IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, March 2022
Impact Factor: 6.252

Categorization of Veterinary data using Multivariable Linear Regression

Ayesha Taranum[1] and Dr. Shanthi Mahesh[2]
Assistant Professor, Department of Computer Science and Engineering[1]
Professor & Head, Department of Information Science and Engineering[2]
Presidency University, Bengaluru, India[1]
Atria Institute of Technology, Bengaluru, India[2]

Abstract: *"Big data" is a significant investigation of the latest vogue in technology and the unexpected impression it will have on the society, economy, and science at giant. This paper explains data is no longer treated as stale or static, but preferably, data has become an unprocessed stuff of business, an essential commercial data used to generate a new form of commercial gain. Machine learning is ideal for exploiting the opportunities hidden in big data. And unlike traditional analysis, machine learning thrives on growing datasets. The more data fed into a machine learning system, the more it can learn and apply the results to higher quality insights. Freed from the limitations of human scale thinking and analysis, machine learning is able to discover and display the patterns buried in the data.*

Keywords: Big Data, Categorization, Linear Regression, Machine Learning, Spark

I. INTRODUCTION

In the last two decades, the continuous increase of computational power has produced an overwhelming flow of data. Big data is not only becoming more available but also more understandable to computers. (Changqing Ji et al. 2012). "Big data" is a significant investigation of the latest vogue in technology and the unexpected impression it will have on the society, economy, and science giants. It aspires to describe where we stand, track down how we turned here, and provide a desperately required guide to the dangers and benefits which lie in future. Big data refer to the authors' newly discovered ability to crunch a large amount of information, disintegrate it in no time, and draw sometimes amazing interpretations from it. Big data will reorganize the techniques we think about business, education, health, politics, and innovation in the future years (V Mayer-Schönberger et al. 2013).

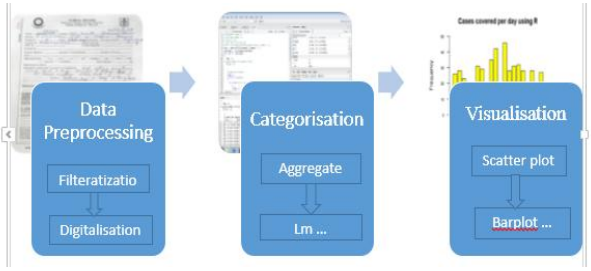
In this paper, we used data of Veterinary College (KVAFSU), Hebbal, Bangaluru. The data is about disease diagnosis and treatment given to animals at the hospital of this college. We digitalized and analyzed only 478 cases of different animals treated in this college and drew some conclusions from it using the statistical language R.

The first section of this paper deals with simple conclusions drawn from the data using an aggregate function of R language. Basic analytical methods used in business intelligence and enterprise reporting tools reduce to reporting counts, simple averages, sums, and running SQL queries. The extension of the basic analytical methods is online analytical processing, which still depends on a human intervention to specify what should be computed [3].

One of the ideal techniques for manipulating the contingencies unexposed in big data is machine learning. It commits to the assurance of eradicating excellence from big and distinct data sources with very less dependence on human supervision. It runs at machine scale and is data driven. It can deal with a large variety of variables with a great amount of data involved in it. It is also well suited for the complexity of dealing with distinct data sources. Not like traditional analysis, machine learning gives astonishing results on growth in the datasets. As the data increases in size, machine learning can apply and learn conclusions with greater quality insights [3]. The second section of this paper deals with applying the machine learning technique (i.e., linear regression) to the data and plotting the results with the help of graphical facilities of R. Third section shows how the same can be done in less time by using high performance computing.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, March 2022
Impact Factor: 6.252

II. BLOCK DIAGRAM

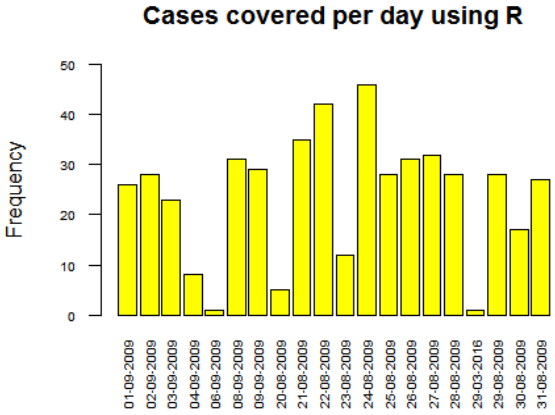


R Code to Categorize Animal Data and Plotting Graph

We categorized animal data based on the number of cases covered in one day. In the first coding, we used the data columns of the dataset and grouped and plotted the graphs based on the frequency. In the second coding, we considered the species columns of the data set to determine the different species covered by the dataset and plotted a bar graph for that.

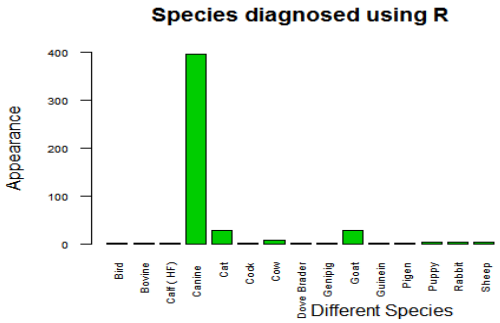
1.

```
> data <- read.csv("aish.csv")
>no.of.cases<- aggregate(x = (data$Date), by = list(Date = data$Date), FUN = length)
>barplot(no.of.cases$x,names.arg = date,las = 2,col = 7,ylim = c(0,50),main = 'Cases covered per day using R' , xlab = 'Date', space = 0.3,cex.axis=.8, cex.names=.8).
```



2.

```
>no.of.appearances.species<-aggregate(x=values, by = list(Species = data$Species),FUN = length)
>barplot(no.of.appearances.species$x,names.arg = no.of.appearances.species$Species, las = 2, col = 5,cex.axis=.8, cex.names=.6,ylim = c(0,400),main = 'Species diagnosed using R',space = 0.3)
```



This two-bar graph shows how the data can be used to draw simple conclusions. Our data contains total 478 cases.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, March 2022
Impact Factor: 6.252

III. Machine Learning

Machine learning is the science of getting a computer to act without being explicitly programmed. In the past decade, machine learning has given us an effective web for searching, practical speech recognition, self-driving cars, and a vastly improved understanding of the human genome. In this paper, we explored the most effective machine learning techniques and gain practically implementing them and graphically showing how it worked [1]. Machine learning is applicable in many areas.

3.1 Types of Problems

Types of problems applicable to Machine Learning:

Regression problem – when we need to predict some continuous output values based on a discrete list of training examples. Example: house price prediction problem (prices are continuous: can be any number 0-200K).

Classification problem – when we need to classify an input value to be part of a finite list of discrete classes. Example: Cancer diagnosis problem (2 classes: either YES or NO). Written number recognition (we have 10 classes: numbers 0-9). Machine learning is to big data as human education is to living experience: We incorporate and hypothesize from former skills to cope with unconventional circumstances. Big data on machine learning photocopy performance with a gigantic scope. Think of machine learning and big data as three stages (and phases of companies that have come out of this space): analyse, collect, and predict. These stages have been detached in the past, because we've been developing the ecological community from the bottom up — investigating different architectural and mechanism options — and developing a set of applications around that (Peter Levine, 2015).

Linear Regression with one variable

Linear regression is the most basic predictive analysis and very commonly used. The relationship between one dependent variable and one or more independent variables can be explained by estimates of regression and are also used to describe the data [4].

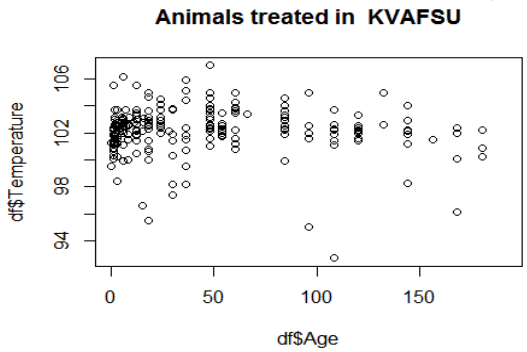
At the heart of the regression analysis is the task of fitting a single line through a scatter plot. The simplest form with one dependent and one independent variable is defined by the formula

$$y = c + b * x,$$

where y = estimated dependent,
c = constant,
b = regression coefficient, and
x = independent variable [4].

Now lets see how this simple linear regression can be applied to our veterinary data set. We have considered an independent variable. The following code shows the scatter plot for our data.

```
plot (df$Age,df$Temperature,main = "Animals treated with KVAFSU")
```

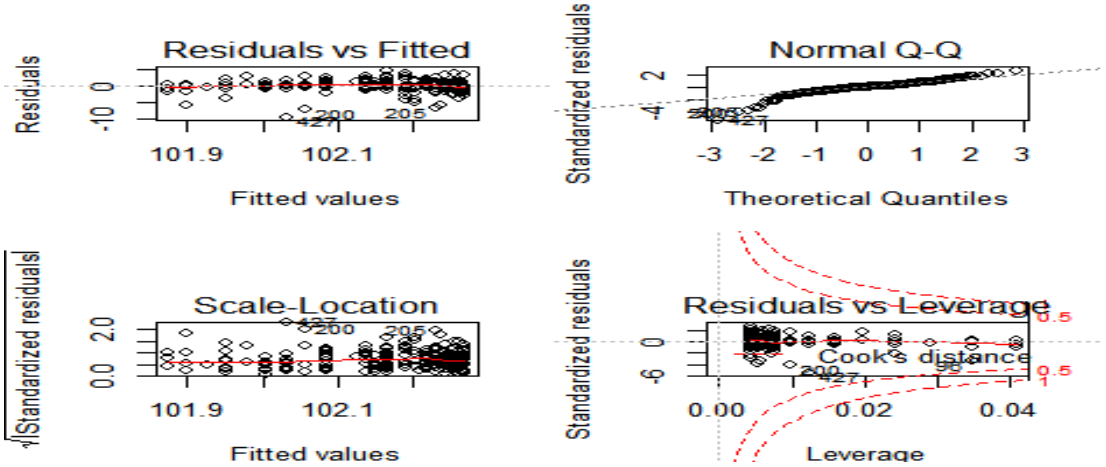



```
>demo.lm<- lm(df$Temperature ~ df$Age,data = df)
>plot(demo.lm)
> summary(demo.lm)
Call:
```

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, March 2022
Impact Factor: 6.252

lm(formula = df$Temperature ~ df$Age, data = df)
Residuals:
Min 1Q Median 3Q Max
-9.3322 -0.6764 0.1940 0.8452 4.8369

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 102.267922 0.151209 676.337<2e-16 ***



df$Age -0.002183 0.002403 -0.908 0.365

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1.726 with 237 degrees of freedom
(239 observations deleted due to missingness)
Multiple R-squared: 0.00347, Adjusted R-squared: -0.000735
F-statistic: 0.8252 with 1 and 237 DF, p-value: 0.3646 Interpreting Simple Linear Model Output
Interpreting Simple Linear Model Output
Below we briefly explain each component of the model output:. (Felipe Rego, 2015).

Formula Call

The first item shown in the output is the formula used in R for fitting the data. Note the simplicity in the syntax: the formula just needs the predictor (Age) and the target/response variable (Temperature), together with the data being used (df).

Residuals

Residuals are nothing but the difference between the original observed response values (Temperature in this case) and the response values predicted by the model. The Residuals section is divided into 5 summary points. When assessing how well the model fits the data, there should be a symmetrical distribution across these points with the mean value zero (0). In this example, the distribution of the residuals does not appear to be strongly symmetrical. Thus, this model predicts some of the points which lie far away from the actual observed points or regression line.

Coefficients

In simple linear regression, theoretically, the coefficients are nothing but two unknown constants representing the intercept and slope terms of the model. To predict the temperature of an animal of a particular age, first get the training

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, March 2022
Impact Factor: 6.252

set and produce an estimate of the coefficients by fitting it into the model formula. Ultimately, the analyst wants to determine the intercept and slope so that the ultimate fitted line is close to 479 data points of the dataset

Coefficient – Estimate

It contains two rows, intercept and slope. Intercept is the expected value of the temperature of an animal considering the average age of an animal in the data set. Second slope is the effect age has on temperature. The slope term means, for 1-month increase in age of an animal the required temperature goes up by -0.0021830F.

Coefficient – Standard Error

It is the measure of the average amount that the coefficient estimate varies from the actual average value of our response variable. In this example, it is previously determined that for 1-month increase in age of an animal the required temperature goes up by -0.0021830F. The standard error is used to calculate an estimate of the expected difference. Especially the required temperature for an animal can vary by 0.002400F. The standard error can be used to find the confidence intervals and to test the hypothesis statistically for the existence of a relationship between age and temperature.

Coefficient – t value

It is a measure of how many standard deviations the model coefficient estimate is far away from 0. If the value is far away from zero, the null hypothesis can be rejected, which means there exists a relation between age and temperature of an animal. These values are also used to calculate p-values.

Coefficient - Pr(>|t|)

This acronym in the model output is the probability of observing any value equal or greater than |t|. If this value is small, then it is unlikely to observe any relationship between the independent (Age) and dependent (Temperature) variables due to chance. The 'signif. Codes' associated with each estimate represent the significance of p-value. Three stars means a highly significant p-value. If the p-value is small, then this indicates that the null hypothesis cannot be rejected.

Residual Standard Error

It is a measure of the quality of a linear model fit. Every linear model has an error term E, because of this error term it is difficult to predict the response variable (Temperature) from the predictor (Age). The Residual Standard Error is the average amount that the response (Temperature) will deviate from the true regression line. In this example, the temperature can deviate from the true regression line by 1.7260F with 237 degrees of freedom. Out of 479 observations, 239 observations are deleted due to missing.

Multiple R-squared, Adjusted R-squared

R-squared statistics gives how well the model is fitting the actual data. This term explains the existence of a relationship between variables of the linear model. R-squared lies between 0 and 1. Quantity near to 0 explains the strong existence of this relation. In this example, R2 is 0.00347, which is near to 0. This term can be improved by increasing the number of variables in the model. This is the reason adjusted R2 value is preferred as it adjusts for the number of variables considered.

F-Statistic

F-Statistic shows whether a relationship exists between the predictor and the response variables. F-statistics depend on the number of predictors and the number of data points (Felipe Rego, 2015).

Interpreting Diagnostic Plots of Linear Regression Analysis

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, March 2022
Impact Factor: 6.252

1. Residuals vs Fitted

This plot shows are there any nonlinear patterns in the residuals. If the residuals are spread equally around a horizontal line without any distinct pattern, which is a good indication of nonlinearity.

2. Normal Q-Q

This plot shows if the residuals are normally distributed.

3. Scale-Location

Moreover, called Spread-Location plot. This plot is used to see if the residuals are spread equally along the ranges of predictors. From this plot, the homoscedasticity can be checked.

4. Residual vs Leverage

Influential cases, if any, were found with the help of this plot. In linear regression analysis, all outliers are not influential. Even if the data have extreme values, the outlier point did not have any effect on the regression line. Such data won't really matter, and are not influential, as they follow the trend in the majority of cases. Whereas some data points are very much influential even if they lie in the range of values. Such influential cases can alter the results if excluded from analysis. These points don't get along with the trend in the majority of the cases. In this example, there are no influential cases or cases as all cases are well inside of the Cook's distance line (Bommae Kim, 2015).

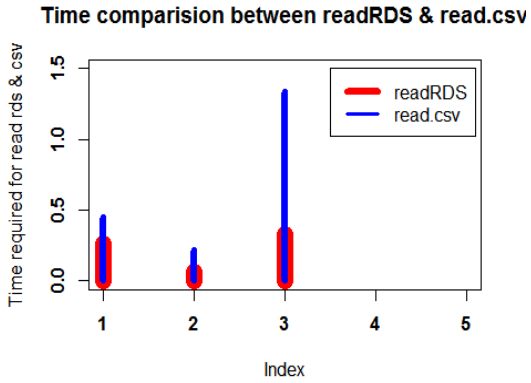
Time comparison between read.csv() and readRDS()

This section graphically shows the time comparison between read.csv() and readRDS(). In the above section, read.csv() is used to import the dataset and linear regression is performed on the same dataset. Here plot the time required to execute the Lm command by importing the dataset using read.csv(). Again plotting the time required to execute the same Lm command with the use of readRDS(), which is a better solution. saveRDS() serializes an R object and saves the representation of the object[2].

```
csvTime<- system.time(
 {
        df<-read.csv("aish.csv",stringsAsFactors=FALSE,dec = '.')
        par(mfrow = c(2,2))
          dataset<-df[,c('Age','Temperature')]
          plot(dataset$Age,dataset$Temperature)
        demo.lm<- lm(dataset$Age~ dataset$Temperature,data = dataset)
          summary(demo.lm)
          plot(demo.lm)
 })
RDSTime<- system.time(
{ saveRDS(df,"KVAFSUdata1.rds")
  KVAFSUdata2 <- readRDS("KVAFSUdata1.rds")
  identical(KVAFSUdata2,df)
  KVAFSUdata2<-KVAFSUdata2[,c('Age','Temperature')]
  KVAFSUdata2$Age <- as.numeric(KVAFSUdata2$Age)
  KVAFSUdata2$Temperature <- as.numeric(KVAFSUdata2$Temperature)
demo.lm<- lm(KVAFSUdata2$Age ~ KVAFSUdata2$Temperature,data =  KVAFSUdata2)
  summary(demo.lm)
  plot(demo.lm)
 })
par(mfrow = c(1,1))
```

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, March 2022
Impact Factor: 6.252

```
plot (RDSTime,type="h",col="red",lwd = 15,ylim = c(0,1.5),ylab = "Time required for reading rds& csv", main = "Time comparison between readRDS&read.csv",font= 2)
lines (csvTime,col="blue",lwd = 6,type = "h" )
legend (3.5,1.5, c("readRDS","read.csv"), lty=c(1,1),lwd=c(6,3),col=c("red","blue"))
```



The above plot shows the time that can be saved with the use of saveRDS() &readRDS (time in red color). The blue color line in the plot is the time required for the read.csv().

IV. Conclusion

In this paper, we used the data digitalized from Veterinary College (KVAFSU), Hebbal, Bengaluru, where sick animals are treated daily. For this paper, we used 478 cases of animals treated in this college. Data analytics is done in R language, which is a statistical language.

Firstly, we have used the dataset for some simple data analytics using aggregate function and plotting the result using the graphical functions of R.

In the second section, this paper shows how linear regression of machine learning technique is applied to the dataset and gives the details of the output of linear regression summary and diagnostic plots.

Third section shows how the serialization interface of a single object is used to minimize the time required by the familiar R functions.

In the future, we can make a comparison of this old world of "small-data" (without machine learning) business intelligence, where it is sufficient to have a small application engine that sits on top of a database, with the world of Big Data. Now, the data being processed is thousand times greater, for such a large data speed is also one factor, so a data engine that's in memory and parallel is needed. To take advantage of machine learning for big data, we're deploying it at the application layer. Which means "big data" not only needs storage but also "big compute" (Peter Levine, 2015).
Again, we can use one of the most popular engines for big data processing is Apache Spark. Spark provides a high-abstraction, platform-independent programming paradigm for very large-scale data processing by leveraging the Java framework.

Acknowledgment

The author wishes to thank Dr. Shanthi Mahesh for her guidance, review, and support.

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article [5]..

References

[1]. https://www.coursera.org/learn/machine-learning?sharebuttons_ref=.
[2]. http://www.fromthebottomoftheheap.net/2012/04/01/saving-and-loading-r-objects/

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, March 2022
Impact Factor: 6.252

[3]. http://www.skytree.net/machine-learning/why-do-machine-learning-big-data/
[4]. http://www.statisticssolutions.com/what-is-linear-regression/FLEXChip Signal Processor (MC68175/D), Motorola, 1996.
[5]. jvi.sagepub.com
[6]. http://www.warse.org/IJATCSE/current/currentDetiles/?heading=Volume%2010%20No.2%20(2021)
[7]. Andrew Ng, Machine Learning | Stanford Online,online.stanford.edu/course/machine-learning-1, 2014
[8]. Bommae Kim, "Understanding Diagnostic Plots For Linear Regression Analysis", September 2015.
[9]. Changqing Ji, Yu Li, WenmingQiu, YingweiJin, Yujie Xu, UchechukwuAwada, Keqiu Li, And Wenyu Qu, "Big Data Processing: Big Challenges And Opportunities" (doi: 10.1142/S0219265912500090), Journal of Interconnection Networks, September 2012, Vol. 13, No. 03n04
[10]. Felipe Rego, "Quick Guide: Interpreting Simple Linear Model Output in R", October 2015.
[11]. Peter Levine "Machine Learning + Big Data Predictive analytics (and where do Hadoop and Spark come in?)" Jan 2015.
[12]. V Mayer-Schönberger, K Cukier, " Big data: A revolution that will transform how we live, work, and think"- 2013 - books.google.com..

BIOGRAPHY

Ayesha Taranum
She is working in Presidency University as Assistant Professor in Computer Science and Engineering Department, Bengaluru. She has completed B.E, M.E and now Pursuing PhD. Under VTU. Her Area of Interest is Big Data and Machine Learning.

Dr. Shanth Mahesh
She is Working in Atria Institute of Technology. She is Professor and Head of Information Science and Engineering.