

Titanic: Machine Learning from Disaster

1. Problem Statement

The goal of the titanic competition is to use machine learning algorithms to create a model that could predict which passengers survived the Titanic shipwreck. Basically, the task is a binary classification problem where given a set of data on passengers on aboard, such as sex, age, and class and the goal of the model is to predict whether a given passenger would have survived in this disaster or not.

2. Data Exploration

a. Load the dataset

As the provided data are in .csv format, we use Using Pandas' **dataframe** to read the train and test files. We peek at the first 10 rows of the training data to get a brief overview of the data (Figure 1).

Out[1]:

	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
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	NaN	C

Figure 1

b. Explore the Dataset

In this section, we explore and analysis the dataset using plots in order to have some insights of the data. We plot the distribution of some features in both training set (Figure 2) and test set (Figure 3).

The training set contains 891 rows and 12 columns. Possible features for each data point are passengers' ID (PassengerId), ticket class (pclass), name, sex, age, number of siblings / spouses aboard (SibSp), number of parents / children aboard (Parch), ticket number (ticket), cabin, port of embarkation (Embarked, C = Cherbourg, Q = Queenstown, S = Southampton). Labels are which passengers survived the Titanic tragedy (Survived, 0 = No, 1 = Yes).

The test set contains 418 rows and 11 columns. Possible features for each data point are passengers' ID (PassengerId), ticket class (pclass), name, sex, age, number of siblings / spouses aboard (SibSp), number of parents / children aboard (Parch), ticket number (ticket), cabin, port of embarkation (Embarked, C =

Cherbourg, Q = Queenstown, S = Southampton). Label that need to be predicted is which passengers survived the Titanic tragedy (Survived, 0 = No, 1 = Yes).

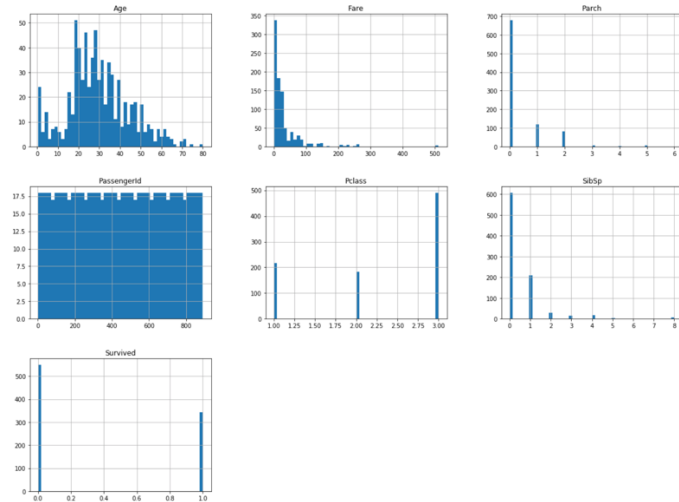


Figure 2

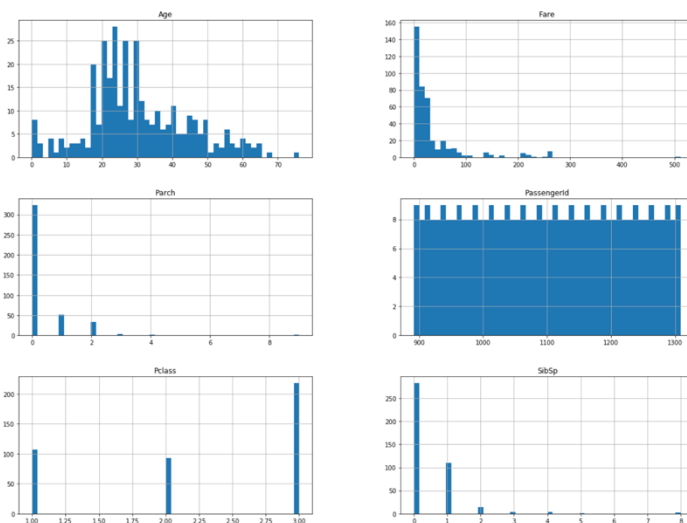


Figure 3

c. Observations

From the above information, we noticed that:

1. Features are in different types:
 - (1). Categorical features: Sex, Embarked, Parch, Pclass and SibSp
 - (2). Continuous features: Age and Fare
2. Some of the features are missing,
 - (1). Training set: Age, Cabin and Embarked
 - (2). Test set: Age, Fare, Cabin and Embarked

d. Missing Data

Check if there is any missing data in specific features. We find that there are three features that are missing data points in training set (Table 1): 'Cabin', 'Age',

and 'Embarked', and test set (Table 2): 'Cabin', 'Age', and 'Fare'. We will describe how to fill in the missing data in the chapter 4.

	Total	Percent
Cabin	687	0.771044
Age	177	0.198653
Embarked	2	0.002245
Fare	0	0.000000
Ticket	0	0.000000
Parch	0	0.000000
SibSp	0	0.000000
Sex	0	0.000000
Name	0	0.000000
Pclass	0	0.000000
Survived	0	0.000000
PassengerId	0	0.000000

Table 1

	Total	Percent
Cabin	327	0.782297
Age	86	0.205742
Fare	1	0.002392
Embarked	0	0.000000
Ticket	0	0.000000
Parch	0	0.000000
SibSp	0	0.000000
Sex	0	0.000000
Name	0	0.000000
Pclass	0	0.000000
PassengerId	0	0.000000

Table 2

3. Feature Correlations

In this chapter, we first use bar charts to show the correlation between survival rate and categorical features, such as Pclass, Sex, SibSp (number of siblings and spouse), Parch (number of parents and children), and Embarked. Then, we use box plots to show the correlation between survival rate and numerical features, such as Age and Fare.

a. Pclass

The bar chart (Figure 4) shows that passengers in 1st class more likely survived than other classes, and passengers in the 3rd class more likely dead than other classes.

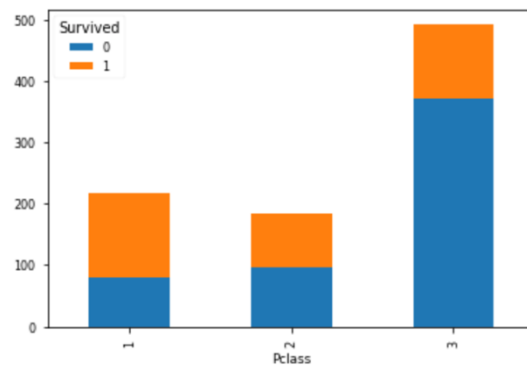


Figure 4

b. Sex

The bar chart (Figure 5) shows that female passengers more likely survived than male.

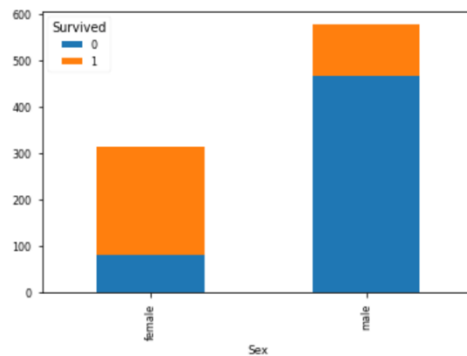


Figure 5

c. SibSp

The bar charts (Figure 6) show that a passenger boarded with more than 2 siblings or spouse more likely survived, and without siblings or spouse more likely dead.

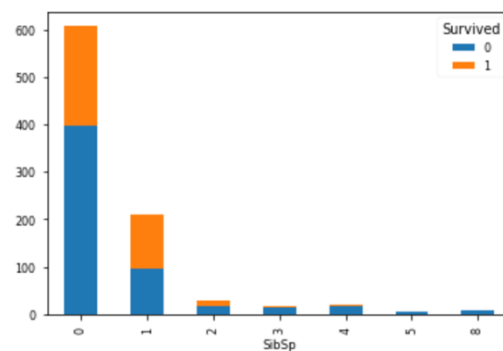


Figure 6

d. Parch

The bar chart (Figure 7) shows that a passenger boarded with more than 2 parents or children more likely survived, and passengers boarded alone more likely dead.

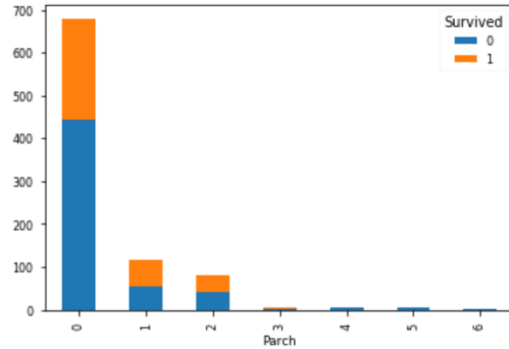


Figure 7

e. Embarked

The bar chart (Figure 8) confirms a passenger boarded from C slightly more likely survived.

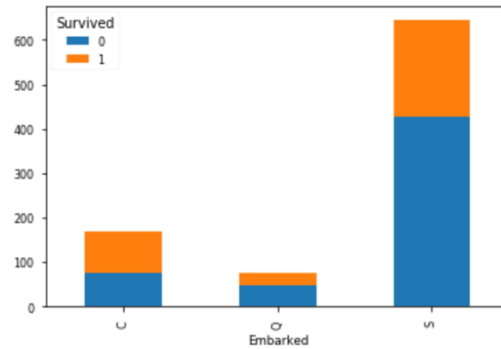


Figure 8

f. Age

The box plot (Figure 9) shows that survived passengers are slightly younger than those who are dead.

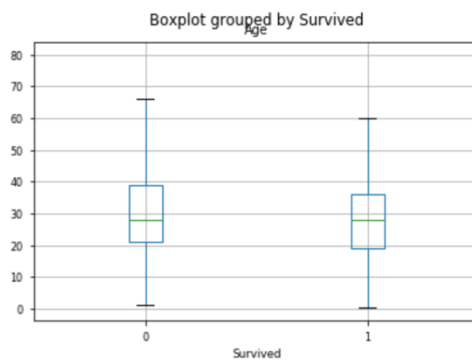


Figure 9

g. Fare

The box plots (Figure 10) show that the fares of survived passengers have higher value than that of dead ones.

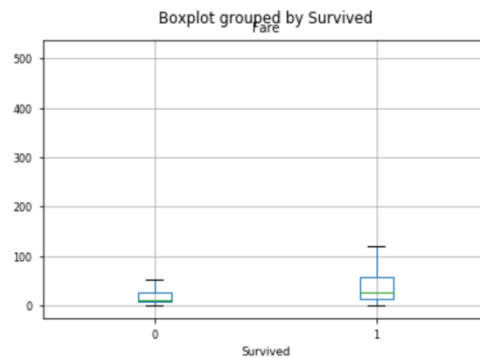


Figure 10

4. Feature Engineering

a. Name

As passengers' names are quite various, which are not significant to predictions; however, from the analysis of the dataset above, we have found that 'Sex' has important correlation with the survival rate. Hence, we keep the title extracted from the 'Name' column and create a new column called 'Title' to save them.

Figure 11 shows the distribution of different titles.

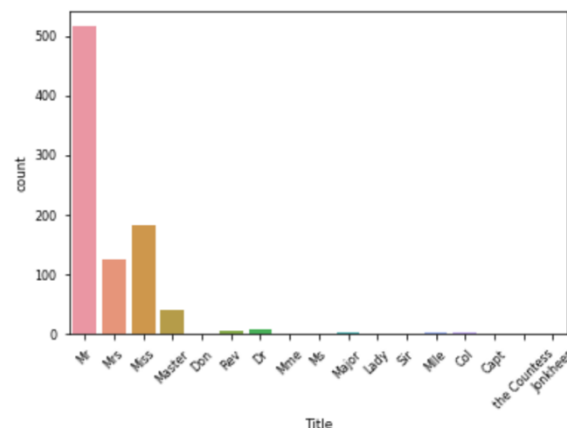


Figure 11

We have seen that Mr., Mrs., Miss and Master are the most common titles, others only have a very low frequency. So we group titles into five categories: Mr., Mrs., Miss, Master and Rare. From the plot below (Figure 12), we could see that 'Mr.' who mainly characterize men has a low survival rate while 'Mrs.' & 'Miss' who mainly characterize women has a high survival rate. At last, we encode title categories into numerical encodings.

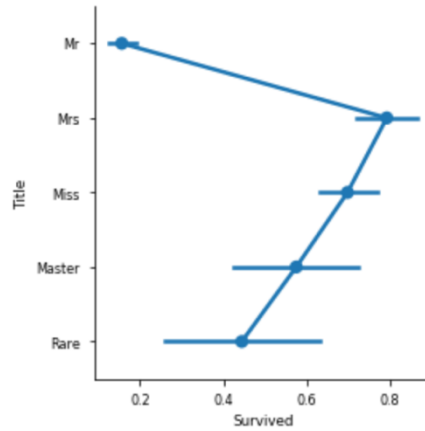


Figure 12

b. Sex

We encode 'Sex' into numerical features (Male:0, Female:1). The plot (Figure 13) shows that female has a higher survival rate than male.

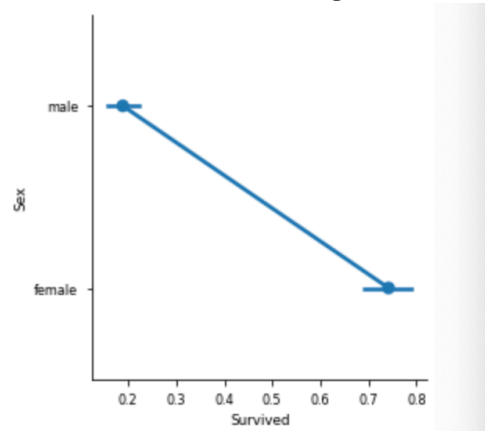


Figure 13

c. Age

We first deal with the missing data point by using the SEX's median age to fill out the missing age, and then converting numerical age to categorical attributes. We set the age bands by cutting the age range into 5 pieces. The plots (Figure 14) shows that younger people has slightly higher survival rate.

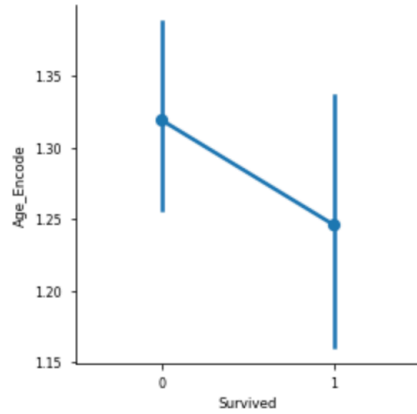


Figure 14

d. Fare

We first deal with the missing data point by using the median to fill out the missing fare, then converting numerical fare to categorical attributes. We set the fare bands by cutting the fare range into 5 pieces. The plot (Figure 15) below shows that people who paid higher fare rate has the higher survival rate.

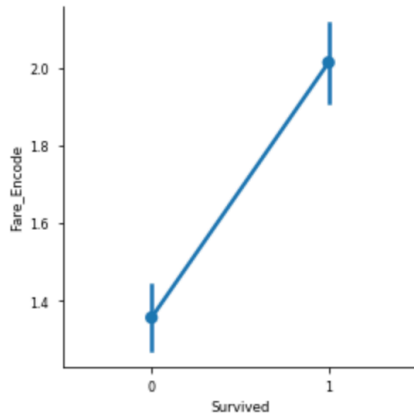


Figure 15

e. Cabin

There are too many missing values in cabin column, we assume that the survival rate might be different for people who has cabin number and people who has not. The plot (Figure 16) shows that people who has the cabin number has the higher survival rate.

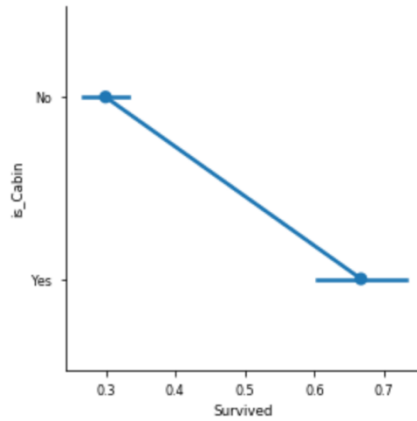


Figure 16

f. Embarked

From the analysis above, we notice that there is only one missing value in Embarked column, and we fill the missing 'Embarked' with the mode. Then we encode 'Embarked' into numerical features S:0, C:1, Q:2). The plot (Figure 17) shows that people embarked on port C has the higher survival rate.

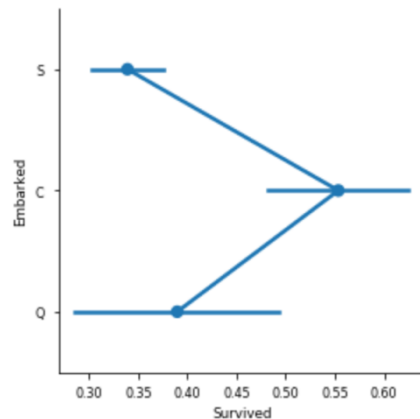


Figure 17

g. Family On board

The 'SibsSP' feature refers to the number of siblings or spouses that a passenger had board with, and the 'Parch' refers to the number of parents or children someone had on the ship. These two features could be combine as a new feature 'Family_Together', where we want to know if a passenger had someone from his/her family onboard. The plot (Figure 18) shows that people who board alone has the higher survival rate.

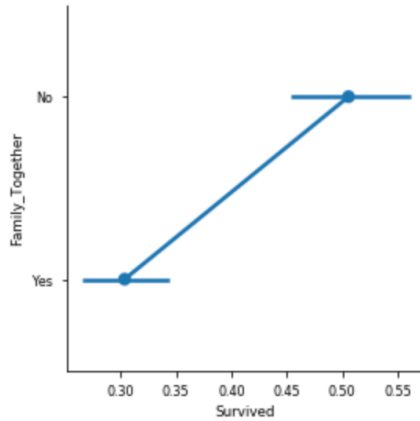


Figure 18

h. Overall Feature Correlations

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare	Family_Size	Family_Tog
PassengerId	1.000000	-0.005007	-0.035144	0.036847	-0.057527	-0.001652	0.012658	-0.040143	0.05
Survived	-0.005007	1.000000	-0.338481	-0.077221	-0.035322	0.081629	0.257307	0.016639	-0.20
Pclass	-0.035144	-0.338481	1.000000	-0.369226	0.083081	0.018443	-0.549500	0.065997	0.13
Age	0.036847	-0.077221	-0.369226	1.000000	-0.308247	-0.189119	0.096067	-0.301914	0.19
SibSp	-0.057527	-0.035322	0.083081	-0.308247	1.000000	0.414838	0.159651	0.890712	-0.58
Parch	-0.001652	0.081629	0.018443	-0.189119	0.414838	1.000000	0.216225	0.783111	-0.58
Fare	0.012658	0.257307	-0.549500	0.096067	0.159651	0.216225	1.000000	0.217138	-0.27
Family_Size	-0.040143	0.016639	0.065997	-0.301914	0.890712	0.783111	0.217138	1.000000	-0.69
Family_Together	0.057462	-0.203367	0.135207	0.198270	-0.584471	-0.583398	-0.271832	-0.690922	1.00

Figure 19

5. Basic Modeling

In this chapter we apply different machine learning algorithms to predict whether a passenger will survive based selected features. We use 861 training points with 8 features each. As some models, such as perceptron, k-nearest neighbors and support vector machine, are sensitive to the scaling of the data, we apply standard-scaling to the dataset. In order to avoid over fitting, we perform stratified cross validation (10 folds) to evaluate our models. We do grid search for some algorithms, such as perceptron, k-nearest neighbors. We use accuracy to evaluate the model performance.

a. Naïve Bayes Classifier

Results: training accuracy 75.66, scaled training accuracy 75.66

b. Perceptron

We do grid search on number of hidden layers and learning rate, and finally set number of hidden layers:32, learning rate:0.001.

Results: training accuracy 75.66, scaled training accuracy 79.58

c. Decision Tree

Results: training accuracy 80.59, scaled training accuracy 80.7

d. Support Vector Machine (SVM)

Results: training accuracy 81.49, scaled training accuracy 81.82

e. K-Nearest Neighbors


We do grid search on number of neighbors

Results: n_neighbors:10, training accuracy 81.36; n_neighbors:6, scaled training accuracy 82.04

6. Testing

Since many algorithms performs better in scaled data, we use scaled data in testing.

According to the results in training set, we select k-nearest neighbors as our model. The final accuracy of my model is 0.79665. Screenshot of submission to Kaggle shows below.

867	WesternNo.1	Palaniraj Rajagopal		0.79665	1	~10s
<p>Your First Entry ↑</p> <p>Welcome to the leaderboard!</p> <p>Your score represents your submission's accuracy. For example, a score of 0.7 in this competition indicates you predicted Titanic survival correctly for 70% of people.</p> <p>What next? You've got a few options:</p> <ul style="list-style-type: none">🧠 Learn skills that can improve your score in our Intro to Machine Learning course by Dan Becker.🔍 Check out the discussion forum to find lots of tutorials and insights from other competitors.🏆 Find a new challenge by entering one of our open, active competitions or searching our public datasets.						