REVOLUTION ANALYTICS

Modeling in Revolution R Enterprise

Module 1: Overview and Review

July 23, 2014

Overview

After completing this course, you will be able to:

- Conduct predictive analysis on your enterprise data using regression and tree-based models.
- Implement models through embedded scoring functions in Revolution R Enterprise.
- Use advanced algorithms for unsupervised learning and data manipulation such as principal components and clustering techniques.
- Understand key concepts in coding big data functions efficiently.

Algorithm and Function Overview

The basic methods we will cover in this course are listed as follows:

- Linear Regression Modeling and Evaluation
  - Simple Regression
  - Multivariate Regression
  - Complex Formulas and Higher Order Terms
  - Stepwise Regression
- Generalized Linear Models
  - Logistic Regression
  - Additional Forms for General Linearized Models
- Data Mining using Trees and Forests
  - Tree Modeling
  - Random Forest

Algorithm and Function Overview

- Unsupervised Model and Other Techniques
  - Clustering
  - Principal Components
  - Running Simulations

The Data

Throughout this course we will be using the following data set:

- Bank Marketing data set from the Machine Learning Repository at University of California, Irvine

[Moro et al., 2011] S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimaraes, Portugal, October, 2011. EUROSIS (http://hdl.handle.net/1822/14838)

The Data: Bank Marketing Data

The Bank Marketing Data Set, which we will refer to as the Bank data, concerns the direct marketing campaigns, or phone calls, of a Portuguese banking institution to its clientele, and the success of those campaigns in causing customers to subscribe to a term deposit. Multiple types of data are collected on each bank client, such as information on one's age and marital status.

Review: Creating a Data Source

We can create the data source using one of the the following commands:

```
# infile <- file.path('data', 'bank-full.csv') BankDS <- RxTextData(file =
# infile)

infile <- file.path("data", "BankXDF.xdf")
BankDS <- RxXdfData(file = infile)
```

Review: Creating a Data Source

We can get basic information about the data using the the following commands:

```
rxGetInfo(BankDS, getVarInfo = TRUE, numRows = 6)
```

```
## File name: D:\Github\Legacy_Course_Materials\modules\Modeling\Overview_and_Review\doc\data\BankXDF.xdf
## Number of observations: 45211
## Number of variables: 21
## Number of blocks: 5
## Compression type: zlib
## Variable information:
...
```

```
rxSummary(~., BankDS)
```

```
## Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.007 seconds
## Rows Read: 10000, Total Rows Processed: 20000, Total Chunk Time: 0.008 seconds
## Rows Read: 10000, Total Rows Processed: 30000, Total Chunk Time: 0.008 seconds
## Rows Read: 10000, Total Rows Processed: 40000, Total Chunk Time: 0.008 seconds
## Rows Read: 5211, Total Rows Processed: 45211, Total Chunk Time: 0.006 seconds
## Computation time: 0.059 seconds.
```

```
## Call:
## rxSummary(formula = ~., data = BankDS)
##
## Summary Statistics Results for: ~.
## Data: BankDS (RxXdfData Data Source)
## File name: data/BankXDF.xdf
...
```

Recap

Let’s review some of the concepts covered in this module:

- How do you create a data source?

Thank you

Revolution Analytics is the leading commercial provider of software and support for the popular open source R statistics language.

www.revolutionanalytics.com, 1.855.GET.REVO, Twitter: @RevolutionR