Extensive data set analysis & prediction using R

IOP Conf. Series: Journal of Physics: Conf. Series **1228** (2019) 012048 doi:10.1088/1742-6596/1228/1/012048

Extensive data set analysis & prediction using R

Padmaja Grandhe[1], Vishnu Priya Damarla[2] and Shaziya Mohammad [3]
[1]Associate Professor, CSE Department, PSCMRCET, Vijayawada, INDIA.
[2,3] B.Tech, III CSE, PSCMRCET, Vijayawada, INDIA.
E-Mail: email:padmajagrandhe@gmail.com

Abstract: Large volumes of data now available in online by several applications. Predictions about future events are difficult in case of Big data. Several applications where these predictions are required are Predicting conformation of waiting list seats in Railway reservations, prediction of some diseases based on health conditions of humans and prediction of students Grades in examination. In the sectors of medical, Railways, airlines and APSRTC fields predictive analysis is useful for taking prevention measures and for future planning. Predictive analytics is a process that comes under the data analysis. Using R we can predict Large data sets in faster manner. This paper predicts the survival of the passengers based on few factors. By considering Titanic data set analysis is performed. Based on the factors gender, class, and age survival of passengers is predicted. Decision Tree and random forest algorithms are used for prediction and for comparing the test data with trained data set.

Key words: Predictive analysis, Titanic data set, Survival Rate, Decision Tree, Random Forest. Medical Applications.

1. Introduction

Predictive analysis is required in many applications to predict the behavior or output of some instances based on the past history and outputs and outcomes already achieved. Manually To perform analytics on large data sets is very difficult and time consuming process. success of prediction is also low when its performed manually. This paper proposes a method to perform Data analysis using statistical analysis tool. Nowadays many Statistical Tools are available in market. By considering R studio here data analytics is performed. Titanic data set is considered for predictive analysis.

2. Literature survey

a. inde proposed couple of algorithms like Random forest(RF) and Latent Dirichlet Allocation(LDA) over R package in order to analyze the large volumes of data using R-studio.

b. Praveena discussed different data visualization techniques data analysis algorithms and issues related to privacy of big data in survey paper. She specified about various big data tools like HADDOP, MAP REDUCE, RAPIDMINER and other business intelligence tools for data visualization.

c. S. Bhanumathi discussed about how predicative analysis is preformed in various applications. Applications considered are health care, education, governance, consumer orientations, and hotel governeance.

d. Hyun Jeong spring proposed methods for text analytics of feedback provided by users on hotel using R studio and analytics tools.

2.1 data set

In this we are going to predict the fate of the passengers aboard the RMS Titanic, which is famously sank in the Atlantic Ocean during its voyage from the UK to New York city after colliding with the ice bar. After colliding with the ice berg 1502 were died out of 2224 passengers and crew. This sensational tragedy shocked the International community and lead to better safety measures for ships. The main reason for loss of lives is that there were less number of life boats which were insufficient for all the passengers. People who are survived in this disaster have some element of luck. Most of them are women, children and upper class passengers. The data set consists of 2 groups of data test data and train data (test.csv and train.csv). Training set is used to build the model. training set is prepared based on known factors and results. For Test data outcome is not known. By using train model and predictive analysis test data outcome is determined based on classification. From that Titanic data set we consider the train data set which consists of 891 observations of 12 variables. The variables in the train data set are Passenger Id, Survived, P-class, Name, Sex, Age, Sib sp, Parch, Ticket, Fare, Cabin, Embarked. [6]. Tabel1 describes the data dictionary of titanic data set.

Table 1: Data dictionary of Titanic data set

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

In pclass variable 1st class denotes upper class, 2nd class denotes middle class and third class denotes Lower class. The following data denotes remaining variables description.

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.

3. Proposed work

Data analyzed based on multiple attributes like age, gender, class. First step is to install R studio for data analysis and visualization. Data downloaded from Kagle[6]. train model was prepared for gender, class, and age basis. Based on decision tree random forest and logistic regression data predictive analysis is done. flow chart describes the functionality of working mechanism.Fig1 shows proposed algorithm for analysis and prediction.

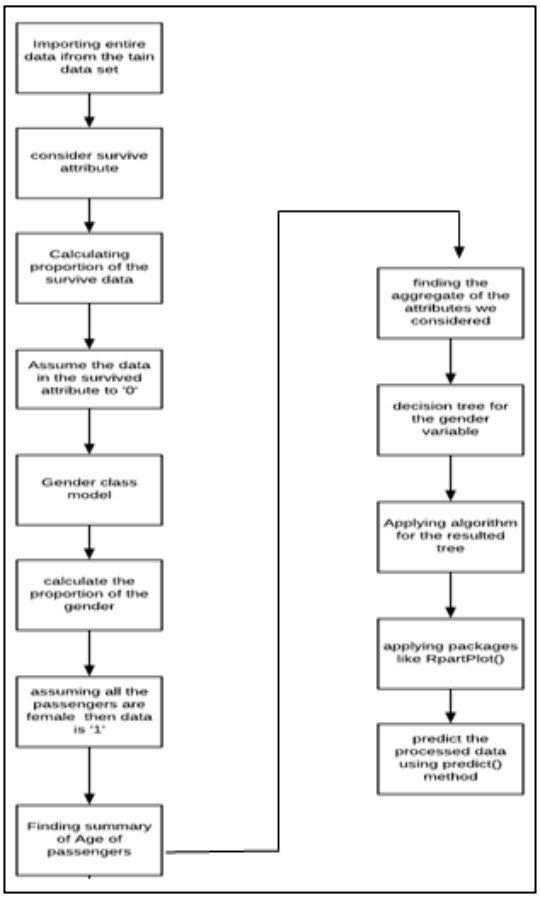

Figure 1: Proposed flow chart for predictive analysis

4. Results

Data analysis is done based on several features .considering relation factor age determining and analyzing survival predictions.Figure2 shows the analysis of survival of passengers based on sex i.e. male or female. More female people are survived than male people.

```
ggplot(data=full[1:LT,],aes(x=Sex,fill=Survived))+geom_bar()
```

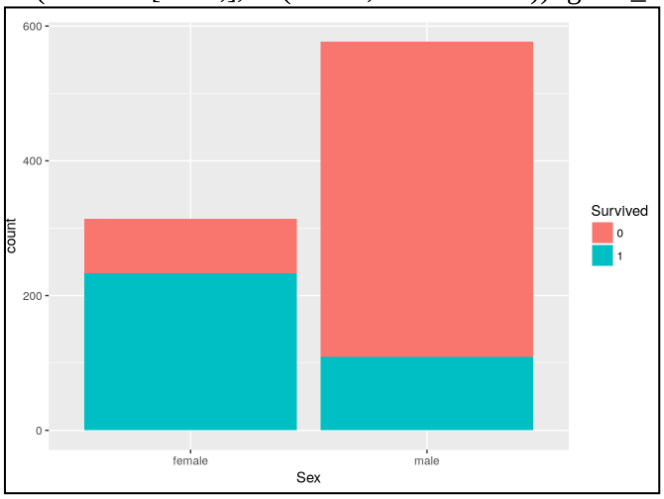

Figure 2: Survival analysis based on sex

Survival rate prediction is done based on embarked variable types c,q,s. and observed that More people are survived in c class.fig3 shows stacked bar plot of titanic passengers survival rate based on embarked.

ggplot(data = full[1:LT,], aes(x=Embarked,fill=Survived))+ geom_bar(position="fill")+ ylab("Frequency")

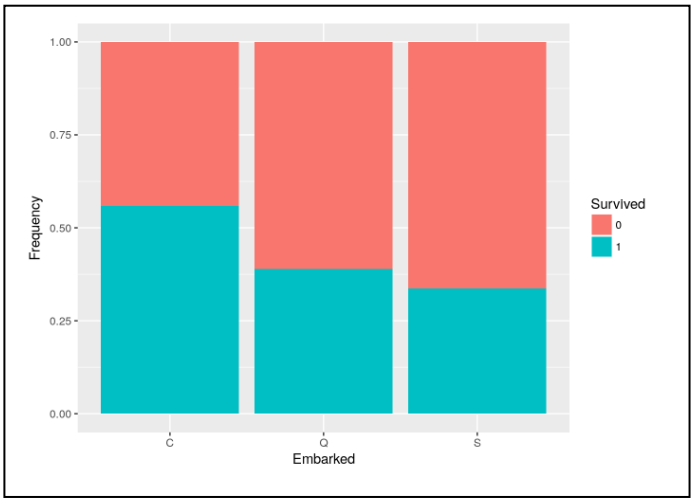

Figure 3: Survival analysis based on Embarked

Survival rate also predicted based on passenger travelling class. People who are travelling in Class 1 survived more when compared to second and third class passengers.
ggplot(data=full[1:LT,], aes(x=Pclass,fill=Survived))+ geom_bar(position="fill")+ ylab("Frequency")

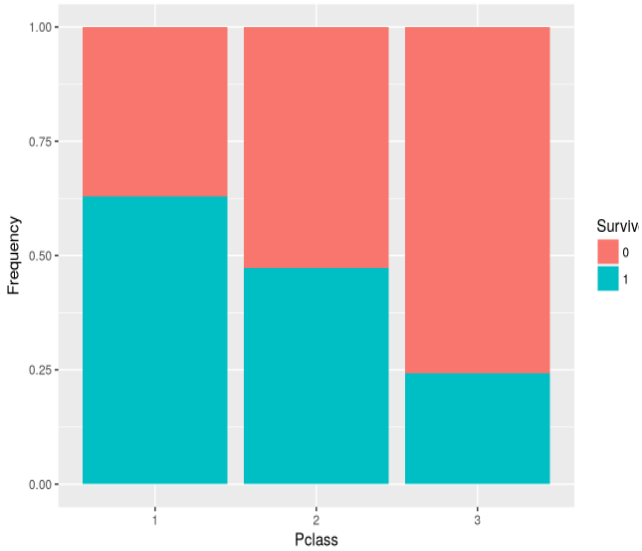

Figure 4: Survival analysis based on Pclass

Prediction is based on decision tree model. model_dt<- rpart(Survived ~.,data=train1, method="class") rpart.plot(model_dt).

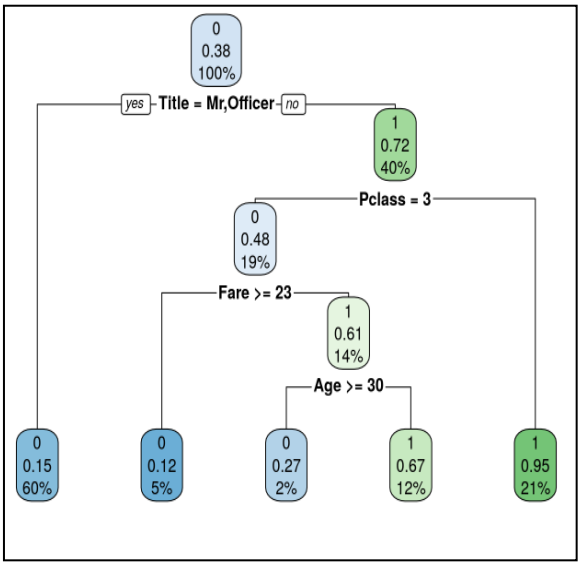

Figure 5: prediction model based on sex, pclass, and age

5. Future work

We will carry out experiment on different types of data for predictive analysis. Other sectors like medical, Automobiles, and Institutional data predictions can be done to prevent some critical situations and plan for future action according to predictions will be done.

6. Acknowledgement

The authors also gratefully acknowledge the Conference team for providing guidelines and suggestions helpful comments and suggestions, which have improved the presentation.

7. Conclusion

The mean of the right predictions that I got on the test set is 0.75 with the decision tree method, 0.78 with the logistic regression model, and 0.81 with the random forest model.

References

[1] Shinde, P. P., Oza, K. S., & Kamat, R. K. (2017). Big data predictive analysis: Using R analytical tool. In 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC). IEEE. https://doi.org/10.1109/i-smac.2017.8058297

[2] Praveena, M. D. A., & Bharathi, B. (2017). A survey paper on big data analytics. In 2017 International Conference on Information Communication and Embedded Systems (ICICES). IEEE. https://doi.org/10.1109/icices.2017.8070723

[3] S. Banumathi, A. Aloysius(2017). PREDICTIVE ANALYTICS CONCEPTS IN BIG DATA-A SURVEY in 2017 International Journal of Advanced Research in Computer Science, ISSN No. 0976-5697.

[4] Zheng Xiang, Zvi Schwartz, John H. Gerdes Jr, Muzaffer Uysal, "What can big data and text analytics tell us about hotel guest experience and satisfaction?", International Journal of Hospitality management, 2016.

[5] Hyun Jeong "spring" Han, Shawn Mankad, Nagesh Gavirneni, Rohit Verma, "What Guests Really Think of Your Hotel: Text Analytics of Online Customer Reviews", Cornell Hospitality Report, 2016.

[6] http://www.kaggle.com/c/titanic-gettingStarted/ data