TITANIC: MACHINE LEARNING FROM DISASTER

Progress Report

Yusi Liu & Yunhao Wu

Table of Contents

Abstract

Keywords: [Tap here to add keywords.]

Introduction

Learning Problem

Learning problem is selected from Kaggle[1].

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships.

One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.

In this challenge, we ask you to complete the analysis of what sorts of people were likely to survive. In particular, we ask you to apply the tools of machine learning to predict which passengers survived the tragedy.

Data Description

This dataset has 11 attributes and 1 target, which includes PassengerId, Name, Sex, Age, Sibsp, Parch, Fare, Cabin, Ticket, Embarked and Survived. And the target for classification is Survived.

Detailed description as follow:

Index	Variable	Definition	Data type	Key
Target	survival	Survival	int	0 = No, 1 = Yes
1	PassengerId		int	
2	pclass	Ticket class	int	1 = 1st,2 = 2nd, 3 = 3rd
3	Name		str	
4	sex	Sex	str	
5	Age	Age in years	float	
6	sibsp	# of siblings / spouses aboard the Titanic	int	
7	parch	# of parents / children aboard the Titanic	int	
8	Fare	Passenger fare	float	
9	cabin	Cabin number	str	
10	embarked	Port of Embarkation	str	C = Cherbourg, Q = Queenstown, S = Southampton
11	Ticket	Ticket number	str	

Table 2.1 Overview of the data

The data has been split into two groups: Train set (train.csv) & Test set (test.csv).

The training dataset has 891 entries, 12 columns, among which there are some data missing in attribute Age (714 entries, 177 missing), Cabin (204 entries, 687 missing) and Embarked (889 entries, 2 missing).

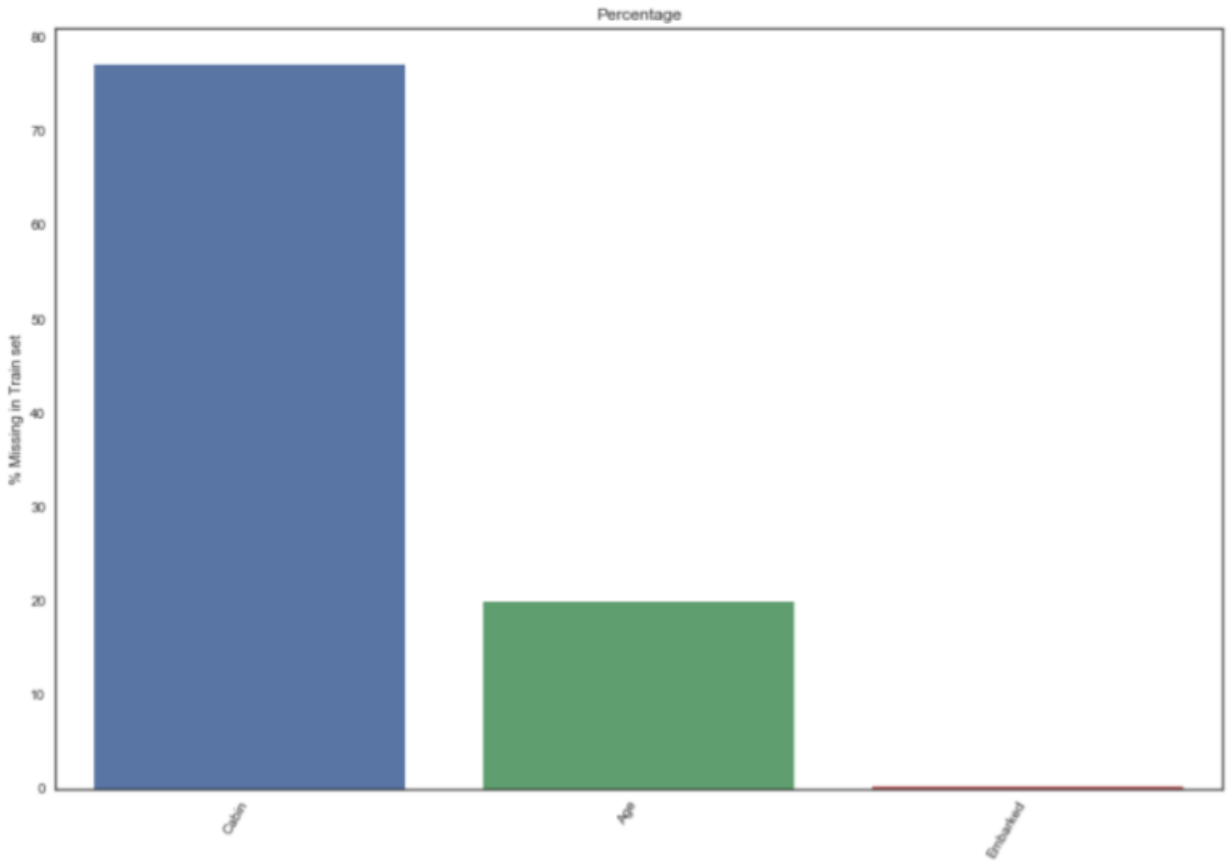

Figure 2.1 Missing percentage in train set

The test set has 418 entries, 12 columns, also containing missing data in attribute Age (332 entries, 86 missing), Fare (417 entries, 1 missing) and Cabin (91 entries, 327 missing).

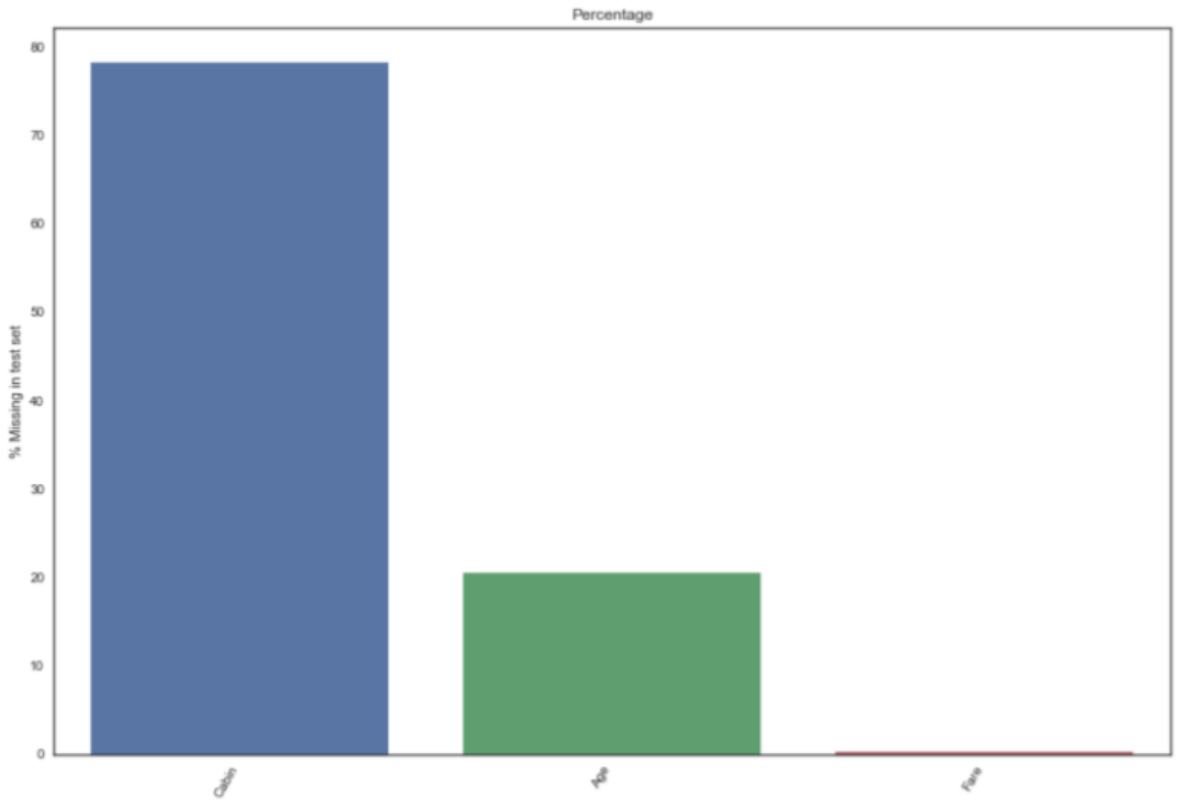

Figure 2.2 Missing percentage in test set

Methodology

Related Work

Many works have been involved before and many models has been adopted. Daniel Manuelpillai used K-Neighbors-Classifier and achieved 65% accuracy on train set [2]. Sagra Jain adoptee the sequential model from keras and achieved a 70% accuracy [3]. A single random forest methot got a result of over 77% [4]. But these single model methods are clearly not the best choices.

Typically, people are using boosting/bagging algorithm to ensemble several models in order to get a higher accuracy. Some favorable choices for ensemble model are GDBT, random forest and SVW [5]. We are also trying to adopt an ensemble model to improve the accuracy of the prediction.

Methodology

We learned following machine learning methods.

Logistic Regression

SVM

K-neighbors

Decision Tree

Random Forest

Gradient Boosting Decision Tree

XGB-GBDT

Experimental Results

Data Overview

Train Set

We describe training set as follows:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

From this table, we can see that the PassengerID is an index, which has little connection to the final survival result. The average age is younger than what we expected. Most passengers

are young adults. And in this accident, most young adults sacrificed their young life to save women and children.

There are some data missing in attribute Age (714 entries, 177 missing), Cabin (204 entries, 687 missing) and Embarked (889 entries, 2 missing).

To make the other information clearer, we draw the Correlation matrix.

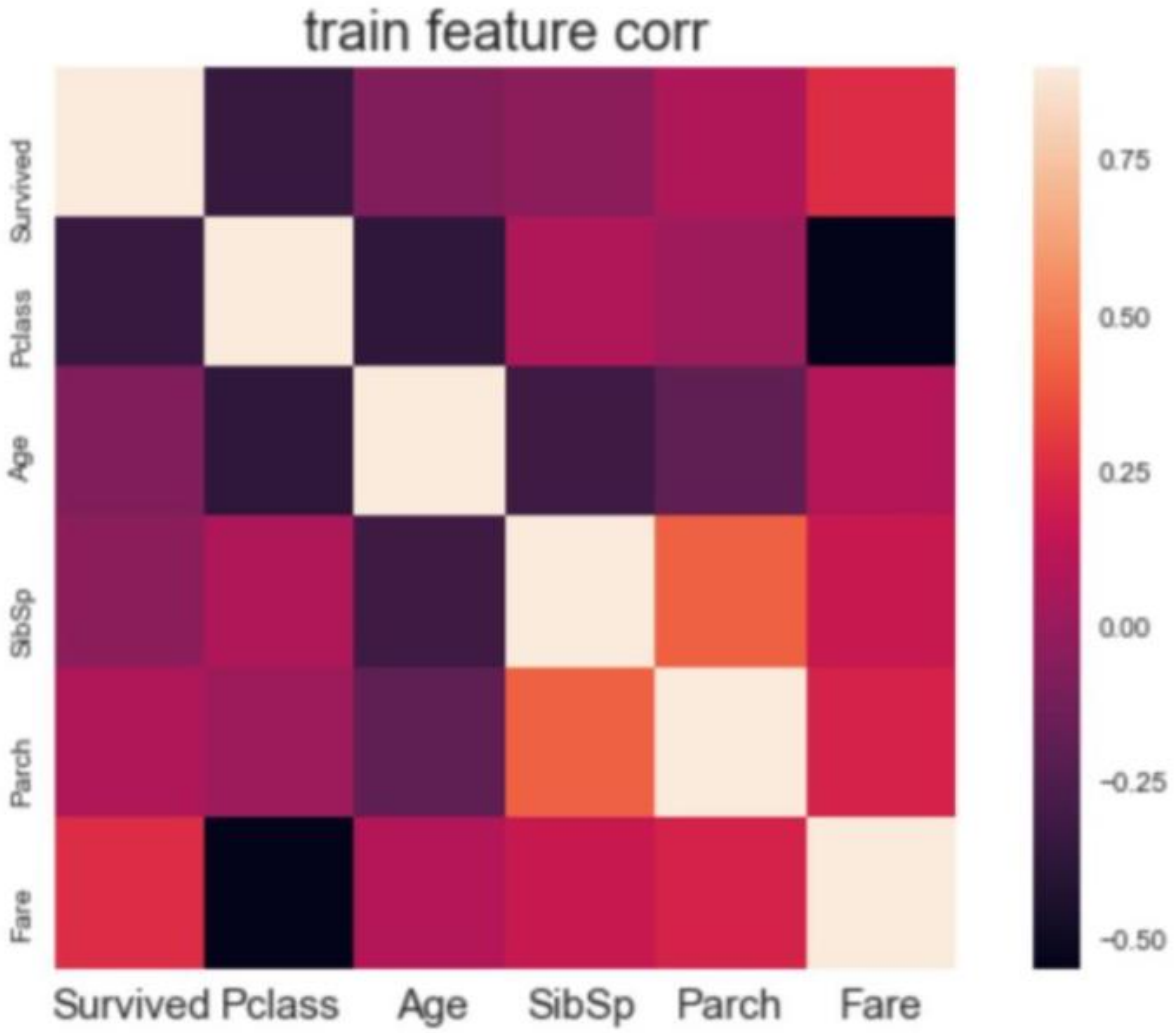

Figure Correlation matrix of training data

From the above matric, we observed something interesting which can be used in the feature selection.

1. The PClass and the Survived are negatively related, which means the more expensive cabin has higher survival rate.
2. The Sex and the Survived are negatively related, which means the women has higher survival rate.
3. The Fare and the Survived are positively related, which provides another evidence of finding 1.
4. The PClass and the Fare are negatively related, which means the higher-class cabin has higher value.

Test Set

We describe test set data as follows.

	PassengerId	Pclass	Age	SibSp	Parch	Fare
count	418.000000	418.000000	332.000000	418.000000	418.000000	417.000000
mean	1100.500000	2.265550	30.272590	0.447368	0.392344	35.627188
std	120.810458	0.841838	14.181209	0.896760	0.981429	55.907576
min	892.000000	1.000000	0.170000	0.000000	0.000000	0.000000
max	1309.000000	3.000000	76.000000	8.000000	9.000000	512.329200

Table Test set description

Attributes Observation

Age

Pclass

Sex

Fare

Sibsp & Parch

Embarked

Cabin

Ticket

Time Line

We've done the relevant works with data set.

What are supposed to do as follows:

Feature extraction & try LR/SVM/K-neighbors/DT model.

Familiar and try XGBoost Algorithm & Ensembling/Stacking.

Try Random forest/GBDT/XGBoot-GBDT & Ensembling model.

Come to a final conclusion.

References

[1] https://www.kaggle.com/c/titanic

[2] https://www.kaggle.com/daniel83fr/titanic-how-to-start-a-beginners-path/notebook

[3] https://www.kaggle.com/sagarjain2030/titanic-machine-learning-from-disaster

[4] https://trevorstephens.com/kaggle-titanic-tutorial/getting-started-with-r/

[5] https://blog.csdn.net/login_sonata/article/details/54315273