Could you Survive? Shipwreck survival analysis

Name: Yukang Xu

Ⅰ. Introduction

The sinking of the Titanic is one of the most infamous shipwrecks in history.

On April 15, 1912, during her maiden voyage, the widely considered “unsinkable” RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren’t enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew.

While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others.

In this project, I will build a predictive model that answers the question: “what sorts of people were more likely to survive?” using passenger data (ie name, age, gender, socio-economic class, etc).

Ⅱ. Data Analysis

1. Identify our data

This is the meet and greet step. Get to know your data by first name and learn a little bit about it. What does it look like (datatype and values), what makes it tick (independent/feature variables(s)), what's its goals in life (dependent/target variable (s)).

Categorical: Name, Sex, Ticket, Cabin, Embarked

Numeric: ID, Class, Age, Sib, Parch, Fare, survival

Features with null value: age (train+ test), cabin(test+ train), embarked(train), fare(test)

Based on that, we analyzed every variable then determined how to handle them.

Survived variable	binary variable: 1 for survide while 0 for not survived
Passenger Id and Ticket number Variable	Random variable: choose to drop from our analysis
Pclass variable	Ordinal variable
Name	Nominal variable: they can be used to find the title of people
Sex and Embarked variables	Nominal Variable: They can be converted to dummy variables
Age and Fare Variables	Continuous Variables
Sibsp and Parch Variables	Discreate Quantitative Variables: they can be used to create new family size variables
Cabin Variables	Nominal Variables: They will be dropped from analysis because of too much missing values

From the analysis above, we will drop Cabin, Passenger, Ticket number variables.

Convert Class, Sex, Embarked, Sibsp, Parch variables to dummy variables and apply Age and fare variables in analysis.

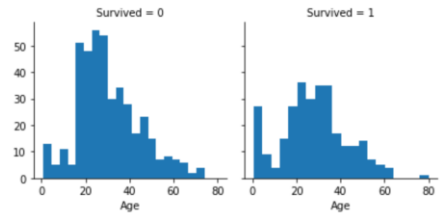
2.Explore the dataset

we will explore our data with descriptive and graphical statistics to describe and summarize our variables. In this stage, you will find yourself classifying features and determining their correlation with the target variable and each other.

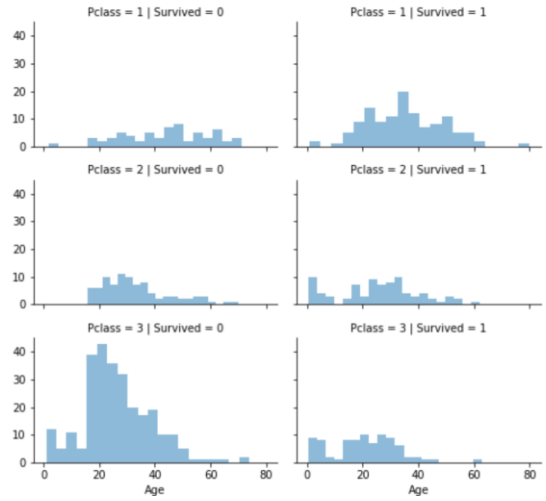
a. (sex)Woman (Female) are more likely to survive

	Sex	Survived
0	female	0.742038
1	male	0.188908

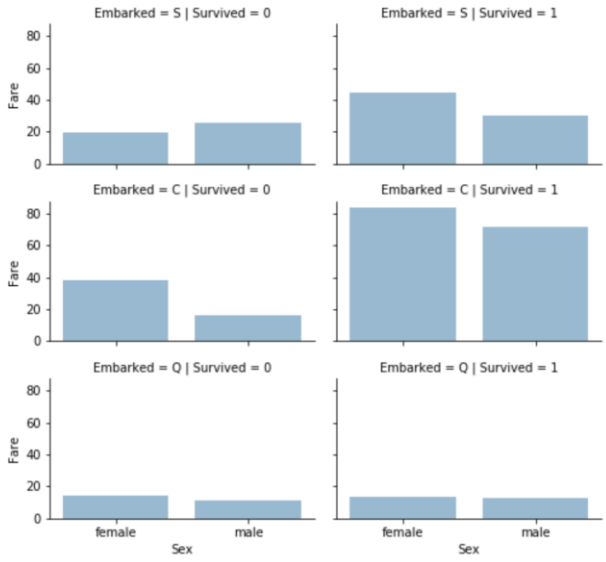
b.(age) Large number of 15-25 year old did not survive while most passengers are in 15-35 age range.



c. (class)Pclass=3 had most passengers, however most did not survive. Most passengers in Pclass=1 survived.



d. Higher fare paying passengers had better survival. And Port of embarkation correlates with survival rates.



3. Clean Data (4' C)

a. Correcting: detect the outliers. Reviewing the data, there does not appear to be any aberrant or non-acceptable data inputs. In addition, we see we may have potential outliers in age and fare. However, since they are reasonable values, we will wait until after we complete our exploratory analysis

to determine if we should include or exclude from the dataset. It should be noted, that if they were unreasonable values, for example age = 800 instead of 80, then it's probably a safe decision to fix now. However, we want to use caution when we modify data from its original value, because it may be necessary to create an accurate model.

b. Completing: fill in the missing value. There are null values or missing data in the age, cabin, and embarked field. Missing values can be bad, because some algorithms don't know how-to handle null values and will fail. While others, like decision trees, can handle null values. Thus, it's important to fix before we start modeling, because we will compare and contrast several models. There are two common methods, either delete the record or populate the missing value using a reasonable input. It is not recommended to delete the record, especially a large percentage of records, unless it truly represents an incomplete record. Instead, it's best to impute missing values. A basic methodology for qualitative data is impute using mode. A basic methodology for quantitative data is impute using mean, median, or mean + randomized standard deviation. An intermediate methodology is to use the basic methodology based on specific criteria; like the average age by class or embark port by fare and SES. There are more complex methodologies, however before deploying, it should be compared to the base model to determine if complexity truly adds value. For this dataset, age will be imputed with the median, the cabin attribute will be dropped, and embark will be imputed with mode. Subsequent model iterations may modify this decision to determine if it improves the model's accuracy.

c. Creating: create a new variable: title to see if it plays a role in survival. Feature engineering is when we use existing features to create new features to determine if they provide new signals to predict our outcome. For this dataset, we will create a title feature to determine if it played a role in survival.

d. Converting: convert ordinal, nominal data to dummy variables. Last, but certainly not least, we'll deal with formatting. There are no date or currency formats, but datatype formats. Our categorical data imported as objects, which makes it difficult for mathematical calculations. For this dataset, we will convert object datatypes to categorical dummy variables.

4. Linear SVC and SVM

Now we are ready to train a model and predict the required solution. There are 60+ predictive modelling algorithms to choose from. We must understand the type of problem and solution requirement to narrow down to a select few models which we can evaluate. Our problem is a classification and regression problem. We want to identify relationship between output (Survived or not) with other variables or features (Gender, Age, Port...). We are also perfoming a category of machine learning which is called supervised learning as we are training our model with a given dataset. With these two criteria - Supervised Learning plus Classification and Regression, we can narrow down our choice of models to a few. These include: From the score, we can draw a conclusion that SVM perform better in our case.

	Model	Score
0	Support Vector Machines	90.57
1	Linear SVC	80.24

Support Vector Machines which are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training samples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new test samples to one category or the other, making it a non-probabilistic binary linear classifier.

5. Model Evaluation

We can now rank our evaluation of all the models to choose the best one for our problem. While both Decision Tree and Random Forest score the same, we choose to use Random Forest as they correct for decision trees' habit of overfitting to their training set.

	Model	Score
3	Random Forest	86.76
8	Decision Tree	86.76
1	KNN	84.74
0	Support Vector Machines	83.84
2	Logistic Regression	80.36
7	Linear SVC	79.12
6	Stochastic Gradient Decent	78.56
5	Perceptron	78.00
4	Naive Bayes	72.28

Ⅲ. Conclusion

Our submission to the competition site Kaggle results in scoring 3,883 of 6,082 competition entries. This result is indicative while the competition is running. This result only accounts for part of the submission dataset. Not bad for our first attempt. Any suggestions to improve our score are most welcome.