Accuracy Analysis of Continuance by using Classification and Regression Algorithms in Python

Swayanshu Shanti Pragnya

Received: 15 December 2017 Accepted: 2 January 2018 Published: 15 January 2018

Abstract

- Reinforcement rate of technics and appositeness towards the convenience of the human being is a perennial mechanism. Mathematics has always been in the root towards the implementation of any algorithm or analysis regarding statistics or language. Extracting more about the data and analyzing them to solve a particular problem is the reason behind any analysis. Scrutiny itself has the different number of outcome which can be predictive or descriptive. Now prediction is how far accurate is tested by using various techniques. The enhancement in problem-solving capability leads to come up with a new aptitude concerning machine learning algorithms. But before prediction of data set collection, exploration, feature extraction, model building, accuracy testing are primarily required to invent. So for explaining all these processes, concept learning is essential. In this paper different algorithms like SVM, Linear and Logistic Regression, Decision tree, and Random forest algorithms will be used to demonstrate the accuracy in titanic data from Kaggle Website with all the required steps by using Python language.

Index terms— data analysis, machine learning, linear regression, logistic regression, random-forest, SVM, pandas and seaborn library, confusion matrix, ROC, precis

1 Introduction

ow a day's statistical analysis in any data is performed just to analyze the data little bit more by using mathematical terms. But only resolving a data is not sufficient when it comes to analysis that too by using statistics. So at this point, predictive audit comes which is nothing but a part of inferential statistics.

Here we try to infer any outcome based on analyzing patterns from previous data to predict for the next dataset when it comes to prediction first buzzword came, i.e., machine learning. So machine learning combine's statistical analysis and computer science for the prediction purpose. Machine learning also introduced to self-learning process from particular data. This learning reduces the gap between computer and statistics. Alarge amount of data prediction can be possible by human interaction as a human brain can analyze the situation with various aspects. Here the partition of algorithms occur, i.e., Supervised (used for labeled data) and unsupervised (data with no tag for learning) algorithm. As the name itself says that machine will learn, but the question arises how that is by using data. In general, by performing mistakes, we learn anything so in Machine learning these mistakes are the data which will be given to the machine to learn. But only learning is not sufficient for a model as again we need to test whatever that machine learned is it accurate or not. Here accuracy testing is required which we are going to measure by creating confusion matrix.

Before building any model in machine learning first, we need to collect the data then few preprocessing is required. Feature extraction is essential to know which features are vital in our model building. After getting the features we can build our model by using different algorithms, depending on our problem statement. Once the model is built, now we need to check its accuracy. Here we will know all the process carried out in model building. Different algorithms used like SVM, K-means, Decision tree, Random Forest, Linear and Logistic regression, from statistics standard deviation, variance analysis, Mean usability, displacement calculation and so on. All the concepts will execute by Python language and code will implement by using Jupyter Notebook.

2 a) The Need of Classification and Regression

Both classification and regression are frequently used in Data mining techniques. Regression comes into eye view when we need to predict dependant (Rely upon other attributes) variable which has relation with other data. Example-In our given Titanic data the number of survived passenger is somehow dependent upon which class the passenger is traveling as well as which cabin they were sitting. So for predicting which person survived is relative upon all these attributes so here we will use regression technique to predict.

As the name itself defines Classification is all about the categorization of data based on condition.

Support Vector Machine algorithm can give high accuracy when the data set is small and as well as less missing values in the given dataset.

Pandas: Highly used library for data analysis. Easy to understand. Open source as well as easy to use in data manipulation. Numpy: Used for scientific computing with python. Matplotlib: It is a mathematical extension from Numpy (Library for mathematical calculation) as well as primarily used for plotting graphs.

3 II.

4 Method

Linear and logistic regression [3] both used for prediction purpose. But what's the difference is much more important to know. These are the following attributes to perceive the difference between these two regression algorithms. Outcome after regression: In linear regression, the result we got is continuous whereas logistic regression has limited number of possible values. Dependent variable: Logistic regression used for the instance of true/false, yes/no, 0/1 which are categorical in nature but linear regression used in case of a continuous variable like a number, weight, height, etc. [4] Fig. **??**

5 : Linear and logistic regression

6 Equation:

Linear regression gives a linear equation in the form of Y = aX + B, means degree 1 equation But, logistic regression gives curved association which is in the form of Y = e^X/1 + e^-X

7 Minimization of error:

Linear regression (LR) uses ordinary least squares method which minimizes the error and, Logistic Regression [5] use the least square method which reduces the error quadratic-ally.

8 Support Vectors

These are the vectors (magnitude and direction) which take support for classification purpose near to the hyper plane. [2] Hyper-plane: Generally plane forms in 2 dimensions but more than 2D it is called the hyper-plane. Though support vectors drawn in more than two extent that's why it splits data through hyper-plane [2]. IV.

9 Code and Explanation

Step 1-Irrespective of any regression or classification algorithm initially need to import libraries like Pandas, Numpy, Matplotlib, Seaborn and from Scikit-learnlinear, logistic regression and SVM module.

Step 2 -Loading data in CSV file format as the data has been taken from Kaggle Titanic competition. Where train and test data set were grasped for regression.

[10]

Step 3-Select required columns in X (mostly independent variable) and in Y take dependant column as per here number of passengers survived is dependant that's why clasped in Y.

Step 4-Data cleaning and fill null values to prepare data.

Step 5-For knowing which column is influenced (value related to other column in data) more on the output column, we need to plot graphs by using regression type. [9] Step 6 -Split the data set into train and test by using Scikit-learn(free software for Machine learning libraries for Python programming).

Step 7-Fill all the null values using Mean or Dummy Values.

Step 8-Finally call regression function whether it is linear, logistic or SVM, KNN, Decision tree. [7].

Step 9-Calculate accuracy of all the algorithms and print it.

Step 10-By importing confusion matrix calculate precision and Recall to Plot the graph.

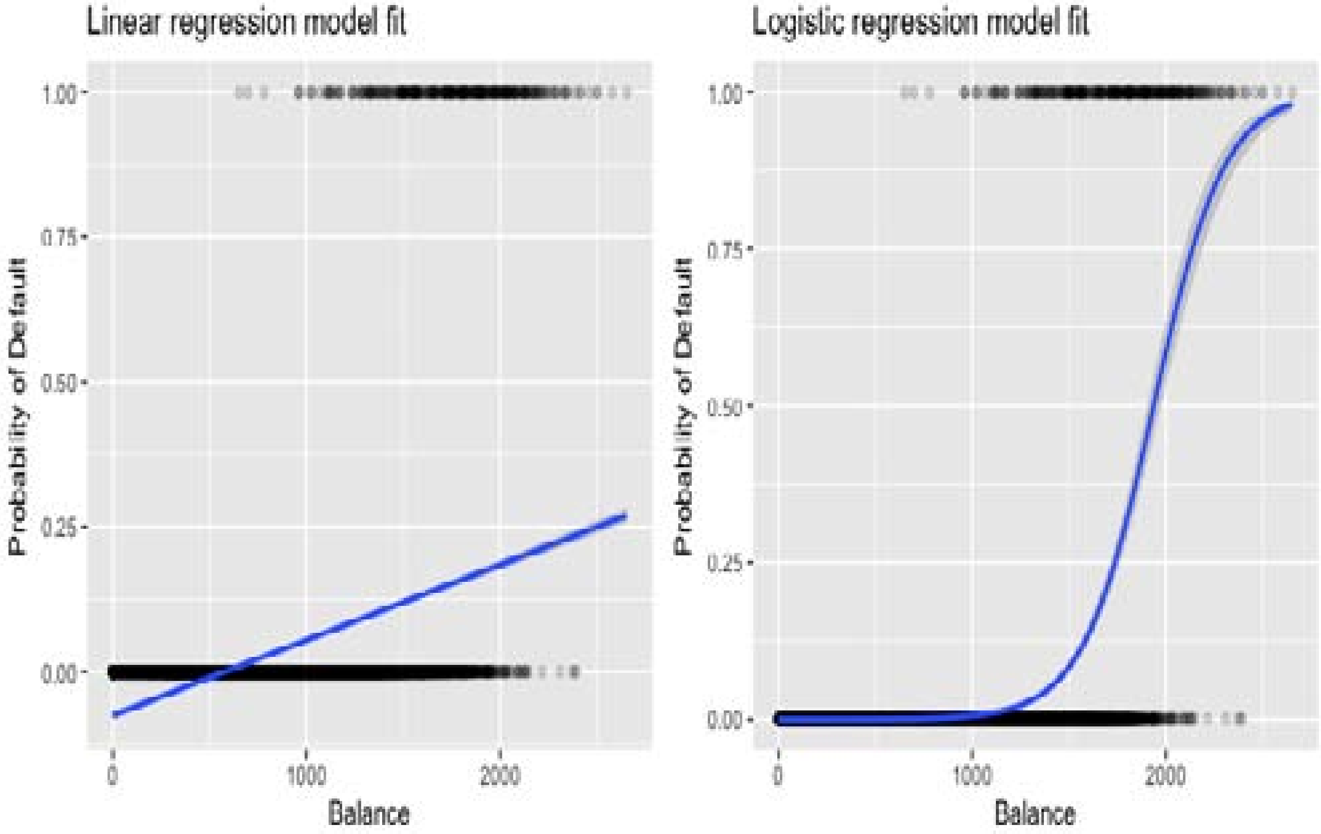
10 Result Analysis

By using the above code, we have already calculated the accuracy of each algorithm. Now by using confusion matrix, we will reckon how many numbers are correctly. Where, TP = Total positive prediction, FP = False positive and FN = False negative.

As per our result, we got Precision as 0.812101910828 and Recall as 0.745614035088. So our models have predicted 81% accurately. From the results, we got both random forest, and decision tree is giving high accuracy.

11 Conclusion

Here we have studied the basic about machine learning, linear regression, logistic regression, SVM, KNN, Decision tree and Random forest tree algorithm. We have executed the code by using python language and got the output successfully by using Confusion matrix, Precision-recall curve. At the end, we have calculated Random forest, and decision tree model are giving a higher accuracy of 92.82 % of data by using modules from scikit learn. As the objective was for knowing all these five algorithms and code execution which is computed with accuracy. We have also performed confusion matrix, for result analysis and got the result by getting the Precision and Recall value. [1] [2]

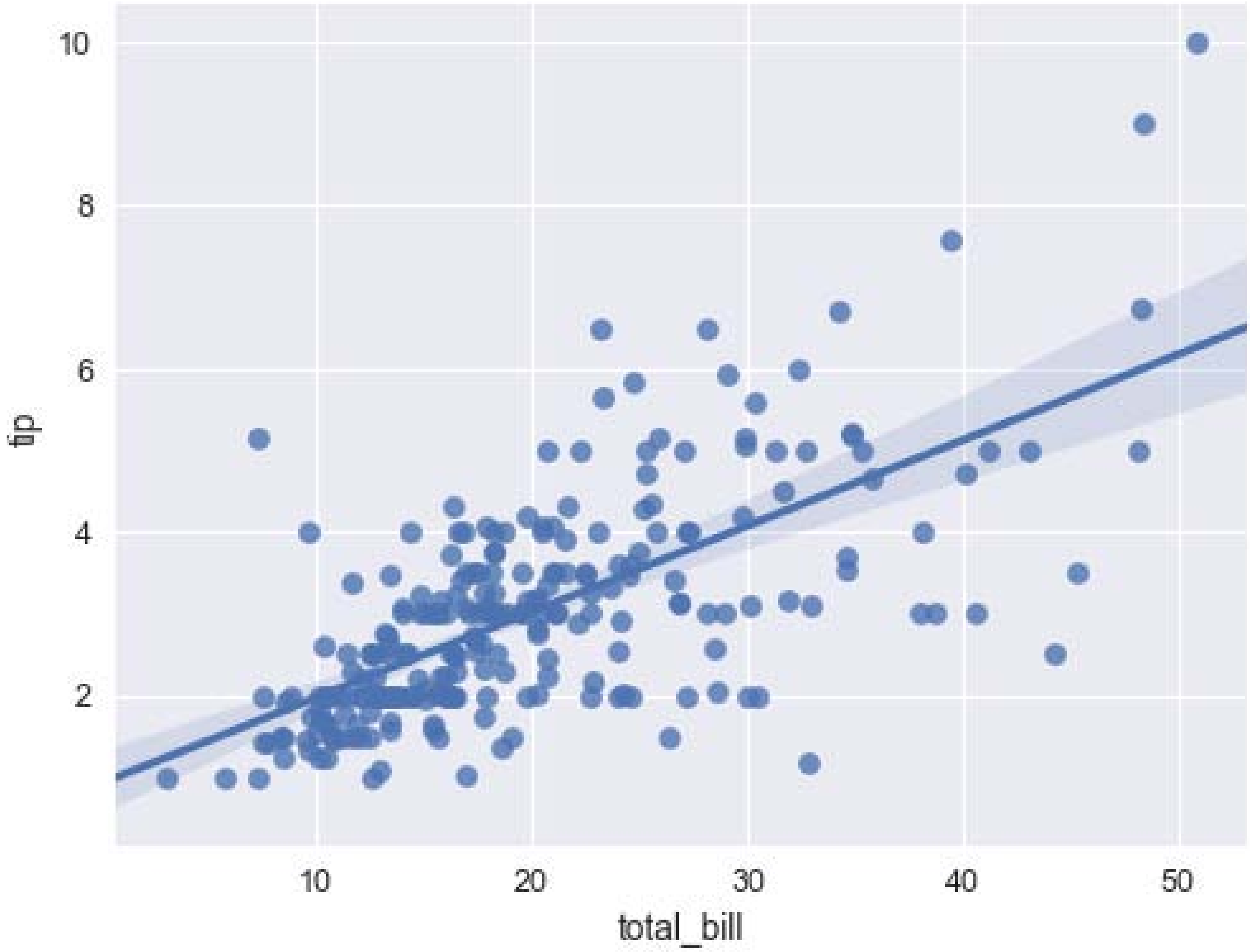


2

Figure 1: Fig. 2 :

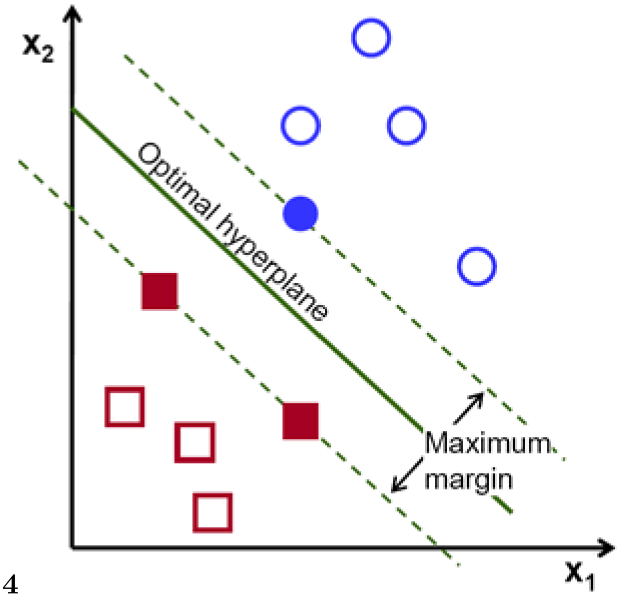
[1] © 2018 Global Journals

[2] © 2018 Global JournalsAccuracy Analysis of Continuance by using Classification and Regression Algorithms in Python



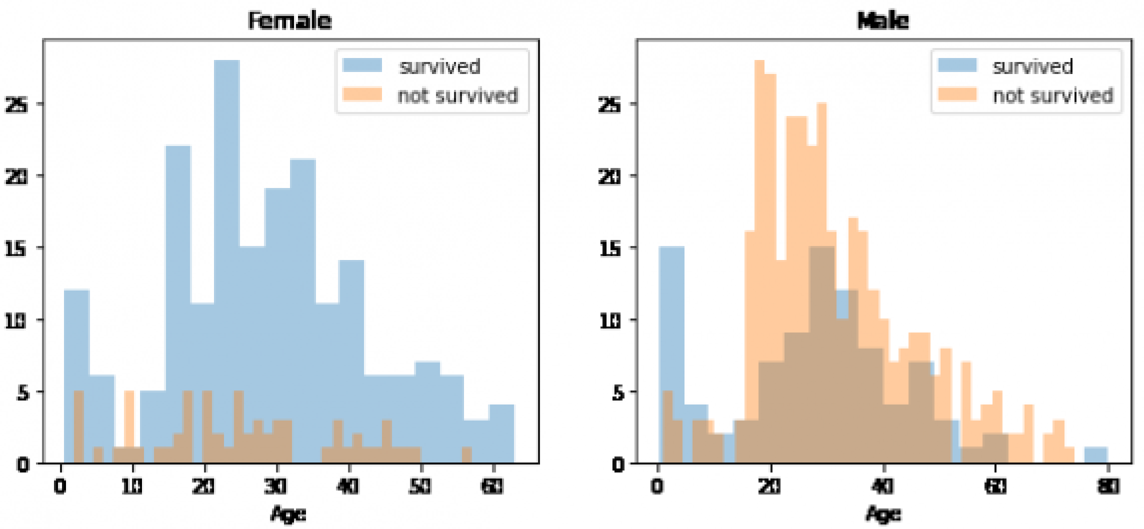
3

Figure 2: Fig. 3 :Accuracy



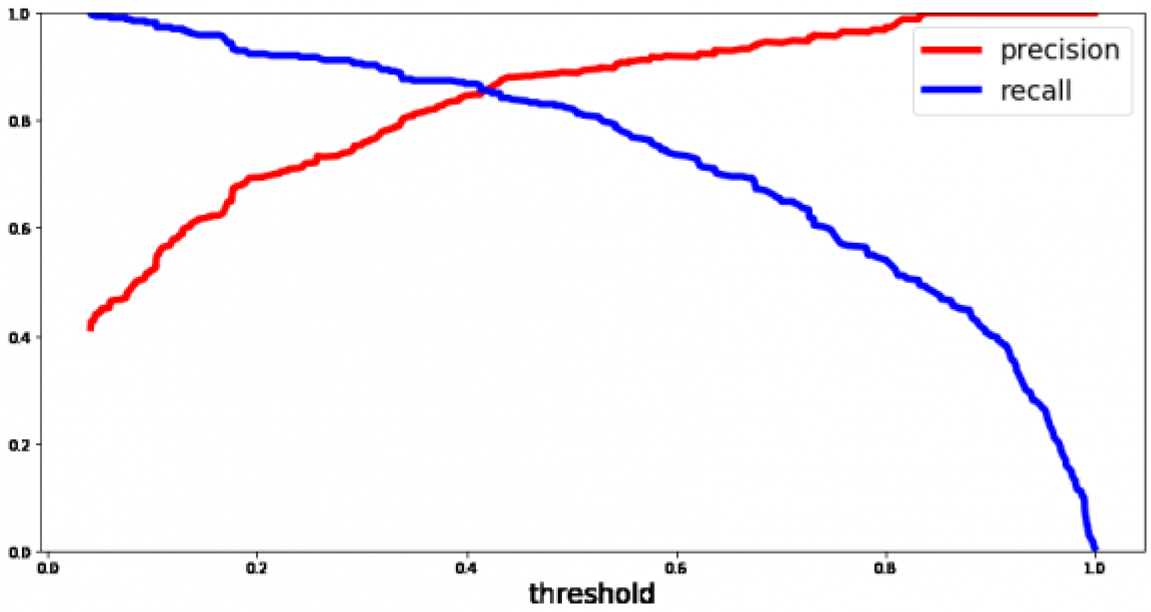
4

Figure 3: Fig. 4 :



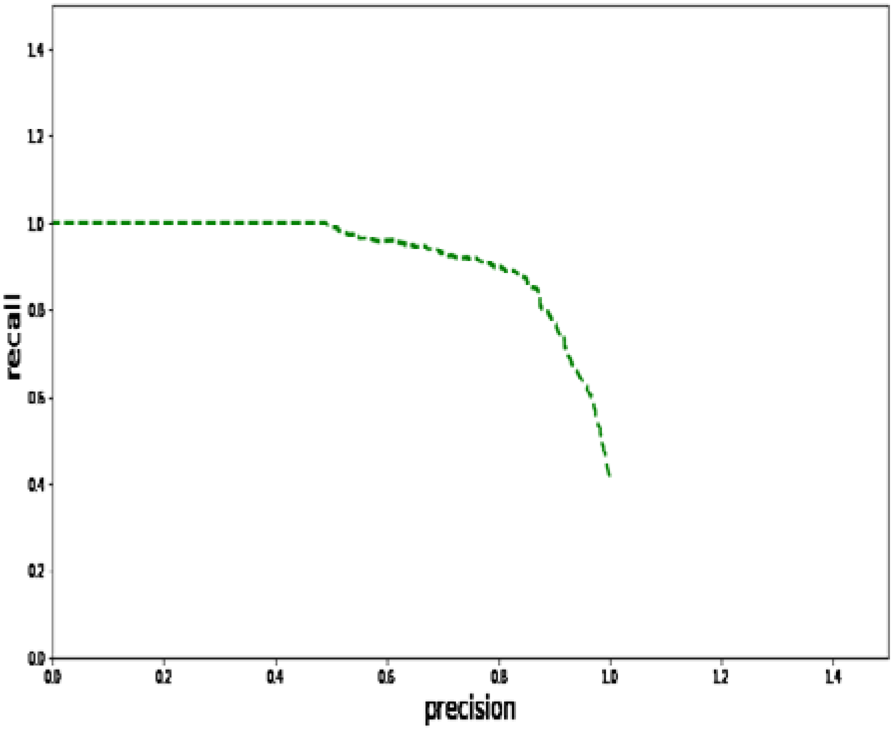
56

Figure 4: Fig. 5 :Fig. 6 :



7

Figure 5: Fig. 7 :



8

Figure 6: Fig. 8 :

[Kaggle] , Kaggle . *Data Science Community*

[A Comparative Analysis on Linear Regression and Support Vector Regression Kavitha S Assistant Professor Computer Science a *A Comparative Analysis on Linear Regression and Support Vector Regression Kavitha S Assistant Professor Computer Science and Engineering Bannari Amman Institute of Technolgy Sathyamangalamkvth*, sgm@gmail.com

[An Introduction to Logistic Regression: From Basic Concepts to Interpretation with Particular Attention to Nursing Domain Pa *An Introduction to Logistic Regression: From Basic Concepts to Interpretation with Particular Attention to Nursing Domain Park*, Seoul, Korea. Hyeoun-Ae College of Nursing and System Biomedical Informatics National Core Research Center, Seoul National University

[Available (2017)] http://www.kaggle.com/ *Available*, Accessed:2-Jun-2017.

[Flight Quest Challenge Kaggle.com (2017)] 'Flight Quest Challenge'. https://www.kaggle.com/c/flight2-final *Kaggle.com* Jun-2017. (2) . GE

[Austin et al. ()] 'Logistic regression for research in higher education'. J T Austin , R A Yaffee , D E Hinkle , S C Bagley , H White , B A Golomb . 379-410. 2. *Handbook of Theory and Research*, 1992. 2001. 8.

[Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain Jo 'Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain'. *Journal of Clinical Epidemiology* 54 (10) p. .

[Bagley et al. ()] 'Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain'. S C Bagley , H White , B A &golomb , V Bewick , L Cheek , J Ball . *Journal of Clinical Epidemiology* 2001. 2004. 54 (10) p. .

[Prediction of Survivors in Titanic Dataset: A Comparative Study using Machine Learning Algorithms Tryambak Chatterjee* De *Prediction of Survivors in Titanic Dataset: A Comparative Study using Machine Learning Algorithms Tryambak Chatterjee* Department of Management Studies*, NIT Trichy, Tiruchirappalli, Tamilnadu, India.

[Receiver operating characteristic curves Statistics review] 'Receiver operating characteristic curves'. 10.1186/cc3000. http://dx.doi.org/10.1186/cc3000 *Statistics review* 13 (6) p. 508512. (Critical Care)

[The Tragedy of Titanic: A Logistic Regression Analysis. Dina Ahmed Mohamed Ghandour1 and (May Alawi Mohamed Abdalla *The Tragedy of Titanic: A Logistic Regression Analysis. Dina Ahmed Mohamed Ghandour1 and*, May Alawi Mohamed Abdalla2.

[Titanic: Machine Learning from Disaster (2017)] Accessed: 2-. http://en.wikipedia.org/wiki/Titanic *Titanic: Machine Learning from Disaster*, Jun-2017. Jun-2017.