Useful R Packages that Aligns with The CRISP DM Methodology

Posted by Sunil Kappal on February 6, 2017 at 8:00am View Blog

2-2 minutes

As we all know CRISP DM stands for Cross Industry Standard Process for Data Mining is a process model that outlines the most common approach to tackle data driven problems. Per the poll conducted by KDNuggets in 2014 this was and “is” one of the most popular and widest used methodology. This method of gleaning insights out of the data is very dear to the industry experts and data miners.

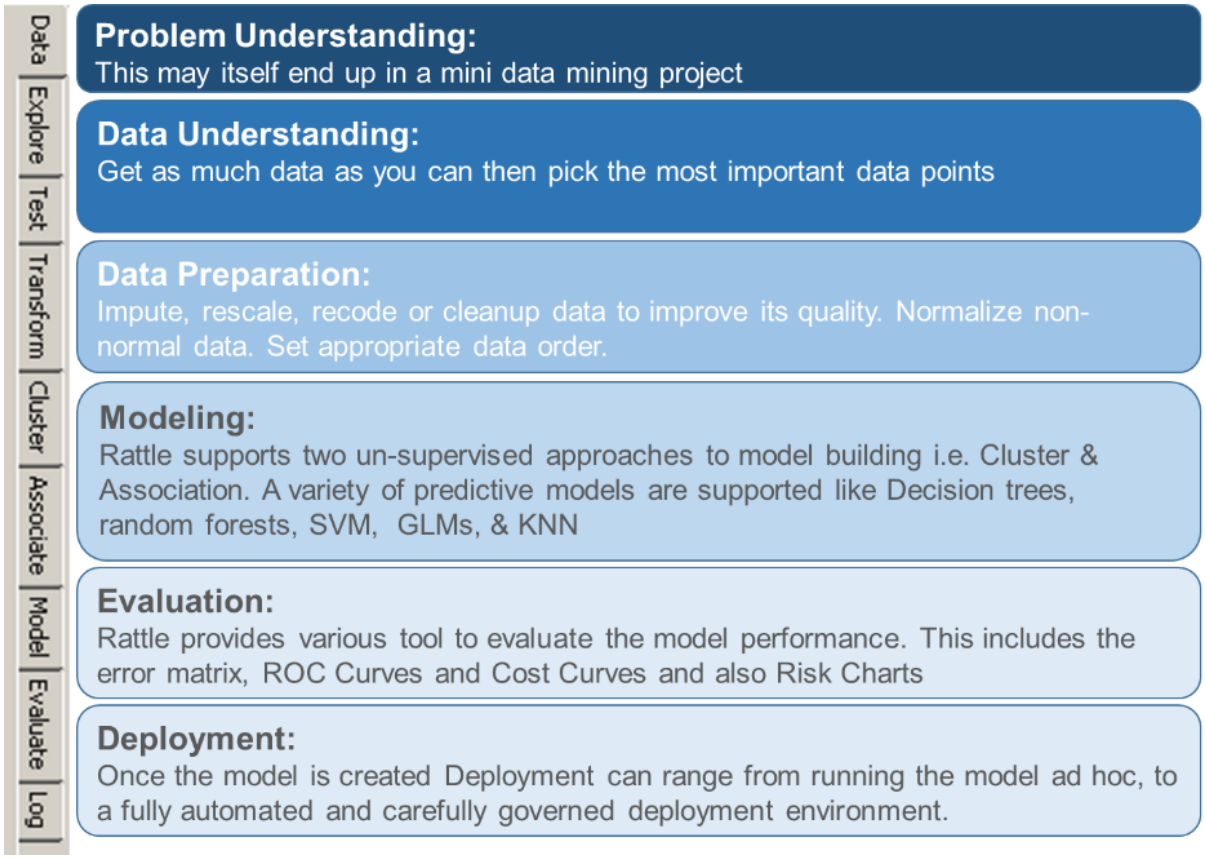
As the title suggest I will align some of the most useful R packages with this most popular and simplistic data processing model and before getting into specific packages, there is one GUI based R Package named Rattle which is very much based around these steps.

Rattle’s tab based interface provides a step by step flow which is a replica of the CRISP DM method. Rattle by itself may suffice the user needs for introducing data mining. However, it also provides stepping stone to a more robust and sophisticated data processing and modelling.

Let’s look at the typical data mining process per the CRISP DM relative to the Rattle Package:

1. Problem Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment

Below is the Rattle Tab that is setup based on the CRISP DM Method of data mining.



Now let's look at some standalone R packages based on the CRISP DM data processing methodology.

CRISP DM Main Activities	PACKAGE NAME & FUNCTIONALITY
Load Data	Use RODBC, RMySQL, RPostgresSQL, RSQLite packages to load data from various databases. To load data from Excel files use XLConnect, xlsx packages and to load data from SAS or an SPSS data set use Foreign
Explore Data	**ggplot2** - R's famous package for making beautiful graphics. ggplot2 lets you use the grammar of graphics to build layered, customizable plots. **ggvis** - Interactive, web based graphics built with the grammar of graphics. **rgl** - Interactive 3D visualizations with R **htmlwidgets** - A fast way to build interactive (javascript based) visualizations with R. Packages that implement htmlwidgets include: **leaflet** (maps) **dygraphs** (time series) **DT** (tables) **diagrammeR** (diagrams) **network3D** (network graphs) **threeJS** (3D scatterplots and globes)
Manipulate Data	**dplyr** – This package has all the essential shortcuts for manipulating, summarizing and joining data sets. Dplyr package provides fast data manipulation. **tidyr** – Change the data format and tidy it up! **stringr** – Tools for regular expressions and character strings **lubridate** – Helps to work with dates and times
Model Data	**car** - car's Anova function is popular for making type II and type III Anova tables. **mgcv** - Generalized Additive Models **lme4/nlme** - Linear and Non-linear mixed effects models **randomForest** - Random forest methods from machine learning **multcomp** - Tools for multiple comparison testing **vcd** - Visualization tools and tests for categorical data **glmnet** - Lasso and elastic-net regression methods with cross validation **survival** - Tools for survival analysis **caret** - Tools for training regression and classification models
Evaluate Data	**modEva** – it is an R package for analyzing and evaluating GLMs with bionomial distribution and a logit link function **ROCR** – is the one of the best R packages to generate sensitivity/specificity curves, lift charts. ROCR is easy to use, with only three commands and reasonable default values for all optional parameters.

Happy Data Munging to All !!!