DATA PREPARATION OF THE TITANIC DATASET FOR TRAINING A RANDOM FOREST MODEL FOR THE PURPOSE OF SURVIVAL RATE PREDICTION

E.I. Gubin Ph.D., Associate Professor
Z. Jifeng, student 8PM0I
Tomsk Polytechnic University
E-mail: czifen1@tpu.ru

Introduction

The sinking of the 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. While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others [1].

For a more efficient assessment of the survival rate, it is necessary to do an analysis based EDA, Feature Engineering and Data Cleaning.

The aim of the work is to prepare data for training the model in order to predict the survival rate based on processed features.

Research methods

There are 2 datasets in our mission:

- Train data, contains all features of data.
- Test data, check the accuracy of the model created.

In the training set we can see the complete data structure and information, from this, we can get the target variable in category type (Survival/Die, 0/1 or etc.). set “Survived” – target variable. If “Survived” = 0, then “Die” and if “Survived” = 1 then “Survival”.

Before to analyze and create predictive model we need to do five steps: 1) Analysis of the features. 2) Finding any relations or trends considering multiple features. 3) Adding any few features. 4) Removing redundant features. 5) Converting features into suitable form for modeling. The reason why we use random forests algorithm is that among all the available classification methods, random forests provide the highest accuracy. The random forest technique can also handle big data with numerous variables running into thousands. It can automatically balance data sets when a class is more infrequent than other classes in the data [2].

Description of the data preparation process

In the Figure 1 show the basic columns of values.

	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

Fig. 1. Data frame of training dataset

Pclass refers to Ticket class (1 = 1st, 2 = 2nd, 3 = 3rd), SibSp and Parch are spouses and children aboard the Titanic and embarked refers to Port of Embarkation (C = Cherbourg, Q = Queenstown, S = Southampton).

Missing data mechanisms can be divided into three categories: 1) missing completely at random (MCAR), 2) missing at random (MAR), 3) Missing not at random (MNAR) [3], in our mission, missing data(Cabin) belongs to MNAR and it is not important feature, so we can start analysis features by dividing features into different types: Categorical Features: Sex, Embarked. Ordinal Features Class. Continuous Features: Age. Then, we can get correlation between the features by using Heat map like Figure 2.

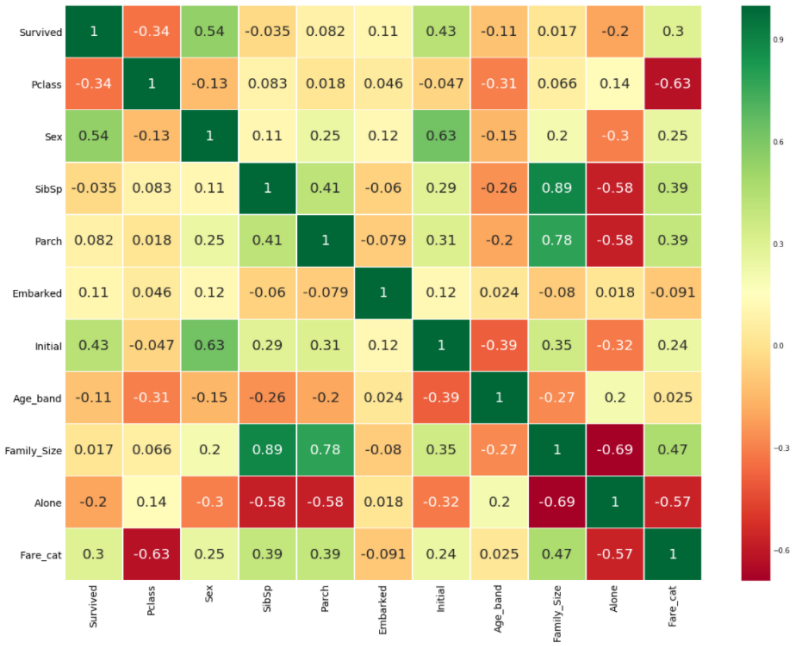

Fig. 2. Heat map of correlation matrix

Now from the above heatmap,we can see that the features are not much correlated. The highest correlation is between SibSp and Parch i.e 0.41. So we can carry on with all features.

Whenever we are given a dataset with features, it is not necessary that all the features will be important. There maybe be many redundant features which should be eliminated. Also we can get or add new features by observing or extracting information from other features. Adding new features like Age_band : group a range of ages into a single bin or assign them a single value, the maximum age of a passenger was 80, so divide the range from 0-80 into 5 bins of size 16. According to the same principle, we add features Fare_cat, Family_Size and Alone. At the same time, we should drop the unneeded features and convert string values into numeric. Here the result in Figure 3.

	Survived	Pclass	Sex	SibSp	Parch	Embarked	Initial	Age_band	Family_Size	Alone	Fare_cat
0	0	3	0	1	0	0	0	1	1	0	0
1	1	1	1	1	0	1	1	2	1	0	3
2	1	3	1	0	0	0	2	1	0	1	1
3	1	1	1	1	0	0	1	2	1	0	3
4	0	3	0	0	0	0	0	2	0	1	1

Fig. 3. Processed data

Conclusion

As a result of data preparation, we gained some insights from the Exploratory Data Analysis. It should be noted that random forest is a kind of bagging algorithm, but random forest uses CART decision tree as a weak learner, and the feature selection of decision tree is also random. Due to the randomness, it is very useful to reduce the variance of the model, so the random forest generally does not need additional pruning, that is, it can achieve better generalization ability and anti-overfitting ability (Low Variance).

References

1. Amy Tikkanen. 2020.“Millionaire’s Special”, “RMS Titanic”, “Royal Mail Ship Titanic”.
2. Leo Breiman. Statistics Department University of California Berkeley, CA 94720. 2001. RANDOM FORESTS.
3. Huang Shan, Gubin E.I. Data cleaning for data analysis // Молодежь и современные информационные технологии: Труды XVI Междунар. научно - практической конференции студентов, аспирантов и молодых ученых. Томск, 2018г. - С. 387-389.