

# APRIL TABULAR PLAYGROUND SERIES – YOUR BASELINE MODEL

Jinzhu Liu<sup>1</sup>

<sup>1</sup> Nanjing University of Science and Technology, China

## Introduction

Predict Whether or not A Passenger Survived the Inking of the Synthanic.

**Evaluation Goal** We task is to predict whether or not a passenger survived the sinking of the Synthanic (a synthetic, much larger dataset based on the actual Titanic dataset). For each row in the test set, you must predict a 0 or 1 value for the target. We score is the percentage of passengers you correctly predict. This is known as accuracy.

**Outlier Detection** aims to identify all possible outliers in the dataset, without explaining why or how they are different.

In this paper, It is important to aim to start somewhere and identify it as the first criterion to compare with existing progress. This helps you create a *baseline model* and get a baseline score.

## Variable Notes

- *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.
- *Loading the Data*,

	Passengerid	Survived	Pclass	Name	Sex	Age	SibSp	Parch
0	0	1	1	Oconnor, Frankie	male	NaN	2	0
1	1	0	3	Bryan, Drew	male	NaN	0	0
2	2	0	3	Owens, Kenneth	male	0.33	1	2
3	3	0	3	Kramer, James	male	19.00	0	0
4	4	1	3	Bond, Michael	male	25.00	0	0

## Data analysis and processing

Inquire NaN Values.

- list out the columns holding Numerical Values – 1.Age 2.Fare
- The remaining columns do not hold numerical values. Let’s explore the distribution of the numerical values a bit before we replace their NaN values

**Fare Chart information** At first, we planned to distinguish classes according to fare, Seems like the ticket to the titanic did not have any fixed price for any class in particular. so it is intended to scale the data and then use the average to impute.

**Age Chart information** It can be seen from the figure that the missing value can be filled by the median.

Count	99866.000000	Count	96708.000000
mean	43.92933	mean	38.355472
std	69.58882	std	18.313556
min	0.68000	min	0.080000
25%	10.04000	25%	25.000000
50%	24.46000	50%	39.000000
75%	33.50000	75%	53.000000
Max	744.66000	Max	87.000000

## Data analysis and processing

Second, We will use *KNN* to perform missing interpolation for Embarked.

PassengerId	0
Survived	0
Pclass	0
Name	0
Sex	0
Age	0
SibSp	0
Parch	0
Ticket	4623
Fare	0
Cabin	67866
Embarked	0

**Imputing Values** In this step, from this we can see that variables SibSp, Parch , PassengerId have a very small correlation coefficient as compared to others. We will be dropping these variables.

## Experiment

**Titanic Dataset** contains 100000 tourists and corresponding 12 attributes.

- First, we set up several classifier models to make a prediction. Then we use the training set to fit the model I built. After fitting, I use the fitted model to predict the remaining data in the training set and calculate its accuracy, weight, etc. Then fuse multiple groups of models, stack the fused model with the logistic regression model, and then fit the training set to get the prediction score. Use this model to predict our test set and see the prediction results of our test set.

we try Stacking Classifier and Logistic Regression on the test data.

	precision	recall	f1-score	support
0	0.79	0.80	0.79	11336
1	0.73	0.72	0.72	8664
accuracy			0.76	20000
macro avg	0.76	0.76	0.76	20000
weight avg	0.76	0.76	0.76	20000

First, lets see if there are missing values in the test, and fill in the missing values in the test in the same way as train. Finally, the well-fitting model is used to make predictions.

## Conclusion

**Prediction effect** Finally, the highest accuracy in LogisticRegression Model is 0.76.

**Project experience** A relatively basic Kaggle project was selected, the purpose is to be familiar with the Kaggle project, deeply analyze and understand each line of the project process, this project has done more processing on the step of data feature processing, and learned a lot from it.

**Direction to be improved** The work can also be further refined to improve the accuracy of prediction, for example, in the process of processing the age column, the age can be segmented according to the size of the age, and it is felt that the size of the age has a certain relationship with the size of the final survival rate.