

Abstract

This report was written to create a model capable of predicting whether a traveler survived or not. In the model, Gender, Age and Pclass will be included, and we will use two machine learning method which is K Nearest Neighbor and Logistic regression to predict survival chance. Two model will be compared to each other and to the base to find predicting accuracy. Lastly, we will try to deduce the result that we have.

1. Introduction

Titanic has always been a disaster when talking about voyage. The incident took place way back on April 15th, 1912, when the Titanic sank during its first voyage after striking an iceberg, resulting in the tragic loss of 1502 passengers and crew out of a total of 2224.This sensational tragedy horrified the international community and led towards action taking by formulating better safety regulations for ships. While it is still controversial about who took responsibility of the incident, we can agree that there are certain characteristics share between survivors. In this report, we try to find out what features will likely make an individual survive in the Titanic incident and perhaps in a disaster.

In this report, we intended to use two kind of classification machine learning method which are KNN and Logistic Regression.

II. Methodology

2.1. Source of data

We will use the provided data which come from Kaggle which is a excel file Train.csv (we also uploaded it on Github). The data was recorded during the incident, and it contains numorous information of 890 passagers on that ship.

2.2. Variable

2.2.1. Data dictionary

|  |  |  |
| --- | --- | --- |
| **Variable** | **Definition** | **Key** |
| survival | Survival | 0 = No, 1 = Yes |
| pclass | Ticket class | 1 = 1st, 2 = 2nd, 3 = 3rd |
| sex | Sex |  |
| Age | Age in years |  |
| sibsp | # of siblings / spouses aboard the Titanic |  |
| parch | # of parents / children aboard the Titanic |  |
| ticket | Ticket number |  |
| fare | Passenger fare |  |
| cabin | Cabin number |  |
| embarked | Port of Embarkation | C = Cherbourg, Q = Queenstown, S = Southampton |

\*Note:

**pclass**: A proxy for socio-economic status (SES)  
1st = Upper  
2nd = Middle  
3rd = Lower  
  
**age**: Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5  
  
**sibsp**: The dataset defines family relations in this way...  
Sibling = brother, sister, stepbrother, stepsister  
Spouse = husband, wife (mistresses and fiancés were ignored)  
  
**parch**: The dataset defines family relations in this way...  
Parent = mother, father  
Child = daughter, son, stepdaughter, stepson  
Some children travelled only with a nanny, therefore parch=0 for them.

2.2.2. Variable Selection

There are many variables provided, but in the model, we decide to use four independent variables which is Sex, Age, Pclass, Embarked. For dependent variable, we will use Survived for our model.

The reason we choose these variables is based on three filters.

First, we check the null value of each column and it turn out there are 177 rows of Age column, 687 rows of Cabin column and 2 rows of Embarkes columns is missing. While we can ‘fix’ Age and Embarked columns, we do not think that is possible to use a column with approximately 80% missing data like Cabin.

Second is the heatmap that we included. From the chart, we can see that Sibsp has low coefficient, so we did not include it in the model. While Fare has a high coefficient, it also has high coefficient with the Pclass so we will also not gonna use that.

*\*Note: The reason we choose Pclass instead of Fare is Fare is continuos variable, which is more suitable for the regression task.*

Third is we have conducted a test logistic regression model annd perform z test amd the test result suggest that we should not use Parch as it is not statistically significant (at z=0.05)

2.3. Preprocessing data

The data that we receive is not cleaned. There are numerous of null value in the Age column, beside the model can run on str data like what we have in Sex column.

For Age column we decide to fill the null value with mean of all the rows in column as there are around 177 missing rows, around 20% of the data, dropping it might lead to less accuracy of the model in general. While filling null value with mean might not improve the accuracy, more accurate age prediction will involve other machine learning model (like random forest), which might extend our presentation’s time too much.

For Embarked, there were only 2 missing data, so we decide to drop it. We also turn the type of data into numerical. S/C/Q is converted into 0/1/2

For Sex column, we will convert the male/female into 1/0 as model we use from sckitlearn only take numerical input.

2.4. Data Visualizing

2.4.1. Pie chart about Sex and Survived

This chart was made to compare the percentage of female and male in the survivor.

A pie chart with a number of people

Description automatically generated

It can be told from the chart that if someone is female, she will have chance to survive in the Titanic incident.

While the chart did not indicate that if you are a male, the percentage that you will live is **20.3%**, it highlights the significant difference in the percent of female to male in the survival data.

2.4.2. Bar Chart about PClass and Survived

A graph of a bar chart

Description automatically generated with medium confidence

This bar chart was made to compare the survival rate of each class in the Pclass with:

* 1 - First class (the highest class)
* 2 - Second class
* 3 - Third class.

It is clearly to see the higher the class ticket that we bought, the more survival rate that we will have.

2.4.3 Density plot

A graph of a number of people

Description automatically generated with medium confidence

The chart shows the age distribution of passengers who survived and those who did not survive in the Titanic case.

Firstly, 80 years old and 4 months old are the oldest and youngest passenger on the ship, with the average passenger's age is nearly 30. From the chart, we can see that the survival probabilities from the age 40 to 60 are similar for both groups. The peak on 'Survived = 1' shows the range is less than the 'Survived = 0' one, which means people who are in the age of 20 to 30 have less chance to survive than expected. We could see the curve of 'Survived=1' is moved towards younger ages than 'Survived=0', it means that younger passengers have higher chance to survive.

2.4.4. Line plots

A graph of a person and person

Description automatically generated with medium confidence

Mostly passenegers boarded from S, especially in Pclass = 3. Female passengers have better survival rate than male, especially in Pclass = 1 & 2. And the chart shows the survival rate of Pclass = 3 is the lowest chance to survive especially the embarked = S, which means it was about money they pay for. The survival in Embarked = Q only/mostly have passengers from Pclass=3, but men in this 'embarked' are unlukiest to survive.

2.5. Models Used

For this project, two models were selected:

1. Logistic regression model

2. K nearest neighbor

The reason while we use these two models in this report instead of linear regression is that we realize our task is classification, which will give the outcome is categorical data not regression which will give a continous outcome.

Before using this model, we used a pair plot to simply visualize the model.

As we can see from the chart, most of the variables that we choose has a normal distribution. While age looks a little bit screwed, we think it is still acceptable.

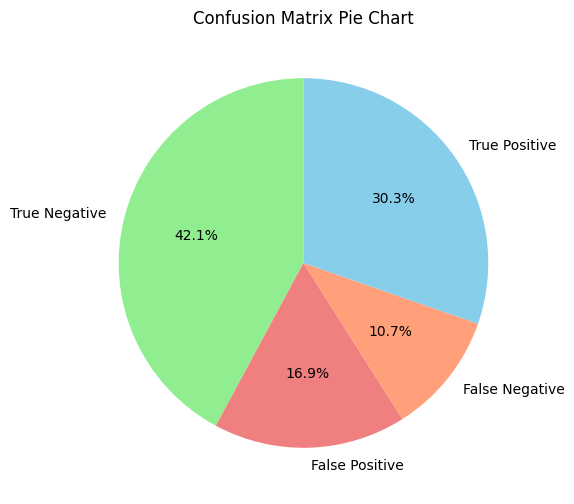
III. Result

3.1. Logistic Regression

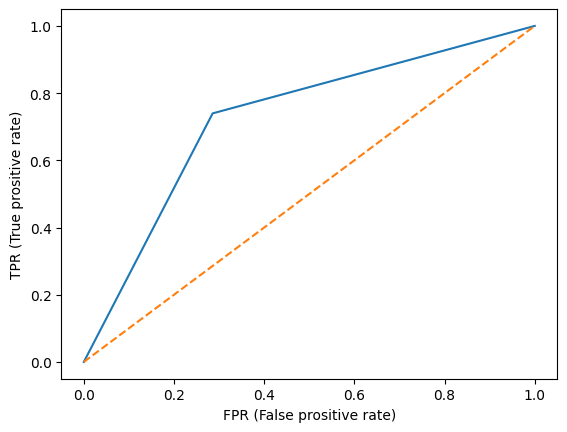
First, we want to discuss that in a real-life situation, to be predicted as alive but actually dead (False positive) or predicted as dead but actually alive (False negative) both have a serious impact but. In the first case, to be predicted as alive but actually dead might expand the cost for a rescuing mission more than it needed. While in the second case, the victim might not receive the rescue need, or the authority might not prepare enough resource for the rescuing misson. Personally, we think the false negative in this case is worst than the false positive.

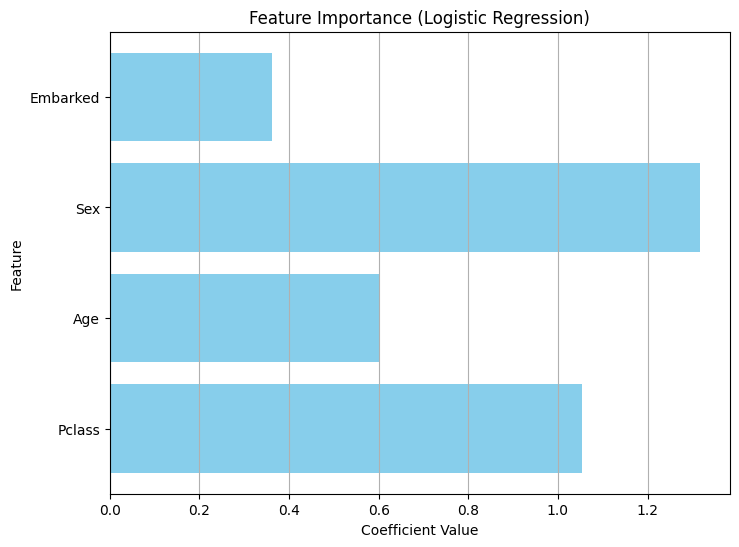
To minimize the false negative, we increased the threshold to 0.7. To come up with this number, we tried some other possibilities higher or lower than 0.7 and 0.7 is the number which give a low false negative while keeping the false positive acceptable.

After choosing a suitable threshold, we create the model, confusion matrix based on the model and visualizing it using a pie chart. As we can see, most of the chart is true positive and true negative, which give us a high accuracy (around 72.4%). False negative occupied 16.9% of the chart which is higher than False positive which is 10.7%. The f1 score, precision, specificity, recall, respectively is 0.69, 0.64, 0.71 and 0.74. We will discuss more about this after caculate the number of KNN model.



We also include the roc curve and caculate the auc which give us a result around 72.7% which we think is enough to predict the right value while not making too many minimal errors.



We create a chart which showing the importance of all the variables that we choose for our model.

\*Note: All the variable is used in their absolute value to make it better for analysing.

As we can see, Sex and Pclass has a big impact on the model and Age and Embarked seem to have less impact on the model. Futher discussion will be included in part IV of the report.

3.2. KNN

For KNN model, we run a few tools to optimize the k values and the k value that we choose is 23.

After testing the model, it gives us a confusion matrix that we will visualize by a heat map.

A graph showing the positive results

Description automatically generated with medium confidence

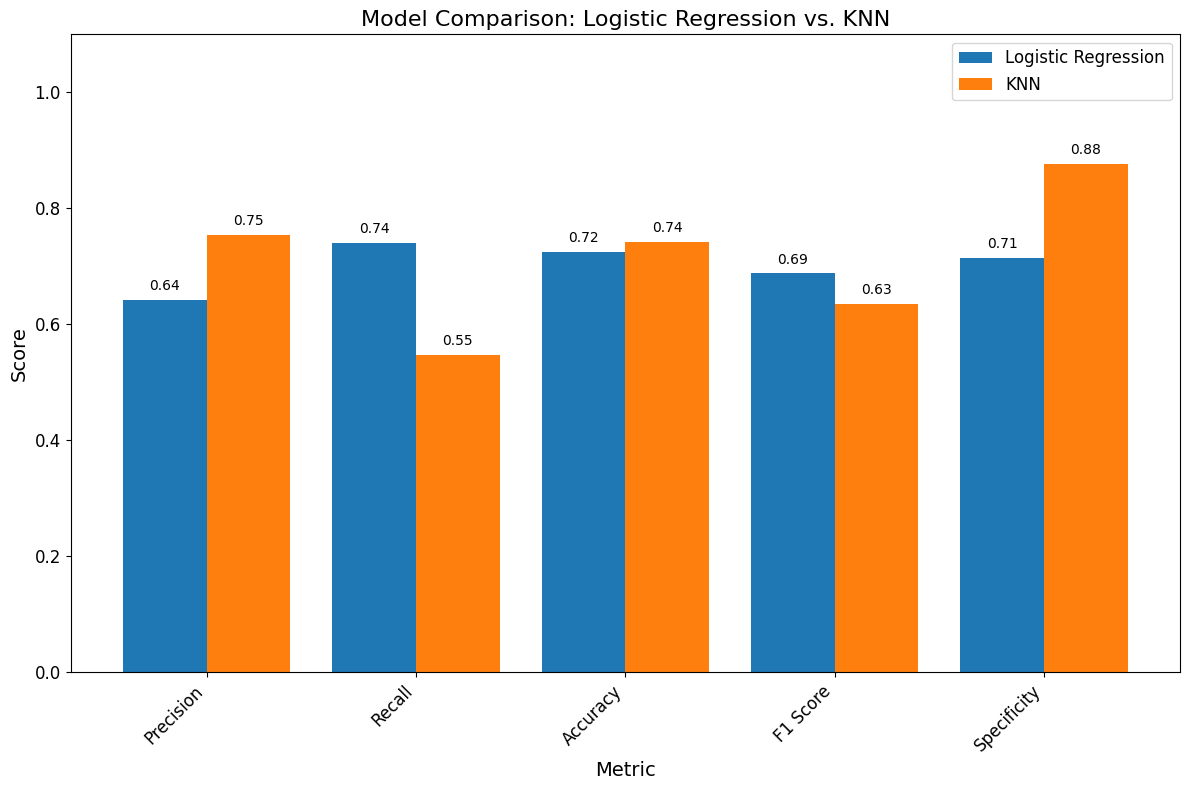
Different color represents the density of the value for it part of the matrix. And similarly, to the logistic regression, most of the predicted value is right. But comparing to the logistic model, KNN has a slightly better accuracy (74%, comparing to 72% of logistic regression model) and it also predicted non-survived (true negative value) better.

We also added a ROC curve, which indicating that the model performance quite well.

A graph with a line and a point

Description automatically generated with medium confidence

For better comparing, we use a barchart to visualize the metrics of two models.



It can tell from the model that KNN is better in recognizing the negative value with specificity higher than the outcome of the logistic regression. It also predicted the positive value with more accuracy, but logistic model makes less error than KNN model, according to the high recall. The result above can be explained as the logistic regression model is created to have at least false negative value. When balancing both precision and recall we got f1 score which the value of logistic regression model is little bit higher indicate that logistic model has a better performs in predict the positive values. But in general, knn seem to be a little bit better than the logistic regression as it has higher accuracy.

IV. Conclusion

As the feature importance analysis of the logistic model indicated that Sex and Pclass have a bigger impact on the model while Age and Embarked. This can be explained as women is often be prioritised in an incident and high social-economic status have more chance to access to the rescue, so it has a strong relationship with the outcome. For Age and Embarked values, it can be explained by the Facetgrid that we make as we can see that the distribution for survived and non-survived is quite similar, so the pattern between the age and the survived rate is not strong. For Embarked, from qualitative perspective, we hardly can see the relationship between the port that start traveling and the survival rate of ourselves in the first place, so deeper research needs to be taken to understand this relationship between it.

From our perspective, to increase the accuracy of both models, the data need to have less null value, as we indicate in the 2.3, there are 177 rows is missing, which is around 20% of the given data, itdecrease the accuracy of both model and might be the reason why the coefficient of Age is low. Beside, the provided data also have a limited variables and some of it can not be used for analysis (like name, ticket number, cabin) make it might make the model underfitting.