www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 6 Issue 9 September 2017, Page No. 22433-22435
Index Copernicus value (2015): 58.10 DOI: 10.18535/ijecs/v6i9.06

Predicting whether songs will be hit using Logistic Regression

Abhishek Chowdhury, Sidhyant Tejas, Thirunavukkarasu K.

Student, School of Computer Science and Engineering, Galgotias University, Greater Noida, India
Student, School of Computer Science and Engineering, Galgotias University, Greater Noida, India
Professor, School of Computer Science and Engineering, Galgotias University, Greater Noida, India

Abstract:
This paper predicts whether a song will be super hit and will it reach it to chart buster based on several music features. It is crucial to predicting the popularity of a song, especially in a business competitive world. We will use Logistic Regression in our model to predict the probability of reaching Top 10 songs in the popularity.

Keywords: Logistic Regression, Classification, Hit Song.

1. Introduction

The music Industry around the globe consists of companies and individuals that make huge money by creating songs and selling them to the people by promoting their songs in live concert, shows etc. The global recorded music Industry reach $16.1 billion in 2016. Still, there is no guarantee of success as the single may become popular which led to high profit while some signal may turn out be not so popular which may lead to losses. So, the big decision problem big Music Label face whether to support an album to get profit as they get a share of profit. Our project is to see if analytics can able to predict if a song can able to make through to the Top 10 of Billboard weekly rating. Usually, a song can be said to be hit if it ever reaches to the Top 10 of Billboard weekly rating. We will use Logistic Regression Algorithm for our prediction. We will feed various song characteristics as input to our algorithm. We will then validate our model using a different method to see our accuracy.

2. Algorithm

Logistic Regression is a classification algorithm which is used to predict Binary Outcome on a given set of independent variables. The binary outcome can be Yes or No, 1 or 0 etc. Logistic Regression belongs to the family of the Generalized linear model. It returns a set of the probability of target class. Then we obtain response label using a probability threshold value.
Logistic Regression made an assumption such as the response variable follow Binomial Distribution, and the dependent variable should have mutually exclusive and exhaustive category. In the logistic regression, we are only concerned about the probability of outcome dependent variable.
The odd of success is by odd = P/1-P. In logistic regression, the dependent variable is a logit that is natural log of odds given by log(odds) = logit(P) = ln(P/1-P). In logistic regression, we find logit(P) = a + bX = y.

$\ln(P/1-P) = y.$

$P/1-P = e^y$

$P = e^y/1+e^y$

A typical logistic function given by

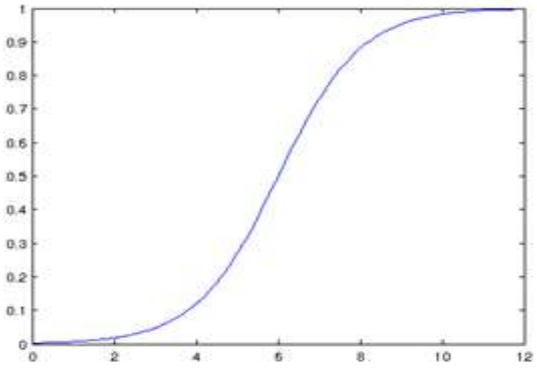

Figure 1: Logistic Function

3.Dataset

We are using the music data from the Million Song Dataset. The Million Song Dataset (MSD) is a musician feature dataset which contains more than one million contemporary songs. It is the joint collaboration between LabROSA and the Echo Nes.The meta data has a huge abstract feature which is generated from Echo Nest which is an intelligent music platform. We are using only a small subset of it which contain data from The year 1990 to 2011. It is available and taken from MIT.X Course 15.071X.

4. Model Implementation

First, we divide our data set into training and test set. Training set contain 75% of the data while testing set contain remaining 25% of the data. The number of observation is

training set = 5680

testing set = 1894

We then remove all the unnecessary variable from our set such as title,year,id etc. which doesn't add any value in prediction.

We can see how different variables are correlated to each other. Now we will feed the training set into

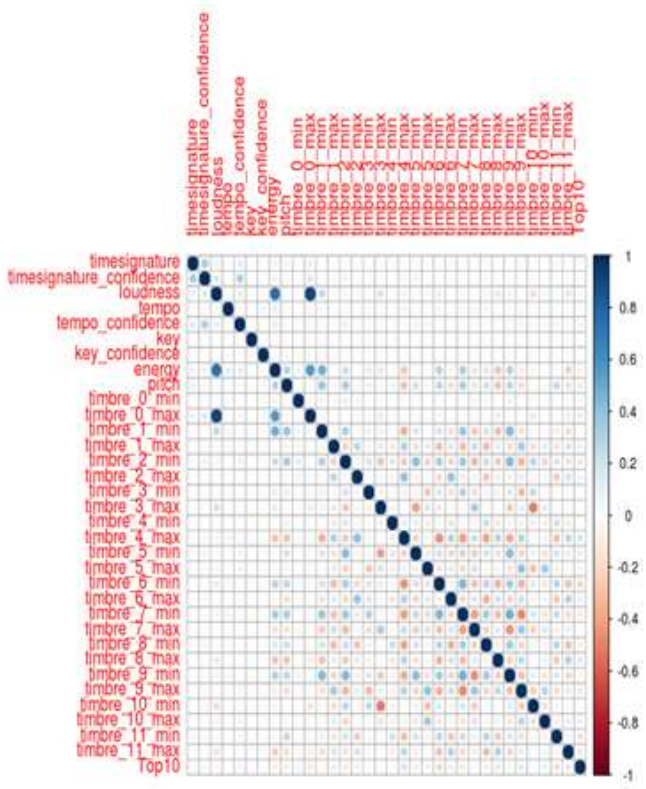

Figure 2: Correlation Between Variables

Generalized Linear Model function to train our model. We may observe different multicollinearity in our model, so we have to adjust that part by removing that variable from our model. Finally, we can ready to make a prediction. We know that our goal is to predict the songs that will make it to the Top 10 of the Billboard weekly rating or not. We will make our prediction in our test set. Once we made our prediction, we need to validate our model.

There are 3 ways to validate our model.

1) Alkaline Information Centre
2) Null Deviance and Residual Deviance
3) Confusion Matrix
4) ROC Curve

We will use confusion Matrix and ROC curve to validate our model.

Confusion Matrix is a tabular representation of Actual vs Predicted values.

We can calculate accuracy by (True positive + True Negative)/ (True positive + True Negative + False Positive + False Negative).

Here we need to know 2 Important term Sensitivity and Specificity which will be used in determining the ROC curve.

Specificity defines how many negative values, out of all the negative values, have been correctly predicted.

Specificity = True Negative/ (True Negative + False Positive).

The Sensitivity defines how many positive values, out of all the positive values, have been correctly predicted.

Sensitivity = True Positive/ (True Positives + False Negative).

And ROC summarises the model performance at a user defined threshold value. It determines the model accuracy using Area Under Curve. Higher the Area, the better is our model. It is plotted between Sensitivity and Specificity.

	PREDICTED TRUE	PREDICTED FALSE
ACTUAL TRUE	1552	62
ACTUAL FALSE	199	81

When we the make confusion matrix of our model, we get 1552 records as True Positive, i.e., we predicted 1552 songs would not make to the Top 10, and it actually didn't. 81 records as True Negative, i.e. we predicted 81 songs would make it through to the Top 10 and it did make it. 199 records were False Positives, i.e. we predicted 199 songs would not make it through to the Top 10, but it did make it. And finally, 62 records were False Negative, i.e. we predicted will make it through the, but it didn't. So finally putting the values to calculate the accuracy we get 0.8621 that is our model has an accuracy of 86.21%.

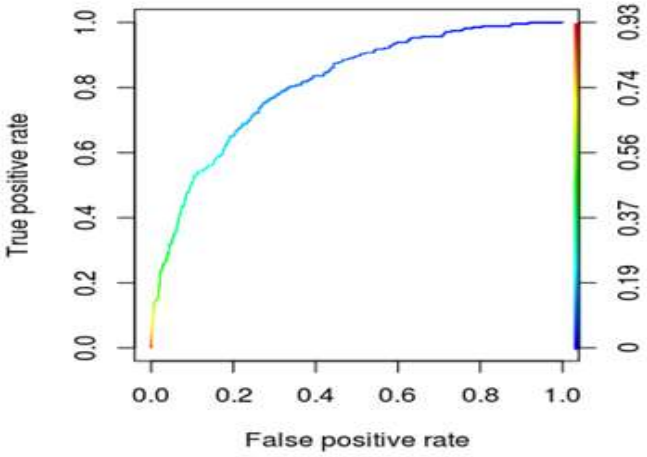

Figure 3: True Positive Rate vs False Positive Rate

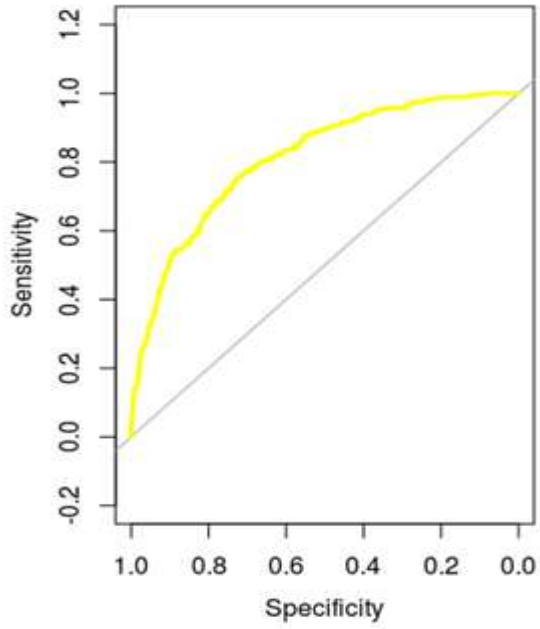

Figure 4: ROC Curve

5. Tool used

We have R Language for our project. It is an open source programming language and is used for statistical computing and graphics. R provides a wide variety of statistical and graphical techniques.

6. Conclusion

Analytics is really shaping the Industry. Through Logistic Regression Algorithm we can able to make a prediction if the song will be hit or not. Although it's not only limited to logistic regression, Various other machine learning algorithms such as Support Vector Machines can be used for prediction.

Music Industry is growing at a fast pace. It may reach $20 billion in revenue by 2020. So, it's really a high profit making a market. By Big Data Analytics many great predictions can be made which will definitely help big music label company. This project aims one such small predictions among them.

7. Reference

1. James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. An Introduction to Statistical Learning. Springer Texts, 2014.
2. David Diez, Christopher Barr, and Mine Çetinkaya-Rundel. Open Intro Statistics 3rd Edition. OpenIntro.org.
3. Jason Brownlee, Logistic Regression for Machine Learning, April 1, 2016,
4. Pachet, Francois, and Pierre Roy. ”Hit Song Science Is Not Yet a Science.” ISMIR. 2008.
5. MITx 15.071x - The Analytics Edge
6. Million Song Dataset | scaling MIR research - LabROSA