

Predicting Risk-of-Readmission for Congestive Heart Failure Patients on big data solutions

Mr. Balachandra Reddy (M.Tech), Dr. Harisekharan (Professor)
SRM University (Cloud Computing), SRM University (Information Technology)

Abstract - Big Data is a collection of data that is large or complex to process using on-hand database management tools or data processing applications. It is becoming very difficult for companies to store, retrieve and process the ever-increasing data. In other words we can say, Big Data is term given to humungous amount of data which is difficult to store and process. The issue lies in using the traditional system is, how to store and analyze Big Data. Risk prediction involves integration of clinical factors with socio-demographic factors like health conditions, disease parameters, hospital care quality parameters, and a variety of variables specific to each health care provider making the task increasingly complex. Unsurprisingly, many of such factors need to be extracted independently from different sources, and integrated back to improve the quality of predictive modeling. Such sources are typically voluminous, diverse, and vary significantly over the time. This project takes Apache Hadoop, an intrinsic part for storing, retrieving, evaluating and processing huge volumes of data for processing effectively. In this work, we study big data driven solutions to predict the 30-day risk of readmission for congestive heart failure (CHF) incidents. We will predict this process by using Logistic Regression and Naive Bayes classification on the basis of data collected from patients. The results are remarkable after the comparison between the two techniques and presented through confusion matrix.

Index Terms - Risk prediction, Health care, Logistic Regression, Naive Bayes classification

INTRODUCTION

Hospital readmission is expensive and generally preventable. Reducing preventable readmission is considered a key quality of care parameter that is deemed measurable. Yet, it is still challenging to develop accurate predictive models to predict such risk and the importance of factors that contribute to readmission due to the diversity of data sources even within a single large hospital. Add to this the aspiration of obtaining a holistic view of cause for readmissions by integrating socioeconomic parameters and external data with existing clinical data, and this problem becomes even more challenging and complex requiring significant advances in data integration, discretization, normalization and data Organization to name a few.

Heart failure (HF), often used to mean chronic heart failure (CHF), occurs the heart is unable to pump sufficiently to maintain blood flow to meet the needs of the body. The terms congestive heart failure (CHF) or congestive cardiac failure (CCF) are often used interchangeably with chronic heart failure. Symptoms commonly include shortness of breath, excessive tiredness, and leg swelling. The shortness of breath is usually worse with exercise, when lying down, and at night while sleeping. There is often a limitation on the amount of exercise people can perform, even when well treated. Heart failure (HF) is the most common discharge diagnosis in elderly patients, accounts for almost a quarter of all cardiovascular hospitalizations, and consumes 1% to 2% of total health care expenditures. Although a number of pharmacologic treatments have been shown to improve outcomes in patients with HF, the prognosis of these patients remains poor. Thus, there is a need for other approaches to management. Although the vast majority of HF research has focused on drug or electrical therapies, programs involving multidisciplinary teams are increasingly touted as a potential strategy for further improving outcomes in HF patients. Although some of the purported improvements may arise from better application of the evidence into practice, these multidisciplinary strategies may also better address the complex interplay between medical, psychosocial, and behavioural factors facing patients with HF and their caregivers.

Heart failure is the leading cause of hospitalization among adults >65 years of age in the United States. Annually, >1 million patients are hospitalized with a primary diagnosis of heart failure, accounting for a total Medicare expenditure exceeding \$17 billion.

Congestive Heart Failure (CHF) has been identified as one of the leading causes of hospitalization, especially for adults older than 65 years of age. Furthermore, studies show that CHF is one of the primary reasons behind readmission within a short time-span. Based on the 2005 data of Medicare beneficiaries, it has been estimated that 12.5% of Medicare admissions due to CHF were followed by readmission within 15 days, accounting for about \$590 million in health care costs. The Centre for Medicare and Medicaid Services (CMS) has started using the 30 day all cause heart failure readmission rate as a publicly reported efficiency metric. All cause 30 day readmission rates for patients with CHF have increased by 11% between 1992 and 2001. Risk prediction involves integration of clinical factors with socio-demographic factors, health conditions, disease parameters, hospital care quality parameters, and a variety of variables specific to each health care provider making the task increasingly complex. Unsurprisingly, many of such factors need to be extracted independently from different sources, and integrated back to improve the quality of predictive modelling. Such sources are typically voluminous, diverse, and vary significantly over the time.

In a recent research study, we have proposed a risk calculator tool that capable of calculating 30-day readmission risk for Congestive Heart Failure based on incomplete patient data.

The existing proposed system is predicting heart failure data using decision tree and neural network, used when historical fraud data is available and labelled; unsupervised approach, such as clustering, used when there is no labelled historical fraud data and hybrid approaches in the literature can be classified as three categories: supervised approach, such as approach, which combine supervised and unsupervised approaches and usually use unsupervised approaches to improve the performance of supervised approach. Traditionally, heuristic rules are used to predict. The intention of this paper is to ascertain risk prediction of heart failure, analyse the characteristics of heart failure data, and review and compare currently proposed prediction approaches using health care data as well as their corresponding data pre-process and discuss the future research directions.

II. DATA PERSPECTIVE

The data that is considered in the CSV (Comma Separated Value) Format .This may be either the text or any kind of format .For simplicity most of the analysts have prefers the CSV Format. The data reveals the following Information:

- Name of the patient
- sex
- age
- B.p
- sugar
- cholesterol
- chest pain
- breathlessness
- Giddiness
- palpitation
- oedema

Heart attack mainly based on cholesterol; here I am giving a small description about cholesterol

These plaques are the main causes of heart attacks, strokes, serious medical problems, leading to the association of so-called LDL cholesterol. Measures used to lower the plasma lipids in patients with hyperlipidaemia will lead to reductions in new events of coronary heart

Disease. Another formulation is that "decreasing blood cholesterol...significantly reduces coronary heart disease events", this discussion is also referred to as the "cholesterol controversy". It is closely related to the saturated fat and cardiovascular disease.

	A	B	C	D	E	F	G	H	I	J	K	L	M
	Name of the Patient	Sex	Age	Date of Joining	Heart Beat per min	B.P	sugar	cholesterc	Chest Pair	Breathless	Giddiness	Palpitation	Oedema
1	Alla Baksh	male	52	12-09-2012	92		95 yes	yes	yes	yes	no	yes	yes
1	ALLI BABU PINJARI	male	49	08-07-2009	98		130 yes	yes	no	yes	yes	yes	yes
1	RAMANAMMA	female	57	06-05-2008	82		85 yes	no	yes	no	yes	yes	no
1	SARATH REDDY	male	54	08-06-2010	95		80 no	no	yes	yes	no	yes	no
1	Sri Hari	male	58	07-07-2011	80		150 yes	yes	yes	no	yes	no	no
1	zaheera pandlapuram	female	51	02-01-2012	76		140 yes	no	no	no	yes	yes	no
1	Devaiah	male	60	09-11-2012	78		100 yes	yes	yes	yes	no	yes	yes
1	Narasamma	female	60	10-01-2012	65		120 yes	no	yes	no	yes	no	yes
0	Kannan Kotadi	male	53	09-02-2013	74		130 yes	yes	yes	yes	yes	yes	no
1	p.venkatesh	male	55	10-10-2012	68		152 yes	yes	yes	yes	yes	yes	no
2	S.Gowri Shankar	male	48	02-01-2013	73		80 yes	yes	no	no	no	yes	yes
3	Nazem basha	male	53	02-01-2012	76		152 no	no	yes	no	no	yes	no
4	Manasa Gurapadu	female	50	02-04-2010	82		125 yes	yes	no	no	yes	yes	no
5	Sabhanna Begam	female	52	08-02-2012	70		100 yes	yes	no	yes	no	yes	yes
6	Narayana	male	50	04-04-2012	98		150 no	no	no	yes	no	yes	yes
7	chenga reddy	male	49	05-07-2014	78		120 yes	yes	no	no	yes	yes	no
8	venkata ramana	male	52	09-08-2012	84		120 yes	no	yes	yes	no	no	no
9	aruna kumari	female	48	08-07-2008	74		110 yes	yes	no	yes	no	yes	yes
0	anil goud	male	53	08-07-2007	76		122 yes	no	yes	yes	no	no	no
1	rajasehar	male	47	07-11-2009	84		95 yes	yes	no	yes	no	yes	no
2	mallikarjuna	male	53	01-12-2011	82		120 yes	no	yes	yes	no	no	no
3	sekhar	male	54	08-11-2012	88		140 yes	yes	yes	yes	no	no	yes
4	mahesh	male	58	04-08-2007	90		98 no	yes	no	yes	yes	no	no
5	wasim	male	50	09-08-2010	74		110 yes	no	yes	no	yes	no	yes
6	basha	male	53	09-08-2009	82		115 yes	no	no	yes	yes	no	no
7	aswathi	female	48	08-08-2011	84		110 yes	no	no	yes	yes	no	no
8	sudheer	male	52	09-07-2012	88		140 yes	no	yes	no	no	no	yes

Fig. 1 Dataset table

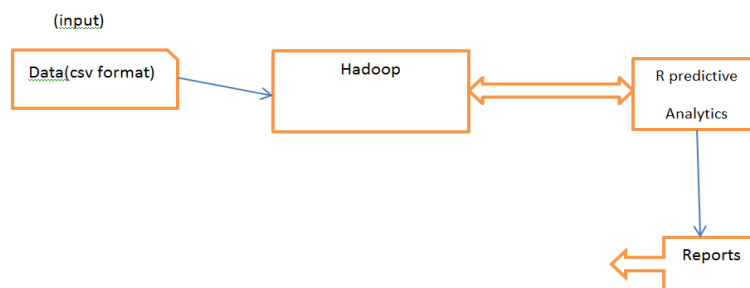


Fig. 2 Architecture diagram

The above figure tells about flow of data from system to another system and its processing

HDFS:

The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware. HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets.

Applications that run on HDFS have large data sets. A typical file in HDFS is gigabytes to terabytes in size. Thus, HDFS is tuned to support large files. It should provide high aggregate data bandwidth and scale to hundreds of nodes in a single cluster. It should support tens of millions of files in a single instance.

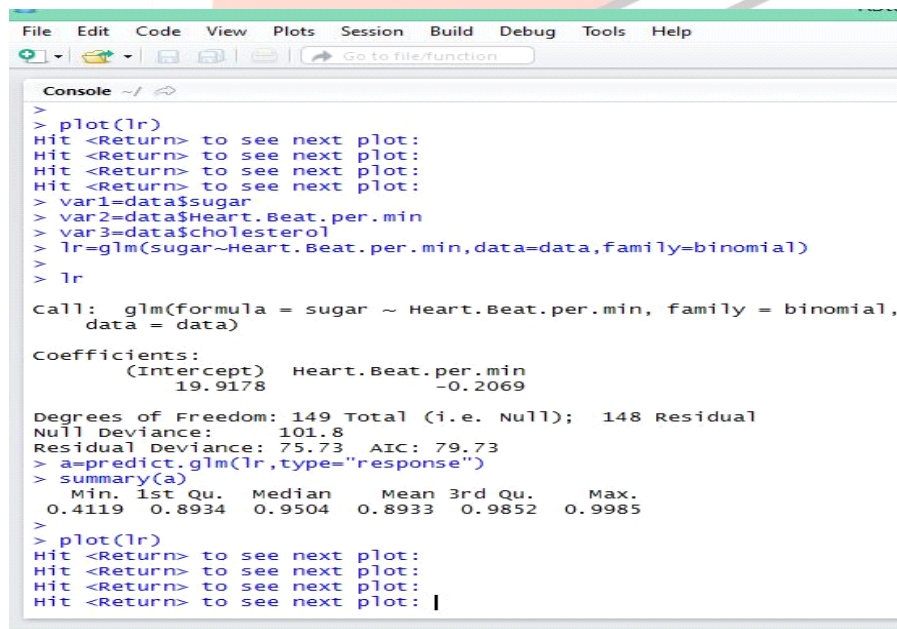
RHadoop:

R is a programming language and a software suite used for data analysis, statistical computing and data visualization. It is highly extensible and has object oriented features and strong graphical capabilities. At its heart R is an interpreted language and comes with a command line interpreter – available for Linux, Windows and Mac machines – but there are IDEs as well to support development like RStudio or JGR. R and Hadoop can complement each other very well; they are a natural match in big data analytics and visualization. One of the most well-known R packages to support Hadoop functionalities is RHadoop that was developed by Revolution Analytics. RHadoop is a collection of three R packages: rmr, rhdfs and rhbase. Rmr package provides Hadoop Map Reduce functionality in R, rhdfs provides HDFS file management in R and rhbase provides Hbase database management from within R.

Techniques**a) Logistic regression**

Logistic regression, or logit regression, or logit model is a type of probabilistic statistical classification model. It is also used to predict a binary response from a binary predictor, used for predicting the outcome of a categorical dependent variable (i.e., a class label) based on one or more predictor variables (features). That is, it is used in estimating the parameters of a qualitative response model. The probabilities describing the possible outcomes of a single trial are modelled, as a function of the explanatory (predictor) variables, using a logistic function. Frequently (and hereafter in this article) "logistic regression" is used to refer specifically to the problem in which the dependent variable is binary—that is, the number of available categories is two—while problems with more than two categories are referred to as multinomial logistic regression or, if the multiple categories are ordered, as ordered logistic regression.

Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables, which are usually (but not necessarily) continuous, by using probability scores as the predicted values of the dependent variable.[3] Thus, it treats the same set of problems as does probate regression using similar techniques; the first assumes a logistic function and the second a standard normal distribution function.



```
File Edit Code View Plots Session Build Debug Tools Help
Go to file/function

Console -/
>
> plot(lr)
Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot:
> var1=data$sugar
> var2=data$Heart.Beat.per.min
> var3=data$cholesterol
> lr=glm(sugar~Heart.Beat.per.min,data=data,family=binomial)
> lr

Call: glm(formula = sugar ~ Heart.Beat.per.min, family = binomial,
data = data)

Coefficients:
      (Intercept)  Heart.Beat.per.min
           19.9178             -0.2069

Degrees of Freedom: 149 Total (i.e. Null); 148 Residual
Null Deviance: 101.8
Residual Deviance: 75.73 AIC: 79.73
> a=predict.glm(lr,type="response")
> summary(a)
      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.4119  0.8934  0.9504  0.8933  0.9852  0.9985
>
> plot(lr)
Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot: |
```

Fig. 3 Logistic regression

(b) Naive Bayes Classification:

In machine learning, naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features.

Naive Bayes has been studied extensively since the 1950s. It was introduced under a different name into the text retrieval community in the early 1960s, 488 and remains a popular (baseline) method for text categorization, the problem of judging documents as belonging to one category or the other (such as spam or legitimate, sports or politics, etc.) with word frequencies as

the features. With appropriate pre-processing, it is competitive in this domain with more advanced methods including support vector machines. It also finds application in automatic medical diagnosis.

Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.

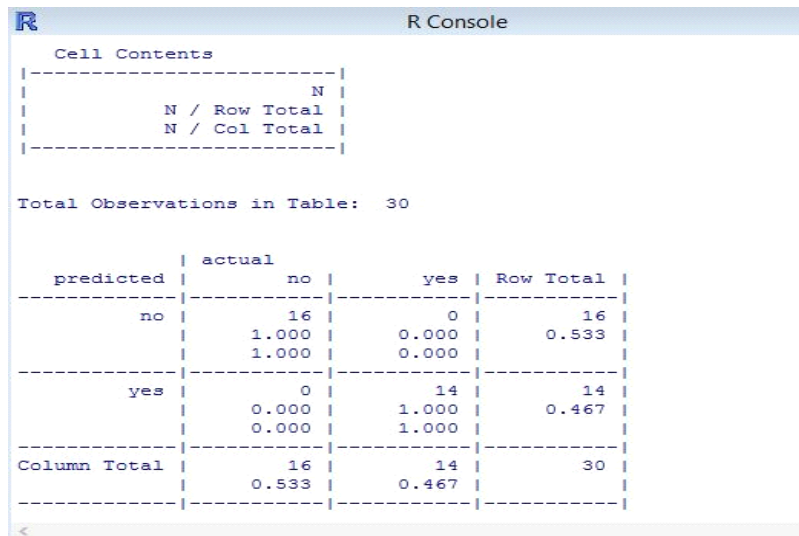


Fig. 4 Bayes classification

Confusion Matrix:

In the field of machine learning, a confusion matrix, also known as a contingency table or an error matrix, is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one (in unsupervised learning it is usually called a matching matrix). Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. The name stems from the fact that it makes it easy to see if the system is confusing two classes.

It is mainly used for predictions; here comparison will be shown between Logistic regression and Naive Bayes classification algorithms. Mainly, which one is more predictable than other. Here we will show some predictable graphs

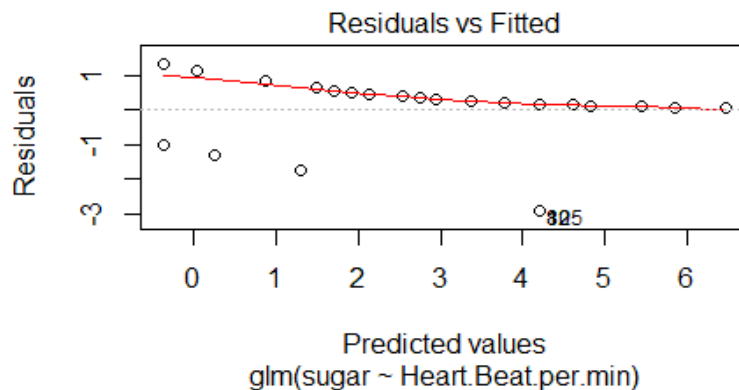


Fig. 5 The graph between sugar and heart.beat.per.min in Logistic regression

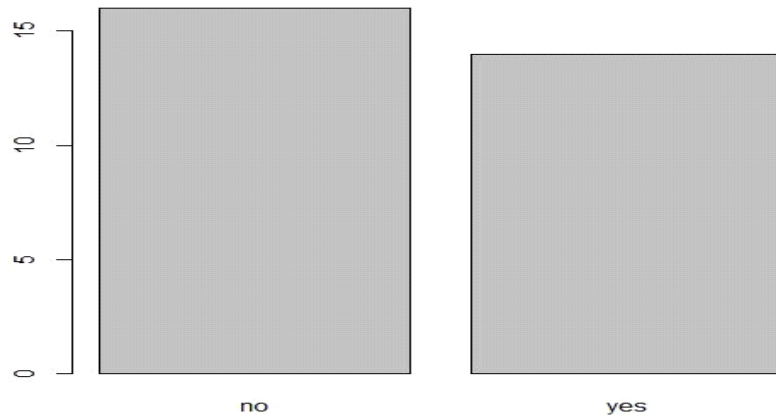


Fig. 6 The graph between actual values and predicted values in Naive Bayes classification

```

Console - 1
>> library(ggplot2, lib.loc = ~/K/WH-library/3.2)
> library(caret)
Error: could not find function "library"
> confusionMatrix(data$Palipitation, sample(data$Palipitation))
Confusion Matrix and Statistics

      Reference
Prediction no yes
no       1     1
yes      1    12

      Accuracy : 0.8667
      95% CI   : (0.5954, 0.9834)
      No Information Rate : 0.8667
      P-Value [Acc > NIR] : 0.6771

      Kappa : 0.4231
      Mcnemar's Test P-Value : 1.0000

      Sensitivity : 0.50000
      Specificity : 0.92308
      Pos Pred Value : 0.50000
      Neg Pred Value : 0.92308
      Prevalence : 0.13333
      Detection Rate : 0.06667
      Detection Prevalence : 0.13333
      Balanced Accuracy : 0.71154

      'Positive' Class : no

> fit <- lda(Palipitation ~ ., data = siva)
Error in is.data.frame(data) : object 'siva' not found
> fit <- lda(Palipitation ~ ., data = data)
Warning message:
In lda.default(x, grouping, ...) : variables are collinear
>> model <- predict(fit)$class
>> baba <- table(model, data$Palipitation)
>> sensitivity(model, data$Palipitation)
[1] 0
>> sensitivity(baba, "Fastbeat")
    
```

Fig. 7 Sensitivity code

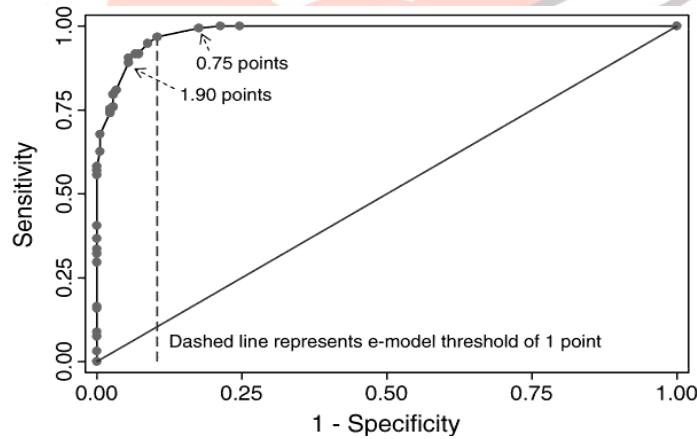


Fig. 8 The graph between Sensitivity and Specificity in Confusion matrix

III. Conclusion

In this work, we study the big data solution for predicting the 30-day risk of readmission for the CHF patients. Our proposed solution leverages big data infrastructure for both information extraction and predictive modelling. We study the effectiveness of our proposed solution with a comprehensive set of experiment, considering quality and scalability. As ongoing work, we aim at leveraging big data infrastructure for our designed risk calculation tool, for designing more sophisticated predictive modelling and feature extraction techniques, and extending our proposed solutions to predict other clinical risks.

References

[1] Krumholz H. M., Normand S. L. T., Keenan P. S., Lin Z. Q., Drye E.E., Bhat K. R., Wang Y. F., Ross J. S., Schuur J. D., and Stauer B. D..Hospital 30-day heart failure readmission measure methodology. Report prepared for the Centres for Medicare & Medicaid Services.

[2] Amarasingham R, Moore BJ, Tabak YP, Drazner MH, Clark CA, Zhang S, Reed WG, Swanson TS, Ma Y, Halm EA. An automated model to identify heart failure patients at risk for 30-day readmission

or death using electronic medical record data. *Journal of Medical Care*, 10:981-988, Feb. 2010.

[3] An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data.

[4] Koelling T. M., Johnson M. L., Cody R. J., and Aaronson. K. Discharge education improves clinical outcomes in patients with chronic heart failure. *Circulation*, 111(2):179-185, Jan. 2005.

[5] Impact of prior admissions on 30-day readmissions in medicare heart failure inpatients.

[6] Meadam N., Verbiest N., Zolfaghar K., Agarwal J., Chin S., Basu Roy S., Teredesai A., Hazel D., Reed L., Amoroso P. Exploring Pre-processing Techniques for Prediction of Risk of Readmission for Congestive Heart Failure Patients. In *Data Mining and Healthcare Workshop*, in conjunction with the 19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2013

