A REPORT
ON THE TITANIC ACCDIDENT SURVIVOR PREDICTING COMPUTER PROGRAM BASED ON MACHINE LEARNING APPROACH

Submitted to

Abdus Salam Azad

Lecturer, CSE, BUET

Submitted by

Md. Mesbahur Rahman

Mahtab Noor Shaan

Muhammad Ali

31 October 2017

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1

Introduction

In this work we represent our work on predicting survival person of the tragic accident of the Titanic. For this purpose we were provided with the information about the on board passengers. Among the information there were data about names, the boarding classes of the passengers, sex, sibling-spouse, parch, ticket no., fare, cabin and information about the information of embarking. List of survivalists was also given. We were asked to run machine learning algorithm on them to predict the survivalists of the Titanic given the above information about the passengers.

CHAPTER 2

Handling Missing Data

Missing values of an attribute of a feature vector and handling them efficiently are common issues in data science. Also in this case some of the data in the ‘Age’ and ‘Embarked’ attributes. Managing those missing values was a major issue regarding the increase of the efficiency of the predicting algorithm. In this work we basically managed the missing values in four different ways. The measures available regarding the management of the missing values are given below.

2.1 Throwing out the missing Values

Throwing out the data containing missing values can be an option when the portion containing the missing values is very small compared to the whole data set. But in our case the number of the records containing missing values in the age column was 177 which are very significant compared to the number of the total number of data which is 891. So we could not follow this method at least for the ‘Age’ attribute.

2.2 Assigning a value that indicates the missing values

The missing data itself can mean a significant importance and hence can create a class of its own. So for categorical data the missing values can be represented by a constant value (like 0 or 1) to represent a class of its own. Though this method can produce great results f or categorical data, it is not suitable for use in the continuous data. Despite this fact, we used this method for ‘Age’ attribute and the result was not satisfactory as expected. The efficiency for this method

using the Decision Tree algorithm was 79.3% (Filling the ‘Embarked’ attribute by maximum occurring value, and using one-hot-encoding for the ‘Pclass’, ‘Sex’, ‘Embarked’.)

2.3 Assigning the average values

Assigning the mean average values for the missing values is a very common approach. Using this approach only, for the ‘Age’ attribute the result was very satisfactory. The efficiency for this method using the Decision Tree algorithm was 80.8% (Filling the ‘Embarked’ attribute by maximum occurring value, and using one-hot-encoding for the ‘Pclass’, ‘Sex’, ‘Embarked’.)

2.4 Using Interpolation Function

Using a interpolating function, available in Pandas, on the ‘Age’ column is another measure used by us to manage the missing data. But as the function interpolates the missing values with respect to the index of the data-frame it does not the help to increase the efficiency.

2.5 Reverse Machine Learning

In this method we used reverse machine learning method to predict the missing values of the ‘Age’ attribute. By inserting all the attributes except the passenger ID, Survived, Age attributes to the linear regression algorithm we predict the missing values of the ‘Age’ attribute. Then we use this complete ‘Age’ attribute for machine learning algorithm for predicting the survivalist of the Titanic. In this method we got 81.8% average accuracy for Decision Tree algorithm.

CHAPTER 3

Feature Engineering: variable Transformation

Scikit-learn requires everything to be numeric so we'll have to do some work to transform the raw data. All possible data can be generally considered as one of two types: Quantitative and Qualitative. Quantitative variables are those whose values can be meaningfully sorted in a manner that indicates an underlying order. In the Titanic data set, Age is a perfect example of a quantitative variable. Qualitative variables describe some aspect of an object/phenomenon in a way that can't directly be related to other values in a useful mathematical way. This includes things like names or categories. For example, the Embarked value is the name of a departure port.

3.1 Dummy variables

Also known as Categorical variable or Binary Variables, Dummy Variables can be used most effectively when a qualitative variable has a small number of distinct values that occur somewhat frequently. This is also known as one hot encoding. For this purpose we used LabelEncoder.fit_transform() in those case where class values are smaller in numbers and pandas.get_dummies() function for those case where the values of the attributes are larger in number such as name i.e.

3.2 Scaling

Scaling is a technique used to address an issue with some models that variables with wildly different scales will be treated in proportion to the magnitude of their values. For example, Age values will likely max out around 100 while household income values may max out in the millions. Some models are sensitive to the magnitude of the values of the variables, so scaling all values by some constant can help to adjust the influence of each variable. Additionally, scaling

can be performed in such a way to compress all values into a specific range (typically -1 to 1, or 0 to 1). We have used mean normalization and min-max normalization for obtaining better performance in linear regression model. By using these normalization method we obtained a efficiency increase from 70% to 75% for KNN algorithm.

CHAPTER 4

Feature Engineering: Derived Variable

An important aspect of feature engineering is using insight and creativity to find new features to feed the model. The basic transformations and interaction variables that we can automate (more on that later) don't take too much time, so that leaves us with efforts to creatively find new variables from the raw data. We derived variable from Age, Sex and Fare etc. For, example an analysis made by us is given below from which we have decided which derived variables are to make and how to make.

Table 1 Importance of various values of 'Name' attribute on classifying

Name	Survived	Died
Mr.	81	436
Mrs.	99	26
Miss	127	55
Master	23	17
Don	0	1
Rev	0	6
Dr	3	4

Name	Survived	Died
Mme	1	0
Ms	1	0
Major	1	1
Lady	1	0
Sir	1	0
Mlle	2	0
Col	1	1
Capt	0	1
The Countess	1	0
Jonkheer	0	1

From the table we can see that only title “Mr.” has a significant amount of capability to predict survival rate (84.3% chance that he has not survived) and also it contains a significant population. Other titles were discarded because it didn’t improve the result. Again, W observe from the data that the group consists of fare>50, 12<Age<50 has a great feature significance. For this reason, we have also used this as a derived feature.

4.1 Tuning Hyper parameter

We observed that ‘max_depth’=4 gives the best result for Decision Tree algorithm for our case.

CHAPTER 5

Results

The results obtained from different machine learning algorithm using various data-preprocessing are given in the table below.

Table 2 Average efficiency obtained on different algorithm

Description of Data Engineering	Average Efficiency in Decision Tree Algorithm	Average Efficiency in Perceptron Algorithm	Average Efficiency in KNN Algorithm
Replacing Missing values with -1	79.50%	55.23%	61.61%
Replacing Missing Values With Average	80.95%	55.17%	62.79%
Replacing Missing Values With Interpolated Values	79.48%	55.00%	62.84%
Replacing Missing values with values obtained from Reverse Machine	81.95%	61.86%	71.76%

Learning (Linear regression)			
Normalizing/Scaling all Features	-	-	75.36%
Considering Pclass=1 and Sex= 'female' as a new feature	81.46%	62.44%	70.72%
Including the Title of the name as a feature	82.13%	65.09%	72.21%

We also tried different combination of values of the attributes to be a new feature of the feature vector. The average efficiency obtained by doing so are given in the table below

Table 3 Effect of different variable on average efficiency

Description of the given and derived feature	Average efficiency on Decision Tree Algorithm
Sex	78.6%
Age	62.74%
Age + Sex	78.3%
Pclass	67.57%
Pclass + Age + Sex	79.35%

Pclass + Age + Sex + PassengerId	80.4%
Pclass + Age + Sex + PassengerId + Parch + SibSp	80.25%
Pclass + Age + Sex + Parch + SibSp	81.24%
Pclass + Age(ML) + Sex + Parch + SibSp + Fare	82.05%
Pclass + Age(ML) + Sex + Parch + SibSp + Fare + Name(isMr)	82.5%
Pclass + Age(mean) + Sex + Parch + SibSp + Fare +Name(isMr) + Cabin(hasCabin)	82.6%

CHAPTER 6

Conclusions and Further Works

In future work, the feature significance of different feature can be calculated bu doing 2D plot of data analysis and visualization. Many derived feature can be made by interaction of different variables. Also we can use Ensemble prediction models such as Random Forest or Neural Networks to further improve our results.