International Journal of Emerging Engineering Research and Technology
Volume 4, Issue 4, April 2016, PP 18-24
ISSN 2349-4395 (Print) & ISSN 2349-4409 (Online)

Machine Learning: Generalized Linear Modeling using R Language and the CARET Package

Amogh Tewari[1], Dr. Bindu Garg[2]

[1]*Computer Science Department, Bharati Vidyapeeth College of Engineering. New Delhi, India*
[2]*Computer Science Department, Bharati Vidyapeeth College of Engineering. New Delhi, India*

ABSTRACT

Academic performance is a very important aspect in our society. A good formal education is considered a cornerstone to any individual's development. Many factors influence the academic performance of a student in a positive or a negative way. The objective of this paper is to study generalized linear models (GLM) using the CARET (collection and regression training) package with R language. The modelling is done on student performance data with 33 variables. The modelling provides an insight into the effect of various factors in the life of a student on their academic performance.

Keywords: Predictive Modelling, Weight analysis, Statistical Analysis, Caret package, Generalised Linear Models.

INTRODUCTION

Education is an important part of modern day human's life. The value of education is signified by the emphasis put on it by the most successful people in all walks of life. This paper attempts to use a generalized linear model (GLM) fit to evaluate the weight-age of various factors that contribute positively or negatively to a student's academic performance. A generalized linear model has an advantage over the other machine learning tools that it is easier to interpret and the coefficients used within the model are directly interpretable. Many more accurate prediction models can be used to reduce error rates, but the ability of GLM to provide clarity and be more effective to this particular case is a strong reason why this technique is used.

SIZE OF A DATA SET, SEPARATION OF TRAINING AND TESTING SET IN THE DATA, OVER-FITTING AND UNDER-FITTING

A large data set is likely to have lesser noise. For example in an experiment of tossing an unbiased coin, if we consider only 2 coin toss' the possibility of getting 2 heads or 2 tails (a case where the bias of the coin would seem 100%) is 50%. However if we consider 100 coin toss', we see a more smoothened curve.

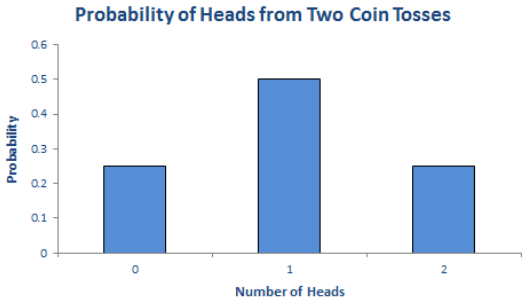

Figure1. *Bell curve for two coin tosses.*

Address for correspondence:

Amogh_tewari@yahoo.co.in

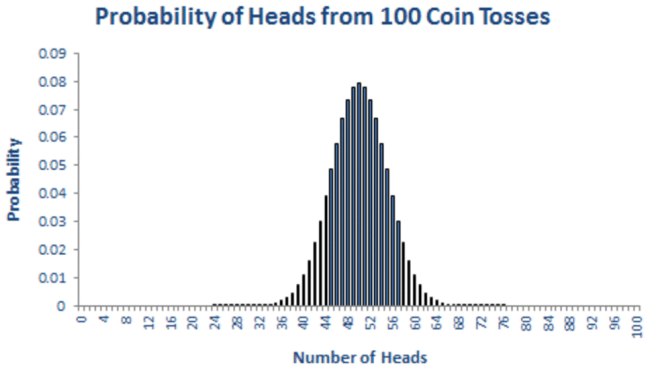

Figure2. *Bell curve for one hundred coin tosses.*

Both the figures show a Bell Curve. The first of the two figures shows more sharp characteristics as compared to the second one. Prediction models using GLM can be understood in the form of curves in n dimensional graphs, where n is the number of variables of the data set that model our graph. New predictions are done on the basis of what values the continuous curve return to the algorithm in use. So if we decide to use the first 2-dimensional graph for designing a prediction model, we would likely be subjecting it to more errors as compared to a prediction model that would be derived from the second figure. The predictions of the first model would not be inaccurate, but they would be non-precise. The more the desired precision, the bigger the data set has to be. A bigger data set does not necessarily guarantee more accuracy, but it does ensure better precision as the model has to assume lesser values. A prediction model that fits to a smoothened curve is likely to have much less errors but it can't be assumed as an error less model. Every data set can be understood in the form of a signal component and a noise component. Our objective is to build a prediction model which fits best to the signal and avoid extrapolating to the noise or fitting too loosely to the data. However, it is more or less practically impossible to avoid some noise in a data set. So, it is important for us to ensure that the data set we have is significant in volume and to choose a model that fits the data set just right.

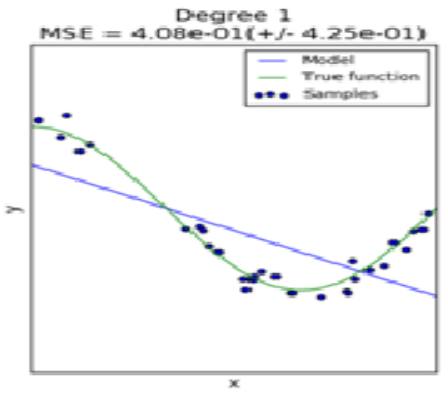

Figure3. *An under-fit curve.*

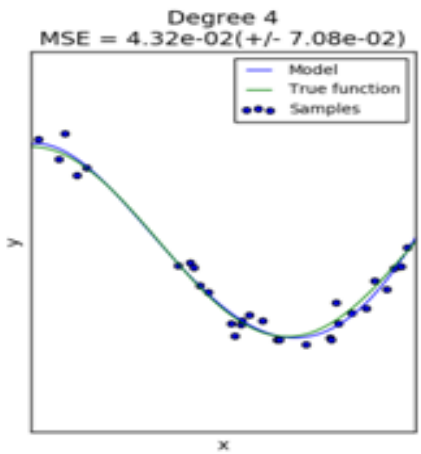

Figure4. *The right fit model.*

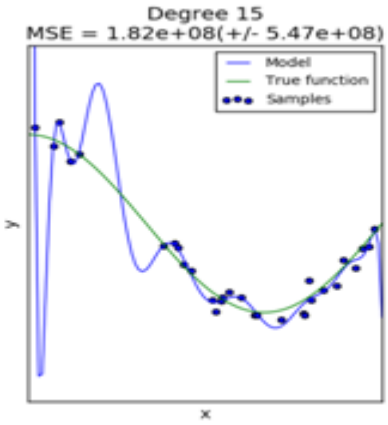

Figure5. *An over fit curve.*

We can say that an under-fit model is too generalized to be able to predict accurately and an over-fit model is closely following the data curve by including out of sample error to be able to predict accurately. It is also important while developing a prediction model to ensure that the model does not predict and realize accuracy on the very data set that it trained to. This practice would lead to an inaccurate measurement of the ability of the model due to over fitting. That is why division of the data set into separate training and testing data sets is an important aspect while a model is being configured and tested.

LITERATURE SURVEY

R is an object-oriented, procedural, multi-paradigm, functional language widely used by data miners and statisticians. R was developed by Ross Ihaka and Robert Gentleman, and named R for the first name in both developers' names. R is a very popular language which has many different tools usable to various industries but its ability to provide ease of data manipulation and statistical and graphical analysis makes it a very favorable tool. [4]

The caret package is one of the many R language libraries used for predictive modeling. Caret stands for Classification and Regression Training. It was developed and maintained by Max Kuhn. It allows us to use direct functions to train, predict and analyze our prediction models. [2]

Paulo Cortez and Alice Silva's paper on student performance data focuses on classification (binary and 5-level) and regression. [1]

Mohammad Amran Hossain and Fabio Pagnotta's paper realizes affect of alcoholism through decision tress. [6]

Education based data mining and machine learning systems for factor analysis are also the subject of research for another paper which uses several other variables more relevant to Indian educational system. [7]

Further reading on GLMs is available in John Nelder and Peter McCullagh's book on Generalized Linear Models [8]

GENERALISED LINEAR MODELS

All paragraphs must be justified alignment. With justified alignment, both sides of the paragraph are straight.

Generalized linear models as described by J.A. Nelder and R.W.M. Wedderburn in their paper on GLM's are an iteration based technique to find weights of a linear model by regression. The estimates of the coefficients obtained by the technique are the most likely values of those weights. [3]

Simply speaking, a GLM takes into account the weight of each variable in a model with the power of one (linearly) and a coefficient specific to that variable. A GLM can be represented by:

$$Y_i = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \ldots\ldots \beta_n x_n + \varepsilon_0$$

Here,

Yi is the response of the GLM, x_n is the value of nth variable for the particular tuple in the data set, B_n is the coefficient associated with the nth variable and e_0 is the error term.

In a two variable system, this is what GLM would look like,

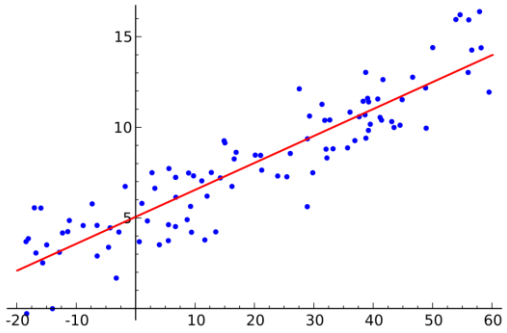

Figure6. *Two variable GLM.*

THE DATA SET

The data set was collected by Paulo Cortez and Alice Silva of University of Minho, Portugal for their paper, "Using data mining to predict secondary school student performance." [1]

The pre-processed data set data set consists of 33 variables and 395 entries with no missing data.

Variables are the following:

1. School, binary, GP or MS for Gabriel Pereira or Moisinho da Silveira
2. Sex, binary, F or M for Female or Male
3. Age, 15 to 22, defines the student's age.
4. Address, binary, R for rural and U for Urban.
5. Famsize, binary, LE3 or GT3 for less than 3 or more than 3.
6. Pstatus, binary, T or A for together or apart.
7. Medu, 0 to 4 for increasing education level.
8. Fedu, 0 to 4 for increasing education level.
9. Mjob, mother's job.
10. Fjob, father's job.
11. Reason, reason for choosing GP or MS.
12. Guardian, guardian of the student, mother, father or other.
13. Traveltime, 1 to 4, increasing time taken to travel to school.
14. Studytime, 1 to 4, increasing time of study per week.
15. Failures – 1 to 4, increasing number of times failed in classes in past.
16. Schoolsup, binary, y for extra school support otherwise n.
17. Famsup, binary, y for extra financial support from family, otherwise n.
18. Paid, binary, y for taking extra paid classes, otherwise n.
19. Activities, binary, y if taking extracurricular activities, otherwise n.
20. Nursery, binary, y for having gone to nursery school, otherwise n.
21. higher, binary, y for wanting to take higher education, otherwise n.
22. internet, binary, y if having a internet connection at home, otherwise n.
23. romantic, binary, y for being involved in a romantic relationship, otherwise n.
24. famrel, 1 to 5, increasing for better family relationships.
25. freetime, 1 to 5, increasing free time after school.

26. gout, 1 to 5, increasing with how many times student goes out with friends.
27. Dalc, 1 to 5, increasing alcohol consumption in week days.
28. Walc, 1 to 5, increasing alcohol consumption on weekends.
29. health, 1 to 5, increasing health condition of the student.
30. absences, 0 to 93, increasing number of absences from school.
31. G1, 0 to 20, grade in the first period.
32. G2, 0 to 20, grade in the second period.
33. Finalgrade, 0 to 20, target variable.

MODELLING USING CARET

First we load the student data file, which is a CSV, a comma separated variable file.

```
> studentData <- read.csv(file.choose(), header=T, sep=";")
> dim(studentData)
[1] 395  33
```

Figure7. *Loading student data file.*

As discussed earlier, the next step is an important one with which slit our data set randomly into two sets of training and testing.

```
> splitting <- createDataPartition(y=studentData$G3, p=0.7, list=FALSE)
> trainingSet <- studentData[splitting,]
> testingSet <- studentData[-splitting,]
```

Figure8. *Splitting into training and testing sets.*

The split sets contain 116 and 279 tuples, which is 0.7 and 0.3 of the total set respectively.

```
> dim(testingSet)
[1] 116  33
> dim(trainingSet)
[1] 279  33
```

Figure9. *Display of distributed data.*

To allow, reproducible results, setting the seed is an important step. Random functions in any computer system are never quite absolutely random and are always based on algorithms. The exact same random numbers can be achieved if we use the set seed function and provide it with the same value each time.

```
> set.seed(1221)
```

Figure10. *Setting the seed.*

The next step pertains to using the train function on the training set. It uses the method GLM and is modelled to predict values of the variable G3, our target variable. With it, we also generate the summary of the GLM.

```
> modelfit_generalised_linear_model <- train(G3~. , data=trainingSet, method="glm")
>
> modelfit_generalised_linear_model
Generalized Linear Model

279 samples
 32 predictor

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 279, 279, 279, 279, 279, 279, ...
Resampling results

  RMSE      Rsquared   RMSE SD    Rsquared SD
  2.301069  0.7636315  0.2064279  0.03899865
```

Figure11. *GLM.*

We get our Root mean squared error as 2.301069, Rsquared as 0.7636315, Standard Error of the Root mean squared error as 0.02064279 and standard deviation of Rsquared as 0.03899865. The final model of our prediction system gives us the value of each coefficient of the variables used in the system.

(Intercept)	schoolMS	sexM
-1.122951	0.902982	0.075322
Mjobhealth	Mjobother	Mjobservices
-0.167323	0.177596	0.329584
reasonother	reasonreputation	guardianmother
0.380906	0.180263	0.339344
paidyes	activitiesyes	nurseryyes
0.126476	-0.327597	-0.481498
Dalc	Walc	health
-0.218204	0.201359	0.044863

Figure12. *Coefficients-1.*

age	addressU	famsizeLE3
-0.221207	0.108194	-0.032491
Mjobteacher	Fjobhealth	Fjobother
0.019025	0.344201	-0.109272
guardianother	traveltime	studytime
-0.117671	0.074720	-0.135506
higheryes	internetyes	romanticyes
-0.029130	-0.048730	-0.310745
absences	G1	G2
0.059632	0.179314	0.973166

Figure13. *Coefficients-2.*

PstatusT	Medu	Fedu
-0.019532	0.110198	-0.038884
Fjobservices	Fjobteacher	reasonhome
-0.670148	-0.329582	-0.070158
failures	schoolsupyes	famsupyes
0.005536	0.417325	0.184315
famrel	freetime	goout
0.535932	0.011443	0.034931

Figure14. *Coefficients-3.*

CONCLUSION

With the absolute value of root means squared error being about 2.3 approximately when graded out of 20, the error rate of 11.5 is fairly acceptable. With that consideration, when we observe the coefficients of various variables of the data set, it is quite clear how different factors impact a student's performance. One very obvious example is the positive 0.12 factor for the variable no. 18. This variable signifies whether the student get extra paid tuitions. A positive factor indicates that it does have a net positive effect, which is the result one would expect. Other examples include a positive factor for having an urban home address and a negative factor for having access to the internet.

REFERENCES

[1] The CARET package, Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer and Allan Engelhardt (2012). caret: Classification and Regression Training. R package version 5.15-044.

[2] "Using data mining to predict secondary school student performance", Paulo Cortez and Alice Silva.

[3] "Generalized linear models", Author(s): J. A. Nelder and R. W. M. Wedderburn, Journal of the Royal Statistical Society. Series A (General), Vol. 135, No. 3 (1972), pp. 370-38

[4] R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

[5] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[6] "Using data mining to predict secondary school student alcohol consumption", Fabio Pagnotta and Mohammad Amran Hossain.

[7] "Prognostication of student's performance: factor analysis strategy for educational dataset", Aniket Muley, Parag Bhalchandra, Mahesh Joshi and Pawan Wasnik, IJEERT Volume 4, Issue 1, January 2016, PP 12 -21 ISSN 2349-4395 (Print) & ISSN 2349-4409 (Online).

[8] Generalized Linear Models, John Nelderand Peter McCullagh, Second Edition. Boca Raton: Chapman and Hall/CRC. ISBN 0-412-31760-5

AUTHORS' BIOGRAPHY

Amogh Tewari, is a Computer Science Engineering Student at Bharati Vidyapeeth College of Engineering.

Dr.Bindu Garg, is the Head of Department at Bharati Vidyapeeth College of Engineering, New Delhi. Her key research areas are Soft Computing, Analysis & Designing of algorithms and Time Series Prediction. She has more than 22 research publications in her credit.