Predictive Analytics

ANALYTICUS

Oct - Dec (2016)

Contents

- Data Structure in R
- Logistic regression
- SPSS 24 : New Features
- Time Series
- News and Events

DATA STRUCTURES IN R

R is a programming language which is widely used among statisticians and data miner for data analysis. It provides a wide range of statistical and graphical techniques. Now all these analyses not only involve data but also require the data to be in correct format. So it is necessary to understand the data structures in R. Data structure helps us to store different types of data in various formats.

The basic important data structures in R are-

Vector – Vectors are the collection of values of data type stored in a single entity. Vectors are of three types-

i) Numeric vector (contains numerical values)
ii) Character Vector (Contains character names)
iii) Logical vector (elements are of logical type say TRUE FALSE or NA)

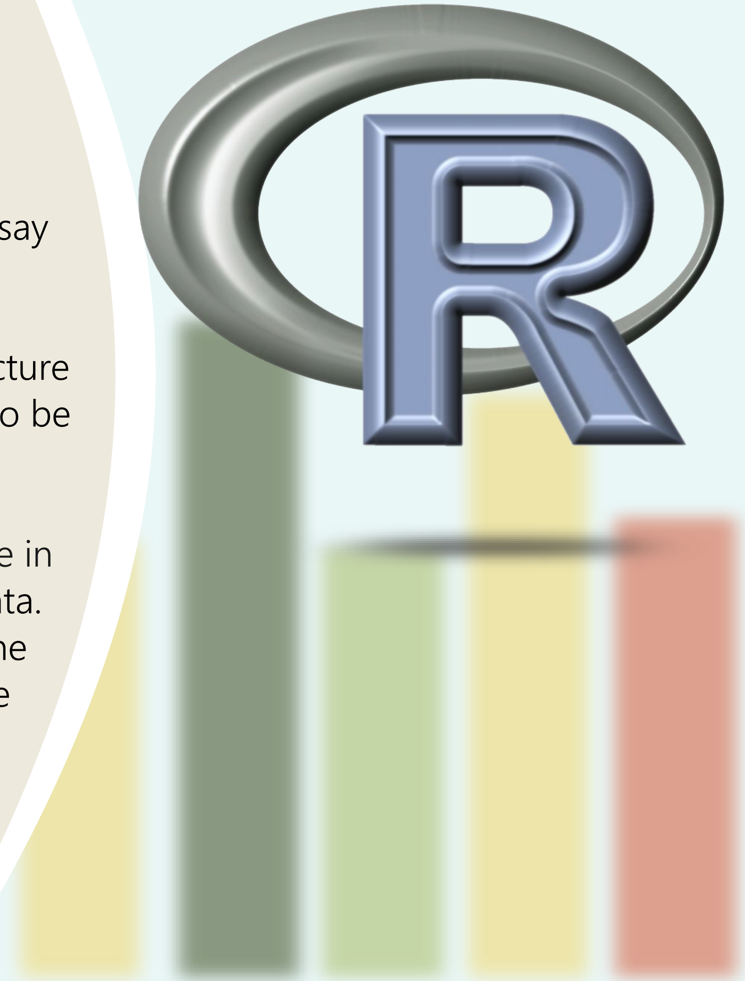
Matrix – It is another basic important data structure in R used to store two dimensional data. It can also be defined as "An array with two subscripts."

Arrays – Arrays are another form of data structure in R which is used to store multi-dimensional data. Arrays are treated as **vectors** when they are one dimensional and treated as **matrix** when they are two-dimensional. In array the product of the dimension should be exactly equal to the length of the object.

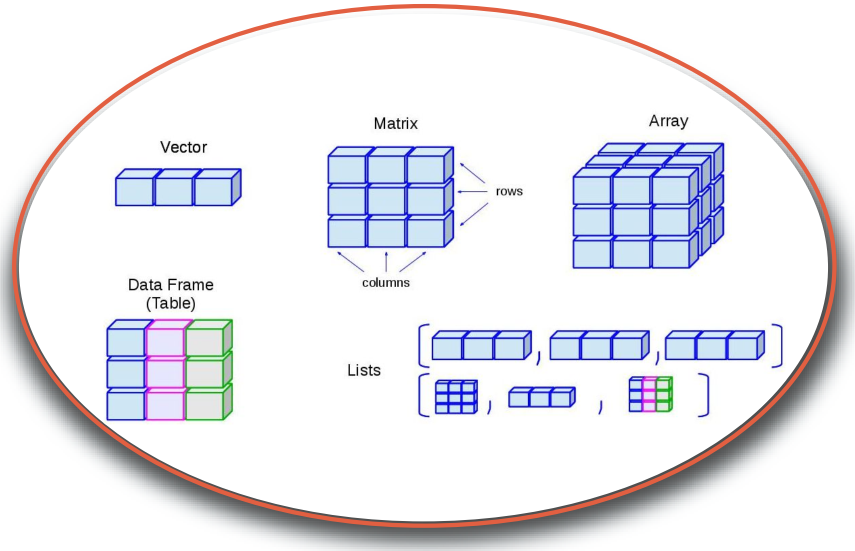
Lists – It is an object consisting of ordered collection of other components which may be numeric, vector, logical values, character array etc.

Factor – It is another basic important data structure in R used to store the categorical data. Unlike vectors, factor contains information about levels present in them.
Factors are also of two types-
i) Ordered factor
ii) Unordered factor

Data frame – It is also an important data structure used to store data tables. It is a list of vectors of equal length and should have unique row names.



Following is a pictorial diagram to better understand the data structures-



The above data structures can also be organised by their **dimensions** and **type**
(i) Homogeneous (all contents are of same type)
(ii) Heterogeneous (all contents are of different type)

The following table helps to understand R data types and structures-

	Homogeneous	Heterogeneous
One dimensional	Vector	Lists
Two dimensional	Matrix	Data frame
Multidimensional	Array	

In R individual numbers are treated as vector of length 1.
After choosing proper data structure, it is easy to proceed for further analysis in R.

LOGISTIC REGRESSION

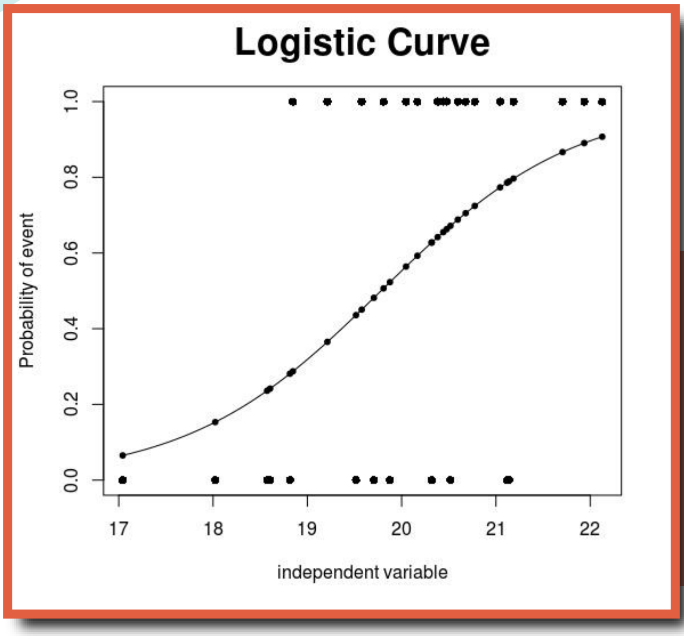
In many situations we are interested to predict an outcome which takes categorical values (e.g-Success/Failure, Good/Better/Best) on the basis of one or more independent variables. In these situations it's appropriate to use logistic regression where the dependent variable is binary (having two outcomes e.g- Good/Bad, Survived/Not). The main goal of using logistic regression is to predict a particular category for one or more independent variables which may or may not be continuous. It is widely used in different field of study such as medical and social sciences,engineering,economics etc.

Some examples where Logistic regression is used-

- Prediction of survival of passengers in a train accident based on features such as class of service,age,sex etc.
- Prediction of probability of failure of a process/ system.
- Prediction of a political candidate to win or lose an election based on the amount of money and time spent during the campaign.
- Prediction of a student to pass or fail in an exam based on the number of hours of study per day.
- Prediction of the survival of a patient after a heart attack.

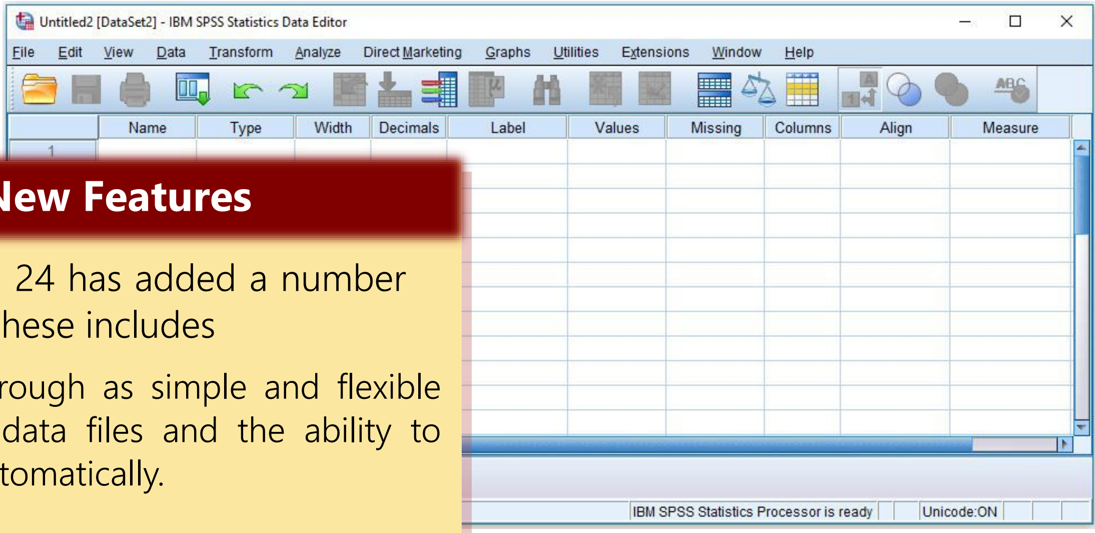
It is clear from the above examples that the outcome variable is of binary nature. In Logistic regression the binary variables are coded as 1 and 0 especially 1 for happening of the event.

In Logistic regression we calculate the probability of the event. Given the values of the independent variables if the outcome of the model is greater than 0.5 (i.e the probability) it is concluded that the event will occur, otherwise it will not.



Logistic curve

The curve shows the probability of happening of the event. If it is greater than **0.5**, then we consider that the event will occur, otherwise it will not.



SPSS 24 : New Features

Newly released version 24 has added a number of improved features. These includes

- Better data import through as simple and flexible interface for reading data files and the ability to detect data formats automatically.
- The ability to extend and customize SPSS functionality with the help of custom dialog builder, newer control options and added properties of existing controls.
- Advanced reporting with rich visual tables and the ability to add advanced output such as significance test results, weighted results and column proportion test to tables.

TIME SERIES

In todays world, predictions are becoming extremely important in the fields of business markets, medicine, economics, politics etc. Everyone would like to know their chance of success before making an investment. In these scenarios time series forcasting helps one to analyse the past occurances to understand the behavior of the data and to predict the future.

There are no models which can forecast the future exactly . However we can reduce the error in our forecasted value by using a range of different techniques with the data and choosing the best model.

For example, we cannot predict the future lucky draw number accurately because we don't know the factors which contributes to the draw of any fixed number, on the other hand we can predict the future electricity bill with more accuracy as we know that electricity consumption is dependent on temperature. The past data of electricity bills are available with the values which can help forcast the future ones . Time series data should be recorded in equal time intervals. It could be recorded hourly, daily, weekly, monthly or yearly. Time series cannot predict random events like terrorist attacks, earthquakes etc.

There are various methods of prediction to be used with different types of data. We will discuss some of them which helps one to forecast.

Time Series Decompositions: Time series data can reveal different types of patterns which gives a clear picture to understand the time series better.

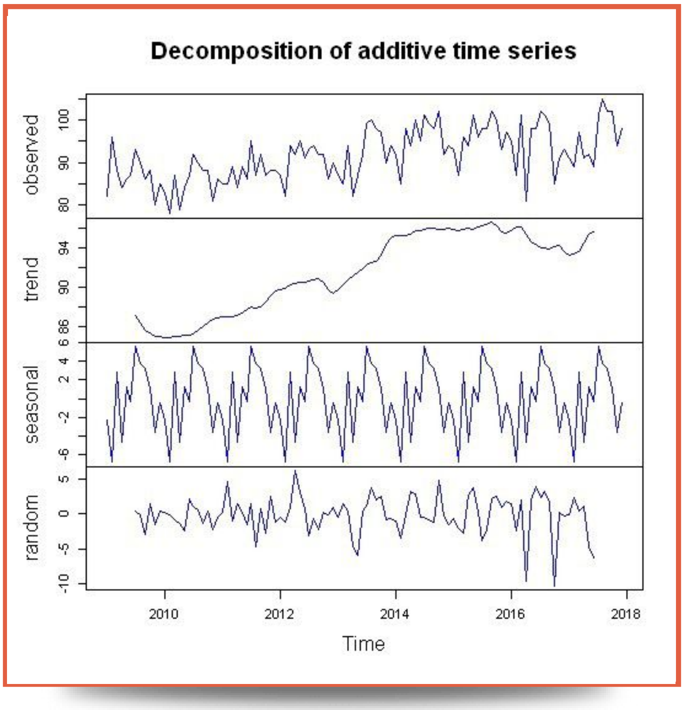
To get the patterns sometimes we need to split the time series data into several components which will be helpful to improve forecasts. These components are Trend, Seasonal and Cyclical.

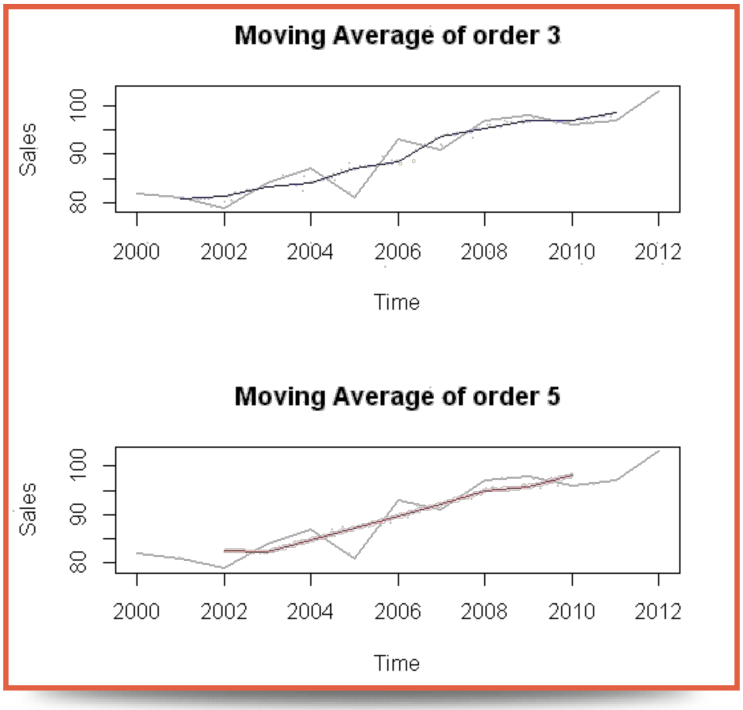
When a pattern exhibits a long term increase or decrease in data, it is called a **Trend.**

When a series is influenced by seasonal factors (e.g. weekly, monthly or quarterly), the data exhibits a **seasonal pattern.** Seasonality is always of a fixed and known period.

When a data shows same rises and falls at least after 2 years like a cycle, the pattern is called **Cyclic**.

Moving Average (MA) : When a time series data has a long term trend or cycle, we use moving average method. The first element of the MA method is calculated by taking the average of the initial fixed subset of the series and the next element will be the average of the second new subset of numbers, excluding the first number and including the next number. This averaging process is repeated over the entire data series. The local averages can smooth out the short term fluctuation for forecasting the long term trends. Thus it is a type of smoothing method.



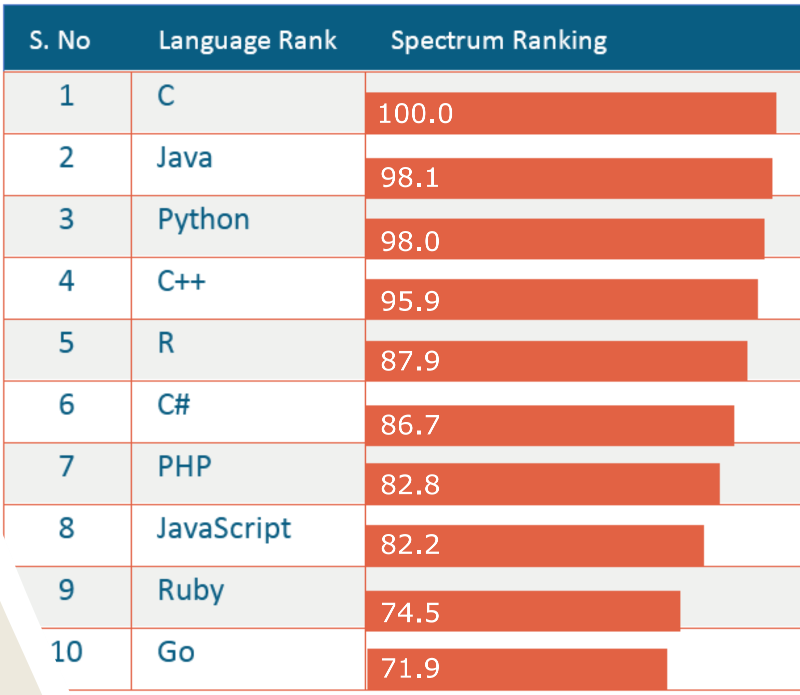


NEWS AND EVENTS

IEEE ranks R language at #5

IEEE Spectrum has just published its third annual ranking with its 2016 Top Programming Languages, and the R Language is once again near the top of the list, moving up one place to fifth position.

R remains the leading tool, with 49% share.



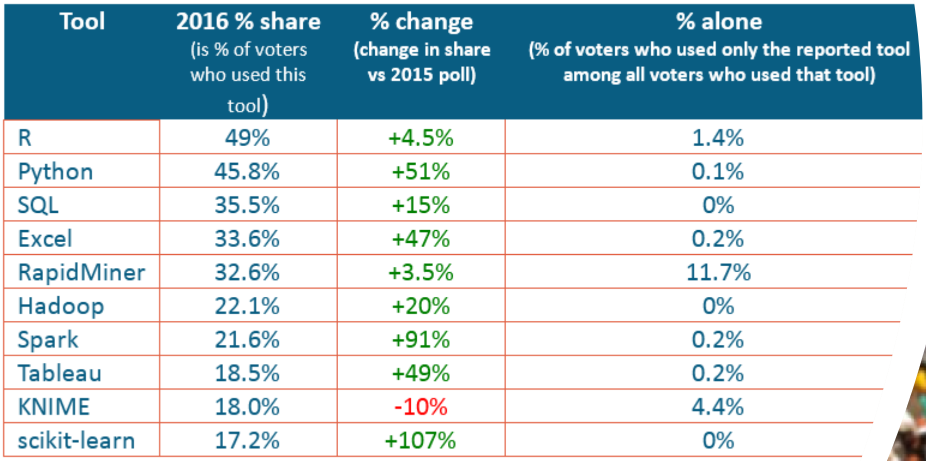
S. No	Language Rank	Spectrum Ranking
1	C	100.0
2	Java	98.1
3	Python	98.0
4	C++	95.9
5	R	87.9
6	C#	86.7
7	PHP	82.8
8	JavaScript	82.2
9	Ruby	74.5
10	Go	71.9

*source-www.r-bloggers.com

IBM Joins R Consortium to Advance the R Programming Language.

In a move to advance data science in the enterprise, IBM has joined the R Consortium to better support the R programming language.

The R Consortium, an open-source foundation to support the R programming language and its user community, today announced that IBM has joined the organization as a Platinum member, the highest level of membership available.



Tool	2016 % share (is % of voters who used this tool)	% change (change in share vs 2015 poll)	% alone (% of voters who used only the reported tool among all voters who used that tool)
R	49%	+4.5%	1.4%
Python	45.8%	+51%	0.1%
SQL	35.5%	+15%	0%
Excel	33.6%	+47%	0.2%
RapidMiner	32.6%	+3.5%	11.7%
Hadoop	22.1%	+20%	0%
Spark	21.6%	+91%	0.2%
Tableau	18.5%	+49%	0.2%
KNIME	18.0%	-10%	4.4%
scikit-learn	17.2%	+107%	0%

*source-www.kdnuggets.com

Events

A number of events on Analytics were conducted over the last few months in some of India's reputed academic and research institutes. Addressing both the student and teaching fraternity, these events were comprised of classroom instructions and hands on sessions followed by access to our web based self-learning portal **Lets Learn Analytics**. Here is what a few senior academicians have to say about them.

AIMA New Delhi

Workshop on research methodology using SPSS and R

Institute of Chemical Technology, Mumbai

Workshop on Analytics with R

Events

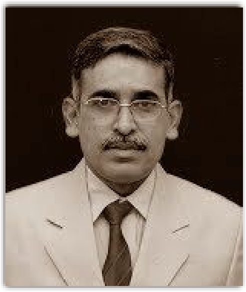
NIT Warangal

FDP on Data science & Big Data analytics

IIITDM Jabalpur

Advanced Data Mining Algorithms & their scalability

"Hands - on experience on R Programming by the instructor is the major strength of the workshop."



Prof D V L N Somayajulu

Chair
E & ICT Academy,
NIT Warangal

"R-trainings provided by Predictive Analytics to our students at ICT Mumbai were very useful and highly appreciated by the participants. "

Dr. Ajit Kumar
Assistant Professor,
Mathematics

Institute of Chemical Technology
Mumbai

"Self paced course on R is quite interactive and help in faster learning for beginners."

Anuja Pandey, AIMA New Delhi