Research Paper

Volume : 4 | Issue : 1 | January 2015 • ISSN No 2277 - 8179

Catastrophic Modelling: Life Risk Due to Cyclone Using Random Forest

STATISTICS

KEYWORDS : Catastrophic Model, Cyclone, Early warning, Predictive Model, Random Forest.

SyluvaiAnthony , M Assistant Professors, Department of Statistics, Loyola College ,Chennai-600034

Sivanagaraju , K Assistant Professors, Department of Statistics, Loyola College ,Chennai-600034

ABSTRACT

This study attempts to create a predictive model which could act as an early warning for the catastrophic effect of a cyclone when it is 300km away from the coast line. This study is based on selected 15 cyclones which had impact in India. The predictive model utilizes mainly three parts of information to predict the life loss of a cyclone, namely Cyclone characteristics ,Weather characteristics and Population characteristics of the geographical area under threat.

1. Introduction

In this study we attempt to build a predictive model based on the historical life loss that has occurred due to 15major cyclones that had affected India. The idea is to use this model to predict the impact of a future cyclone when it is 300km away from the coast line which would give adequate time for the Government to deploy additional forces to reduce life loss.

In this study the explanatory variables for the random forest can be broadly divided into three categories namely 1.Cyclone characteristics, 2.Weather characteristics and 3.Population characteristics .The following is the list of variables under the three broad categories.

a. Cyclone Characteristics:

Cyclone Intensity at 300km,Estimated Pressure of the cyclone at 300km, Surface Wind at 300km from coast line, Pressure Drop observed at 300km from coast line, Rate of change of Cyclone Intensity, Rate of change of Estimated Pressure, Rate of change of Surface wind, Rate of change of Pressure drop, Risk category provided by meteorological department.

b. Weather Characteristics:

Rainfall, Post Monsoon/Pre monsoon Indicator, Day/Night Indicator

c.Population characteristics:

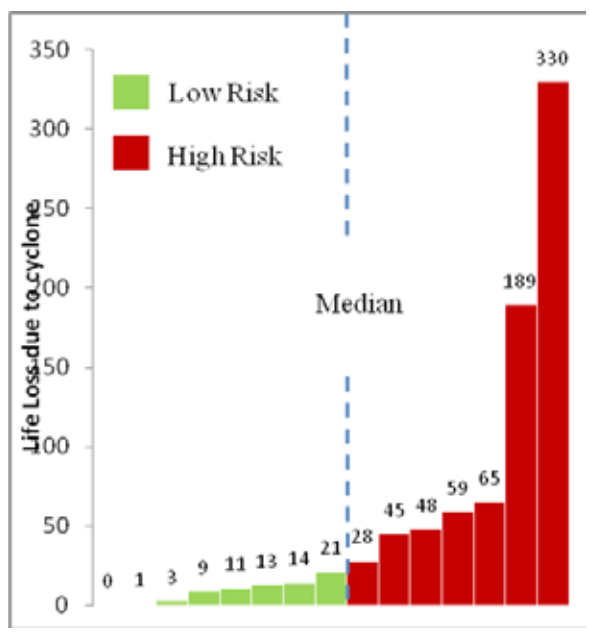
Population GDP, Population, Population per sq/km

The response variable or the dependent variable here is coded as High risk or Low risk based on the historically observed number of life loss due to cyclones.

2.Response variable and randomforest

The distribution of life loss is studied for the 15 Cyclones in the study and the cyclones are classified as high risk or low risk using median life loss as the cut value as shown in Figure1.

Figure 1



The problem reduces to a binary classification problem with 15 predictors.

Since the number of predictors are equal to the number of instances the classical general linear classification model would result in high standard error are hence we have opted for a bootstrap based predictive model namely random forest which is basically works on the principle of developing weak classifiers based on randomly selected instances and predictors.The idea behind random forest is that "Combination of weak classifiers becomes a strong classifier". Thus random forest is a suitable classification model in this scenario.

3. Rate of change measurement model

The rate of change of cyclone characteristic with respect to time is measured from the data obtained from the formation of the cyclone till the point when the cyclone is 300km away from the coast line. The rate of change is obtained using a linear regression model through the slope coefficient of the model and it is obtained for each cyclone and used as predictor variables. The rate of change measures the change in intensity of the cyclone characteristic per unit time.

4. RANDOM FOREST MODEL

The optimal input parameters for the random forest model is selected as mtry=5 and ntree =50.

Total of 50 trees were constructed by taking five predictors at random in each iteration. A sample output of 10 trees with split variables and split values are is given in Table 1.

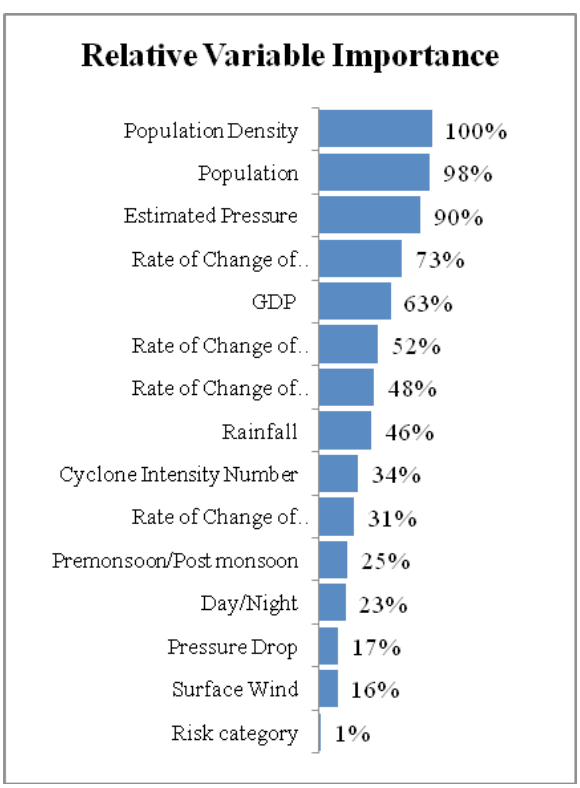
Table 1: Partial Tree Information

Tree Number	Split Variable	Split Value
Tree 7	Rainfall	376.5
Tree 13	Estimated Pressure	994
Tree 20	Pressure Drop	6.5
Tree 28	Rate of change of Pressure drop	0.4255
Tree 32	Density	521.25
Tree 40	Rate of change of surface wind	1.447
Tree 42	Rainfall	145.1
Tree 43	Rate of Change of Estimated Pressure	-1.4225
Tree 46	Density	521.25
Tree 49	GDP	15.71

5. Relative variable importance

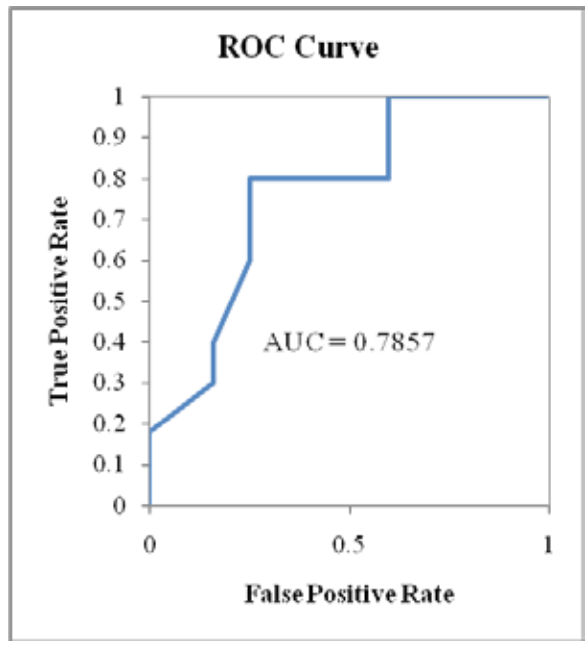
Variable importance in predicting the catastrophic level of cyclone is described in the form of relative variable importance in Figure 2.

Figure 2



From the Relative variable importance chart we can observe that the Population characteristics are most important in predicting the catastrophic impact of the cyclone followed by Cyclone characteristics and Weather characteristics.

6. Model performance



The Area Under the Curve is 0.7857

Performance of Model is observed by the method of leave one out model validation. In leave one out model validation Step i: The ith observation is left out and the remaining 14 observations are used to predict the left out observation. In this similar fashion each of the 15 cyclone is predicted as high risk or low risk. The reason for using leave one out validation is that The classification table based on the leave one out method is given in Table2.

Table 2: Classification Table based on Leave one out prediction

Classification table		Observed outcome	
		High Risk	Low Risk
Predicted Outcome	High Risk	5 (71%)	2 (25%)
	Low Risk	2 (29%)	6 (75%)

Sensitivity = 71%, Specificity=75%

7. Conclusion

The leave one out model performance indicates that the model can be used to predict the risk category of a future cyclone when it is 300 km away from the coast line with 71% accuracy of a predicted high risk and 75% accuracy on a predicted low risk.

Acknowledgement

Authors thank Dr.T.Leo Alexander, Associate professor, Department of Statistics, Loyola College, Chennai-34 for his helpful suggestions and improvement of the manuscript. Also we thank the open sources website namely www.imd.gov.in for accessing data.

Volume : 4 | Issue : 1 | January 2015 • ISSN No 2277 - 8179

REFERENCE

1. AlmuallimH,et.al(1994). Learning Boolean concepts in the presence of many irrelevant features. Artificial Intelligence, 69: 1-2, 279-306, | | 2. Alsabti K et.al(1998). CLOUDS: A Decision Tree Classifier for Large Datasets. Conference on Knowledge Discovery and Data Mining (KDD-98). | | 3. Andy Liaw et.al.(2002). Classification and Regression by randomForest.R NewsISSN 1609-3631, vol2/3,PP 18-21Merck Research Laboratories. | | 4. Breiman.L (2002). Manual on setting up, using, and understanding random forests.3.1,http://oz.berkeley.edu/users/breiman/,Using_random_forests_V3.1.pdf,18,19. | | 5. Kathryn H.,(2007). NASA Satellites Eye Coastal Water Quality, NASA Earth | ObservatoryNewscast.http://www.nasa.gov/centers/goddard/news/topstory/2007/coastal_waters.html. | | 6. Veronica Grasso et.al.(2012). Early Warning Systems: A State of the Art Analysis and Future Directions ISBN: 978-92-807-3263-4, Job Number: DEW/1531/NA, United Nations Environment Programme, Nairobi. |