Titanic: Machine Learning from Disaster

Haoran Duan[1]

170733151, MSc Data Science, School of Computing

Abstract. Use machine learning to tackle the Titanic competition and understand the algorithm and analyze their performance.

1 Introduction

This coursework focuses on using machine learning techniques to tackle a Kaggle competition. There are two data files, the training data file which is split for training model and cross validation, and the test data file is used as unseen data for final evaluation. This work is under supervision, so it has the label which can teach our model to classify if people can survive or not.

1.1 Titanic Dataset In this dataset, the 'Survived' variable is a response variable(label), which '1' represents the survived and '0' represents the not survived.Others are features.The data quality,there some missing values in this dataset, if one feature contains many missing values, I will directly drop this feature.The missing value of other features, I will use some statistical value to fill it, such as median or mean.

1.2 Exploratory Data Analysis With the real life experiences, the 'PassengerID' is allocated randomly and the 'Cabin' has too many missing values , they can not impact the results.The 'Name' variable is not numerical type, but it can be used to calculate family size from surname, and make titles like doctor or master. The 'Sex' and 'Embarked' variables are not numerical data which need to be converted.The 'Age' and 'Fare' variables are continuous and it is possible to check the range of them,which means that different people in different range performed differently in this disaster. The 'SibSp' and 'Parch' are discrete quantitative, and they can be used to create a family size and 'whether a people is alone' variable.Because if family size is big, someone should worry about their family, which can lead to sacrifice. On the contrary, if a people is alone on the 'Titanic', he don't need worry about others and focus on leaving ship. As for the 'Ticket', different type of tickets means the different position on the ship by prefix. All the original not-numerical features will be labeled as numerical and dropped directly in new set of training data.

2 Methods

First of all, I will do some data preprocessing such as process the missing value in each column for each features and process the nominal data to numerical data.And then use three different supervised learning models to learn the classifications and do some predictions. Also, to keep the fair and consistent, I will set the random state value 43 for all of model.

2.1 Data Preprocess

Age First, I change the type of age to be the 'Int' type. And if we want to find the relationship between 'Age' and 'survived', it is necessary to make some constrain for 'Age', because in different range of age, young people are more likely to survive.So I create a new features about the 'Age' named 'AgeR'.The missing value will be filled by the median in 'Age' because of its robust property.

Fare Same as 'Age' feature,different level of price will lead to the different position, different number of people and so on.They can impact the probability of surviving. So I create a new feature 'FareR'.

FamilyN The sum of the value of 'ScibSp' and 'Parch' with the people himself can represent the family size.

Alone If family size is 1, it means that this people is alone. I create new binary feature named 'Alone'.

CallWhat Different prefix leads to different number of survived people,so I create a new features 'CallWhat' about the prefix from 'Name' feature.

Ticket There are two parts for 'Ticket',I will extract the prefix because I think it can represent the position or other things.And just drop the second part because it represent the serial number. And make them as dummy features.
All the features which do NOT have the numerical values will be labeled as numerical features.

2.2 LogisticRegression I think one of the easiest and high efficiency model for this course work is Logistic Regression.Also it is the beginning of many algorithm.The objective is to predict if people can survive, this is a binary question with the output 0 or 1. Although we can't use any linear model on it to get good prediction, fortunately, the logistic regression has the function such as 'Sigmoid', which can map the prediction in specific range like [0, 1], and they all depend on the value the output from the equation like the linear equation.

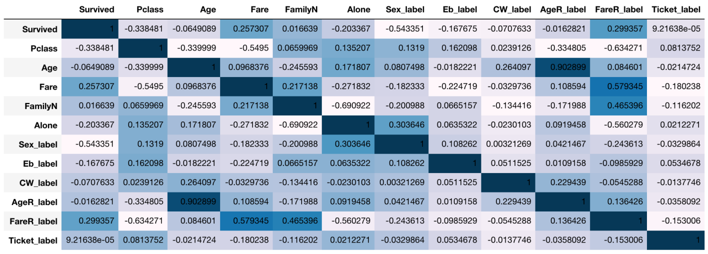
And with the help of the cost function such as least squares error or maximum likelihood estimation, which can use to find a good parameter for our model. So the smallest loss means the best parameter and the model we want. Finally, after filtering by a threshold, the classification results can be obtained.

2.3 RandomForestClassifier In this part, I will start from a DecisionTreeClassifier, and use the cross validation to find the best model with the 'max _ depth' and 'criterion'. If we use the 'Entropy', which is the measurement of uncertainty information impurity, as the criterion, the leaf nodes will have the lowest entropy. In machine learning, one of the objectives of a machine learning model is to generalize well to new data it has never seen before,so the randomforest is a better choice than decision tree, the RF constructed from many decision trees like bagging. And the reason why the RandomForest(RF) call RANDOM, because during the constructing, RF use the random sample of training data, and consider the random subsets of features when split the nodes.

3 Results

3.1 Data Preprocess

After preprocessing our data, the relationship, which between the features and labels, also between the features and features, are shown as Fig.1.



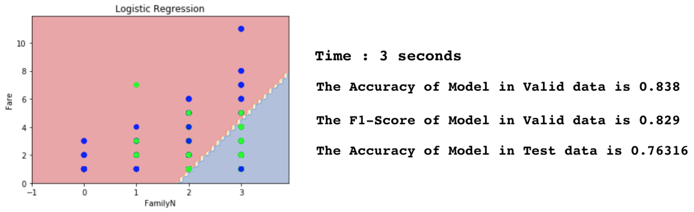
	Survived	Pclass	Age	Fare	FamilyN	Alone	Sex_label	Eb_label	CW_label	AgeR_label	FareR_label	Ticket_label
Survived	1	-0.338481	-0.0649089	0.257307	0.016639	-0.203367	-0.543351	-0.167675	-0.0707633	-0.0162821	0.299357	9.21638e-05
Pclass	-0.338481	1	-0.339999	-0.5495	0.0659969	0.135207	0.1319	0.162098	0.0239126	-0.334805	-0.634271	0.0813752
Age	-0.0649089	-0.339999	1	0.0968376	-0.245593	0.171807	0.0807498	-0.0182221	0.264097	0.902899	0.084601	-0.0214724
Fare	0.257307	-0.5495	0.0968376	1	0.217138	-0.271832	-0.182333	-0.224719	-0.0329736	0.108594	0.579345	-0.180238
FamilyN	0.016639	0.0659969	-0.245593	0.217138	1	-0.690922	-0.200988	0.0665157	-0.134416	-0.171988	0.465396	-0.116202
Alone	-0.203367	0.135207	0.171807	-0.271832	-0.690922	1	0.303646	0.0635322	-0.0230103	0.0919458	-0.560279	0.0212271
Sex_label	-0.543351	0.1319	0.0807498	-0.182333	-0.200988	0.303646	1	0.108262	0.00321269	0.0421467	-0.243613	-0.0329864
Eb_label	-0.167675	0.162098	-0.0182221	-0.224719	0.0665157	0.0635322	0.108262	1	0.0511525	0.0109158	-0.0985929	0.0534678
CW_label	-0.0707633	0.0239126	0.264097	-0.0329736	-0.134416	-0.0230103	0.00321269	0.0511525	1	0.229439	-0.0545288	-0.0137746
AgeR_label	-0.0162821	-0.334805	0.902899	0.108594	-0.171988	0.0919458	0.0421467	0.0109158	0.229439	1	0.136426	-0.0358092
FareR_label	0.299357	-0.634271	0.084601	0.579345	0.465396	-0.560279	-0.243613	-0.0985929	-0.0545288	0.136426	1	-0.153006
Ticket_label	9.21638e-05	0.0813752	-0.0214724	-0.180238	-0.116202	0.0212271	-0.0329864	0.0534678	-0.0137746	-0.0358092	-0.153006	1

Fig. 1. The relationship of Features and Labels

In my work, I will use both F1 score and accuracy to judge the performance of the model, and use the cross validation to find the best model with the best parameters. The test data will be used only one times in the final prediction, and compare with the ground truth.

3.2 LogisticRegression

Logistic Regression(or Linear Regression) is a basic model for many other complex models. For preventing the overfitting and comparing fairly, I use 10 folds cross validation to obtain the best logistic regression. The performance is shown on the right in Fig.2. And if we use the two features which have the most influence to results(Fig.1), and we can see that the blue point is ‘0’ and green point is ‘1’, our model is good for recognizing the ‘Not Survived’ people(0) so that most error is about classify and predicting the ‘survived’ people.



```
Time : 3 seconds

The Accuracy of Model in Valid data is 0.838

The F1-Score of Model in Valid data is 0.829

The Accuracy of Model in Test data is 0.76316
```

Fig. 2. Logistic Regression

3.3 RandomForestClassifier

DecisionTreeClassifier First of all, I train a decision tree classifier and use the cross validation to find the best model. After that I visualize the tree,the part of the tree(Fig.3) shows that atrribute will have the most information and the leaf nodes will has the lowest entropy.

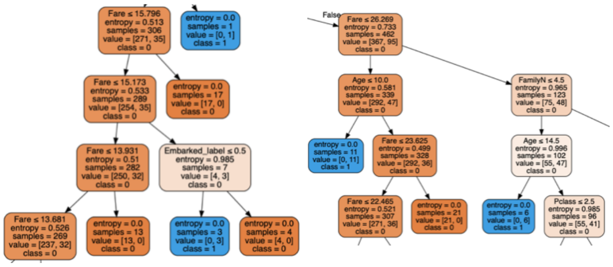
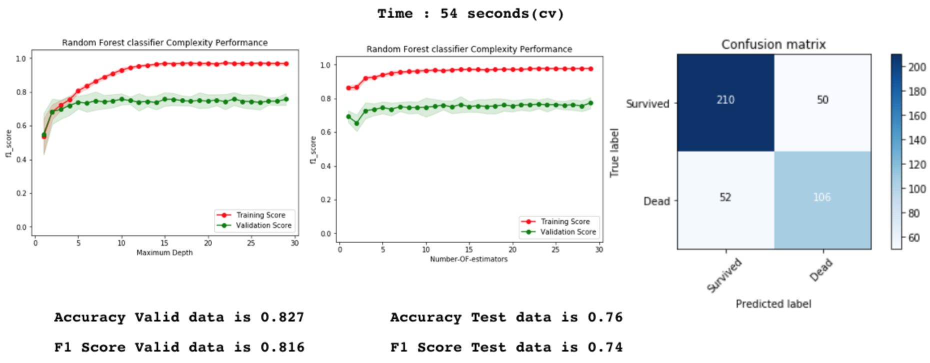


Fig. 3. Part of Decision Tree Visualization

RandomForestClassifier In this part, I will drop the ‘Parch’ and ‘SibSp’ because of the ‘FamilyN’ and ‘Alone’ contains the information. And the different depth and number of estimator will influence the model performance, but they will fit to the specific value finally(Left,middle in Fig.4). Also from the confusion matrix(Right in Fig.4)the model is not sensitive(Recall), but has a bit good precision.



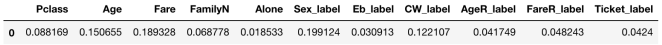
```
Time : 54 seconds(cv)

Accuracy Valid data is 0.827        Accuracy Test data is 0.76

F1 Score Valid data is 0.816        F1 Score Test data is 0.74
```

Fig. 4. The Performance of RandomForestClassifier

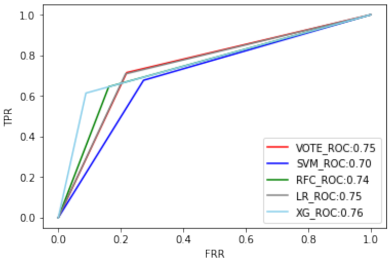
Finally,I output the parameter for the random forest model(Fig.5),the bigger the parameters are the more important the features are for the model.



	Pclass	Age	Fare	FamilyN	Alone	Sex_label	Eb_label	CW_label	AgeR_label	FareR_label	Ticket_label
0	0.088169	0.150655	0.189328	0.068778	0.018533	0.199124	0.030913	0.122107	0.041749	0.048243	0.0424

Fig. 5. Parameter about the features for the model

Other models also perform well in my processed data.The accuracy in test data of ‘SVM’ model is about 71% , 76% in the Votingclassifier and 80% in Xgboost model. Also I plot the ROC curve for these model to compare. Bigger area leads to better performance and smoother curve leads to less overfitting(Fig Below).



4 Discussion

First of all, the running time of some single models on the same dataset are almost same,such as logistic regression and decision tree, because this dataset is not big, all of our method has the same time.And the different running time is because we use cross validation in different number of parameter combination.
In my work, I think models need more than one decision boundaries because of many features, so the decision tree perform better than Logistic regression.But the decision tree is easy to be overfitting, even because it has the ability to fit all the complex the Non-linear features combination. So the RandomForest is another better choice, it use bagging method to combine many tree,and different tree start from different features,this means that they interpret the objective or the problem in different aspect. Because of constructed from many tree, the random forest is robust if there any missing or wrong value, so the generalization ability is also good. But if the random features are same,there will be some same tree, it will influence the whole model to judge the results.
This dataset is very small,so we can get the good performance by just using logistic regression,but I think it is still underfitting because there is no enough data and it has the enough diversity and information. So in the future, I think it is better to get more data and use different more complex models, such as multiLayer neural network.

5 Conclusion

Machine learnig is a tool which is used to make the computer clever by its way. Different machine learning algorithms can be used in different data, and different properties will lead the different results such as Support Vector Machine for the binary classification or linear regression for predicting a specific value. Also when we apply the different machine learning algorithm, there are some trick to make the results better like using cross validation to find the best ‘max _ depth’ for our tree model or using ensemble algorithm to reduce the overfitting and get better accuracy. On the other hand, I think data collection and data preprocessing is a key step before the analyzing, which can lead the way of algorithm.Moreover, I think the data is even the most important stuff. If we have enough data and they have enough representative, it will be that Data more important than Algorithm.