RemixAutoML Library Introduction

Adrian Antico

2019-03-31

Contact Info

LinkedIn: https://www.linkedin.com/in/adrian-antico/

Remix Instute: https://www.remyxcourses.com or https://www.remixinstitute.ai

Vignette Intent

This vignette is designed to give you the highlights of the set of automated machine learning functions available in the RemixAutoML package. To see the functions in action, visit the Remyx Courses website for the free course at **https://www.remyxcourses.com** and walk through them (and check our the other courses too!).

Package Goals

The **RemixAutoML** package (*Remix Automated Machine Learning*) is designed to automate and optimize the quality of machine learning, the pace of development, along with the handling of big data and the processing time of data management. The library has been a development task at Remix Institute over the course of the past year to consolidate all of our winning methods for successfully completing machine learning and data science consulting projects. These were actual projects at Fortune 500 companies, Fortune 100 companies, tech startups, and other consulting clients. We are avid R users and feel that the R community could benefit from its release.

Package Design Philosophy

The two core packages RemixAutoML relies on are H2O and data.table. There are other packages used, for example, the forecast package, but H2O and data.table are used the most. I use data.table for data wrangling of all internal functions due to its ability to handle big data with minimal memory and the speed at which their functions process data. I chose to use H2O and their machine learning algorithms because of their high quality results, flexibility of use, ease of operationalization, and ability to manage big data. I use these functions routinely for machine learning projects and they continue to outperform every other package / software I test them against. Many of the other R packages for modeling or data manipulation have terrible run times and fail once I get going with bigger data. As a simple example, sometimes I do a one-hot encoding for testing out keras models and I can just run my DummifyDT function to create those. I've tried out a few others with fast runtimes but they fail on bigger data, mess up the column ordering, and don't offer the flexibility of creating one-hot encoding features versus standard dichotomized features.

Install H20

Follow this link to install H20 if it isn't on your machine already. Install H20

There are seven categories of functions (currently) in this library I'll go over:

- Automated Supervised Learning
- Automated Unsupervised Learning
- Automated Model Evaluation
- Automated Feature Interpretation
- Automated Cost Sensitive Optimization
- Automated Feature Engineering
- A Few Miscellaneous Functions

Automated Supervised Learning Functions

Functions include:

- `AutoH20Modeler()`
- `AutoH20Scoring()`
- `AutoTS()`
- `AutoRecommender()`
- `AutoRecommenderScoring()`
- `AutoNLS()`

AutoH20Modeler()

The supervised learning functions handle multiple tasks internally. The **AutoH20Modeler** function can build any number of H2O models, automatically compare hyper-parameter tuned versions to baseline versions, selecting a winner, saving the model evaluation and feature interpretation metrics / graphs, along with storing models and their metadata to refer to them later in a production setting.

The models available include:

- Gradient Boosting Machines
- LightGBM (Linux only)
- Distributed Random Forest
- XGBoost (Linux only)
- Deeplearning
- AutoML (for Windows users XGBoost and LightGBM are not available)

For Windows users (Mac?), XGBoost is not available and therefore neither is LightGBM (XGBoost and LightGBM are not utlized in AutoML model selection with Windows).

AutoH20Scoring()

This function is the complement of the **AutoH20Modeler**, **AutoKMeans**, and **AutoWord2VecModeler** functions. Specify which rows of your model metadata collection file to run and **AutoH20Scoring** will return a list of predicted values, where each element of the list is a set of predicted values from the model it ran. For the **AutoH20Modeler** you will generate a file called grid_tuned_paths.Rdata which contains the path to your models (among other items) that you can pass along to the **AutoH20Scoring** function to automatically score your models. For the **AutoKMeans** you will generate a file called KMeansModelFile.Rdata which contains the paths to the models for scoring your GLRM and KMeans models. For the **AutoWord2VecModeler** you will generate a file called StoreFile.Rdata which contains the paths to your word2vec models for scoring.

In total, the AutoH20Scoring function can score

- Regression models
- Quantile regression Models
- Binary classification Models
- Multinomial classification Models
- Multioutcome multinomial classification models
- Generalized low rank dimensionality reduction models
- KMeans clustering models
- Word2vec models

AutoTS()

Another automated supervised learning function we have is an automated time series modeling function (AutoTS) that optimally builds out seven types of time series forecasting models, compares them on holdout data, picks a winner, rebuilds the winner on full data, and generates the forecasts for the number of desired periods. The intent is to make these processes fast, easy, and of high quality. Every model makes use of the optimal settings of their paramters to give them the best chance of being the best. Each model uses a Box-Cox transformation on the target variable and all predictions are back-transformed.

The competing models include:

- ARFIMA (Autoregressive Fractional Integrated Moving Average)
- ARIMA (Autoregressive Integrated Moving Average)
- ETS (Exponential Smoothing and Holt Winters)
- TBATS (Exonential Smoothing State Space Model with Box-Cox Transformation, ARMA Errors, Trend and Seasonal Components)
- TSLM (Time Series Linear Model)
- NN (Autoregressive Neural Network)
- Facebook Prophet

AutoRecommender()

This function builds out several variations of collaborative filtering models on a binary ratings matrix. To automatically build the binary ratings matrix, see **RecomDataCreate**.

The competing models include:

- RandomItems
- PopularItems
- UserBasedCF
- ItemBasedCF
- AssociationRules

AutoRecommenderScoring()

This function will automatically score your winning model. Simply feed in your data and the winning model returned from the **AutoRecommender** function and this function will generate a table of several recommended products (by rank) for each entity.

AutoNLS()

This automated supervised learning function builds nonlinear regression models for a more niche set of tasks. It's set up to generate interpolation predictions, such as smoothing cost curves for optimization tasks. It returns the interpolated data, the winning model name, the model object, and the evaluation metrics table.

The competing models include:

- Asymptotic
- Asymptotic through origin
- Asymptotic with offset
- Bi-exponential
- Four parameter logistic
- Three parameter logistic
- Gompertz
- Michal Menton
- Weibull
- Polynomial regression or monotonic regression

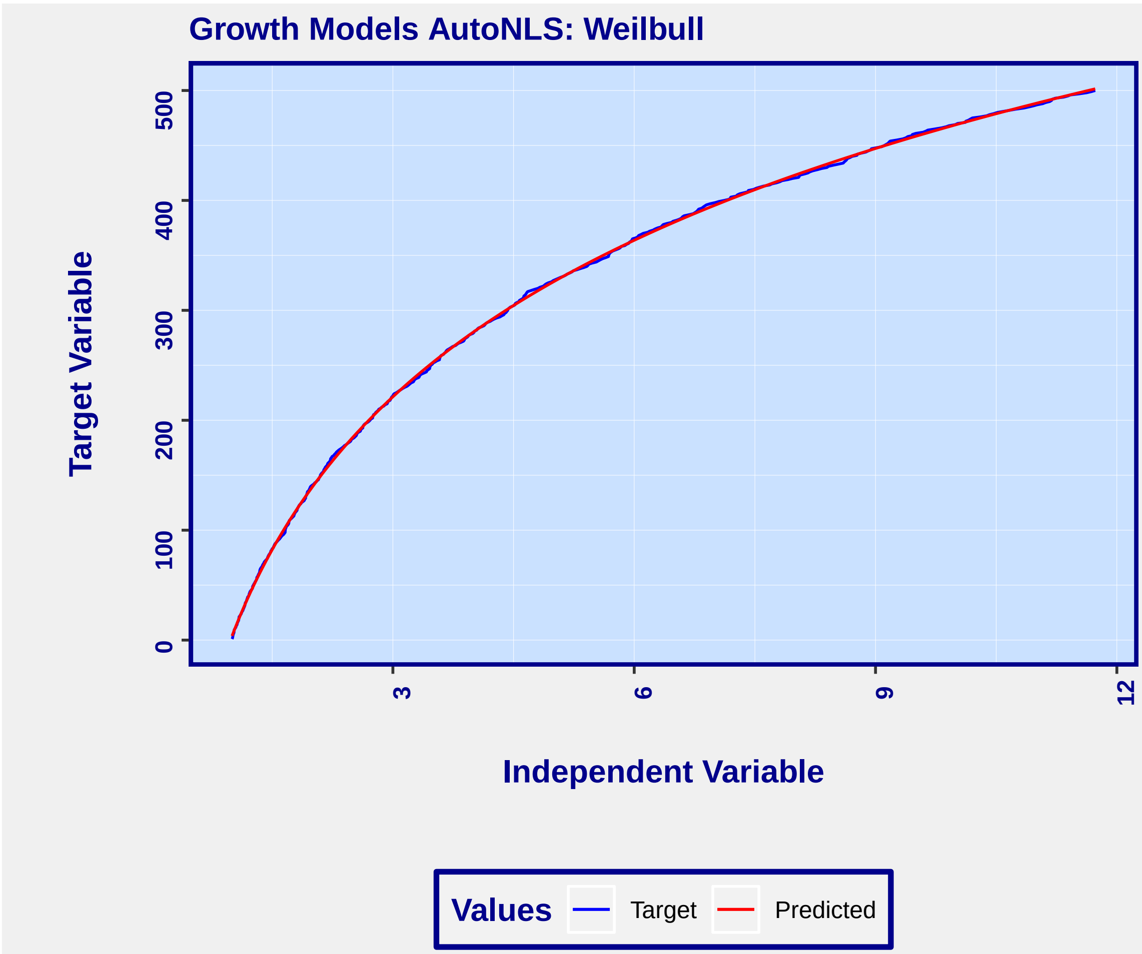
Example of AutoNLS with Plot (simulated data)

Find more demos at **https://www.remyxcourses.com/course?courseid=modeling-tools-library-walkthrough**

```r
library(RemixAutoML)

# Create Growth Data
data <-
    data.table::data.table(Target = seq(1, 500, 1),
                           Variable = rep(1, 500))
  for (i in as.integer(1:500)) {
    if (i == 1) {
      var <- data[i, "Target"][[1]]
      data.table::set(data,
                      i = i,
                      j = 2L,
                      value = var * (1 + runif(1) / 100))
    } else {
      var <- data[i - 1, "Variable"][[1]]
      data.table::set(data,
                      i = i,
                      j = 2L,
                      value = var * (1 + runif(1) / 100))
    }
  }

  # Add jitter to Target
  data[, Target := jitter(Target,
                          factor = 0.25)]

  # To keep original values
  data1 <- data.table::copy(data)

  # Build models
```

```
  data11 <- AutoNLS(
    data = data,
    y = "Target",
    x = "Variable",
    monotonic = TRUE
  )

  # Join predictions to source data
  data2 <- merge(
    data1,
    data11$PredictionData,
    by = "Variable",
    all = FALSE
  )

# Plot output
ggplot2::ggplot(data2, ggplot2::aes(x = Variable)) +
  ggplot2::geom_line(ggplot2::aes(y = data2[["Target.x"]],
                                  color = "Target")) +
  ggplot2::geom_line(ggplot2::aes(y = data2[["Target.y"]],
                                  color = "Predicted")) +
  RemixAutoML::ChartTheme(Size = 12) +
  ggplot2::ggtitle(paste0("Growth Models AutoNLS: ",
                          data11$ModelName)) +
  ggplot2::ylab("Target Variable") +
  ggplot2::xlab("Independent Variable") +
  ggplot2::scale_colour_manual("Values",
                               breaks = c("Target",
                                          "Predicted"),
                               values = c("red",
                                          "blue"))

# Print model makeup and evaluation metrics
summary(data11$ModelObject)
#>
#> Formula: Target ~ SSweibull(Variable, Asym, Drop, lrc, pwr)
#>
#> Parameters:
#>        Estimate Std. Error t value Pr(>|t|)
#> Asym 1254.87640   84.96916   14.77   <2e-16 ***
#> Drop 2415.64331  184.08239   13.12   <2e-16 ***
#> lrc    -0.41877    0.01502  -27.88   <2e-16 ***
#> pwr     0.23216    0.01770   13.12   <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 1.643 on 496 degrees of freedom
#>
#> Number of iterations to convergence: 0
#> Achieved convergence tolerance: 2.479e-07
data11$EvaluationMetrics
#>        ModelName MeanAbsError
#> 1:      Weibull     1.310615
```

```
#> 2:       Gompertz    15.564747
#> 3: Michal_Menton    20.312342
#> 4:       Logistic    21.215046
#> 5:           Poly    80.912907
```



Example of AutoTS with Plot (simulated data)

Find more demos at **https://www.remyxcourses.com/course?courseid=modeling-tools-library-walkthrough**

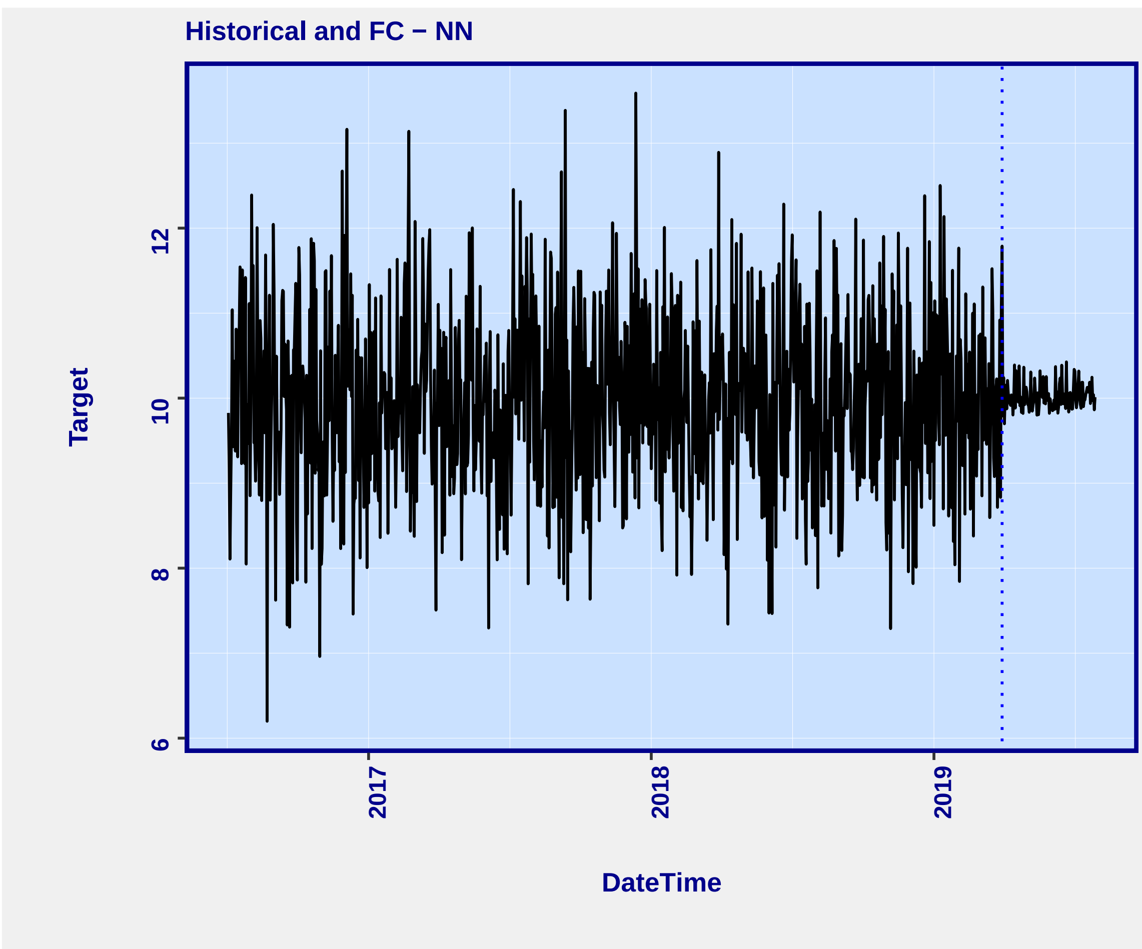
```
library(RemixAutoML)

# Create autocorrelated time series data
data <- data.table::data.table(DateTime = as.Date(Sys.time()),
                               Target = (10 + arima.sim(model = list(2,0,2),
                                                        n = 1000)))
data[, temp := seq(1:1000)][, DateTime := DateTime - temp][, temp := NULL]
data <- data[order(DateTime)]

# Build models and generate forecasts
output <- RemixAutoML::AutoTS(data,
                              TargetName      = "Target",
                              DateName        = "DateTime",
```

```r
                              FCPeriods      = 120,
                              HoldOutPeriods = 30,
                              TimeUnit       = "day",
                              Lags           = 5,
                              SLags          = 1,
                              NumCores       = 4,
                              SkipModels     = NULL,
                              StepWise       = TRUE)
#> [1] "ARFIMA FITTING"
#> [1] "ARIMA FITTING"
#> [1] "ETS FITTING"
#> [1] "TBATS FITTING"
#> [1] "TSLM FITTING"
#> [1] "NNet FITTING"
#> [1] 1
#> [1] 2
#> [1] 3
#> [1] 4
#> [1] 5
#> [1] "PROPHET FITTING"
#> [1] "FIND WINNER"
#> [1] "GENERATE FORECASTS"
#> [1] 1
#> [1] 2
#> [1] 3
#> [1] 4
#> [1] 5

# Collect output and prepare for plotting
data1 <- output$Forecast
maxDate <- data[, max(DateTime)]
data.table::setnames(data1, names(data1), c("DateTime","Target"))
data2 <- data.table::rbindlist(list(data[, Target := as.numeric(Target)],
                                    data1))

# Plot time series data and forecast
ggplot2::ggplot(data2, ggplot2::aes(x = DateTime, y = Target)) +
  ggplot2::geom_line() +
  RemixAutoML::ChartTheme(Size = 10) +
  ggplot2::geom_vline(ggplot2::aes(xintercept = maxDate),
                      linetype = "dotted",
                      color = "blue") +
  ggplot2::ggtitle(paste0("Historical and FC - ",
                          output$EvaluationMetrics[1,1][[1]])) +
  ggplot2::theme(legend.position="none")

# Print the evaluation metric and model makeup
knitr::kable(output$EvaluationMetrics)
```

ModelName	MeanResid	MeanPercError	MAPE	ID
NN	0.0855026	-0.0836938	0.0979466	1
TBATS	0.0533688	-0.0864057	0.0982996	2
ETS	0.0463101	-0.0869940	0.0986655	3

ModelName	MeanResid	MeanPercError	MAPE	ID
ARFIMA	0.0456626	-0.0870479	0.0986990	4
ARIMA	0.0456572	-0.0870484	0.0986993	5

```
summary(output$TimeSeriesModel)
#>             Length Class        Mode
#> x           1000   ts           numeric
#> m              1   -none-       numeric
#> p              1   -none-       numeric
#> P              1   -none-       numeric
#> scalex         2   -none-       list
#> size           1   -none-       numeric
#> subset      1000   -none-       numeric
#> model         20   nnetarmodels list
#> nnetargs       0   -none-       list
#> fitted      1000   ts           numeric
#> residuals 1000   ts           numeric
#> lags           6   -none-       numeric
#> series         1   -none-       character
#> method         1   -none-       character
#> call           4   -none-       call
```



Automated Unsupervised Learning Functions

The suite of functions in this category currently handle optimized row-clustering and anomaly detection. For the row-clustering, we utilize H2O's Generalized Low Rank Model and their KMeans algorithm, with hyper-parameter tuning for both. The function automatically adds the clusters to your data and can save the models for scoring new data with the **AutoH20Scoring** function. We have a few others currently in development and will release those when they are complete. The anomaly detection functions we have currently are for time series applications. We have a control chart methodology version that lets you build upper and lower confidence bounds by up to two grouping variables along with a time series modeling version. The clustering function and the control chart method function update your data set that you feed in with new columns that store the clusterID or anomaly information. The time series function updates your data, supplies you with the final time series model built, and a data.table that only contains anomalies.

Functions include:

- `GenTSAnomVars()`
- `ResidualOutliers()`
- `AutoKMeans()`

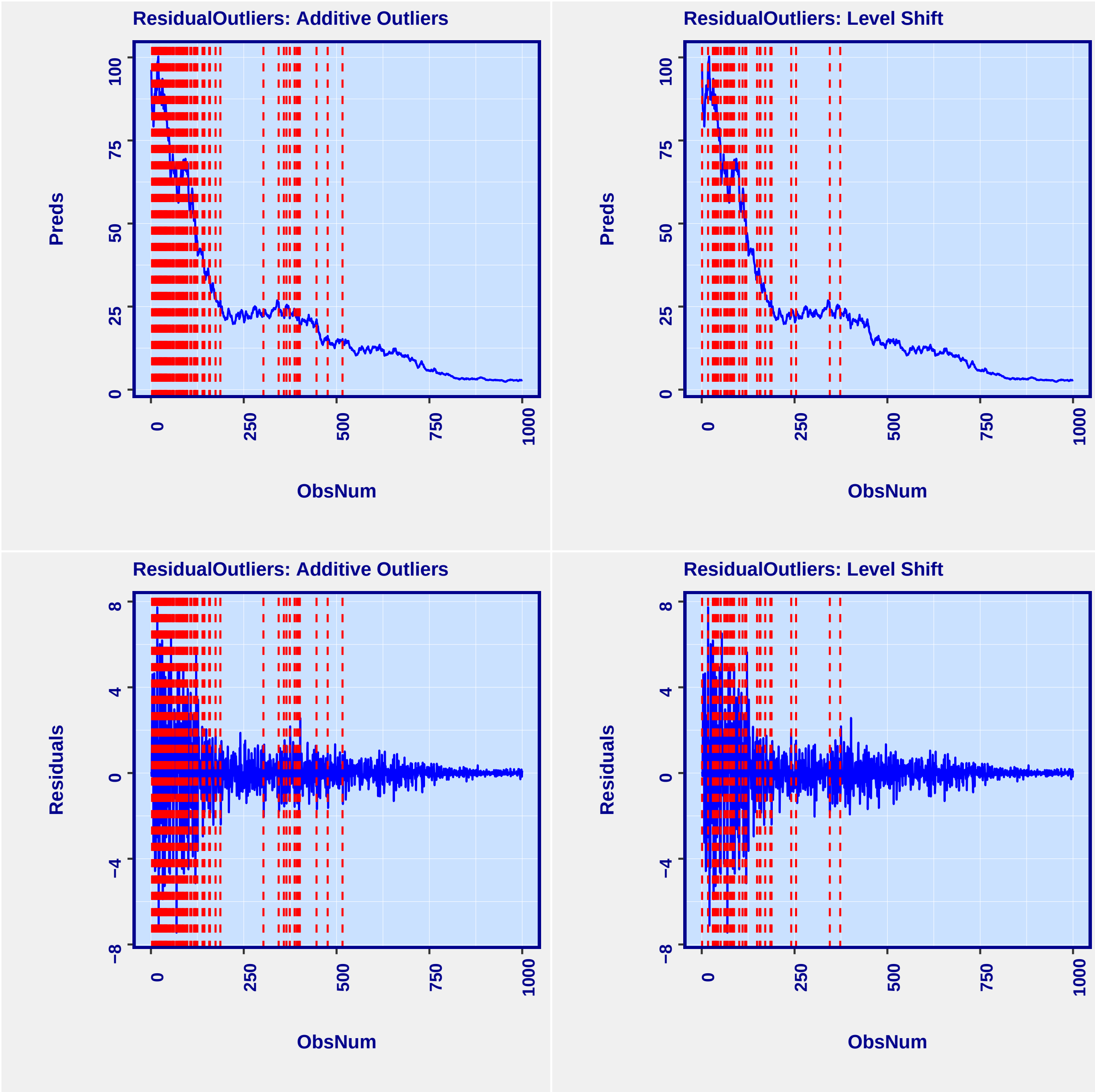
Demo of ResidualOutliers()

Find more demos at **https://www.remyxcourses.com/course?courseid=modeling-tools-library-walkthrough**

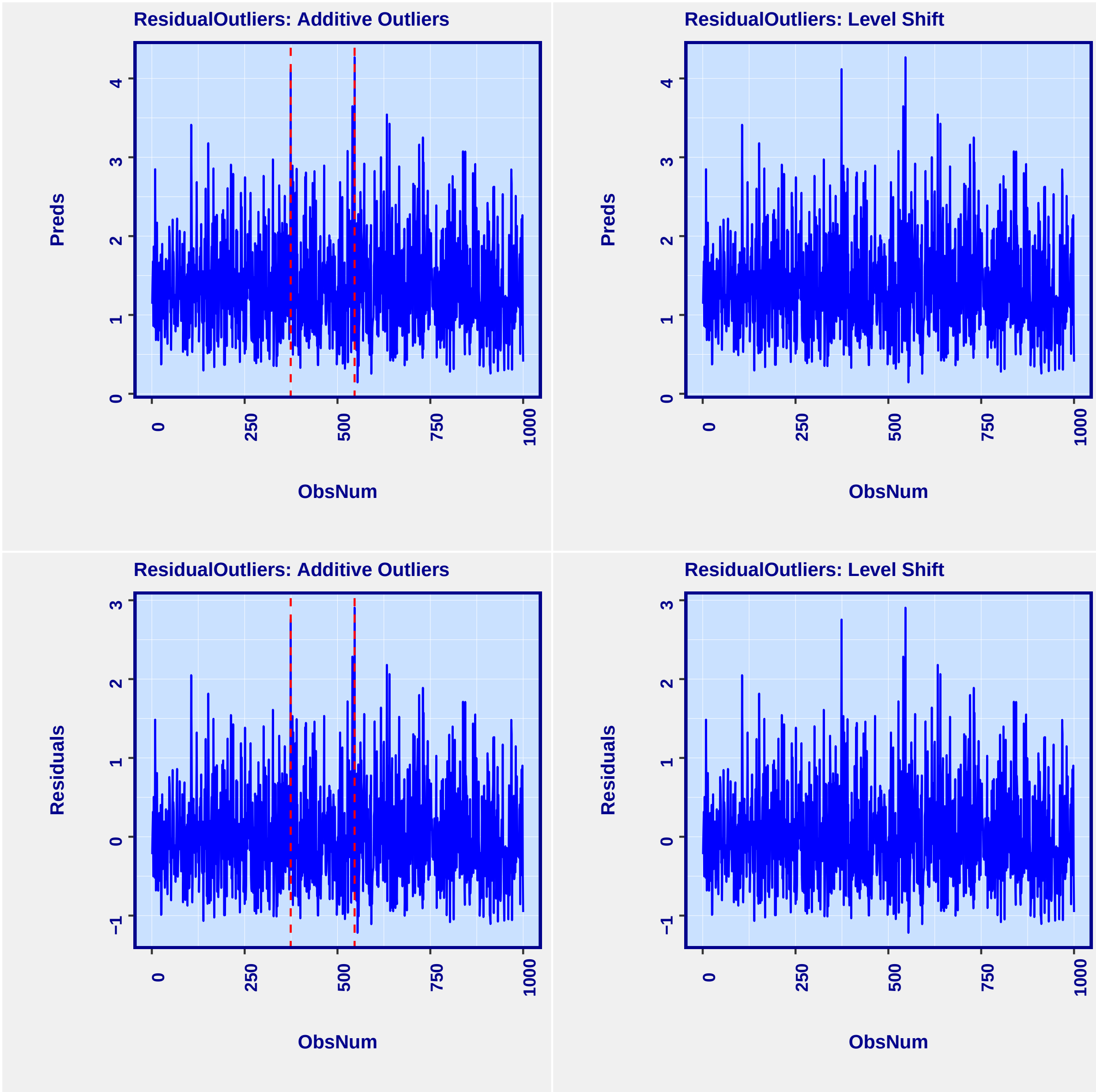
```r
# Create time series data
data <- data.table::data.table(
  a = seq(0,1000,1),
  predicted = sde::GBM(N=1000)*100)[-1,]

# Run function and collect results
stuff    <- ResidualOutliers(data = data, maxN = 5, cvar = 4)
data     <- stuff[[1]]
model    <- stuff[[2]]
resid    <- stuff[[3]]
outliers <- data[type != "<NA>"]

# Create Plots
p1 <- ggplot2::ggplot(data, ggplot2::aes(x = ObsNum)) +
  ggplot2::geom_line(ggplot2::aes(y = Preds),
                     color = "blue") +
  ChartTheme(Size = 10) +
  ggplot2::geom_vline(data = outliers[type == "AO", "ObsNum"],
                      ggplot2::aes(xintercept = outliers[
                        type == "AO"][["ObsNum"]]),
                      linetype = 8, colour = "red") +
  ggplot2::ggtitle("ResidualOutliers: Additive Outliers")

p2 <- ggplot2::ggplot(data, ggplot2::aes(x = ObsNum)) +
  ggplot2::geom_line(ggplot2::aes( y = Residuals),
                     color = "blue") +
  ChartTheme(Size = 10) +
  ggplot2::geom_vline(data = outliers[type == "AO", "ObsNum"],
                      ggplot2::aes(xintercept = outliers[
                        type == "AO"][["ObsNum"]]),
```

```r
                      linetype = 8, colour = "red") +
  ggplot2::ggtitle("ResidualOutliers: Additive Outliers")

p3 <- ggplot2::ggplot(data, ggplot2::aes(x = ObsNum)) +
  ggplot2::geom_line(ggplot2::aes(y = Preds),
                     color = "blue") +
  ChartTheme(Size = 10) +
  ggplot2::geom_vline(data = outliers[type == "LS", "ObsNum"],
                      ggplot2::aes(xintercept = outliers[
                        type == "LS"][["ObsNum"]]),
                      linetype = 8, colour = "red") +
  ggplot2::ggtitle("ResidualOutliers: Level Shift")

p4 <- ggplot2::ggplot(data, ggplot2::aes(x = ObsNum)) +
  ggplot2::geom_line(ggplot2::aes( y = Residuals),
                     color = "blue") +
  ChartTheme(Size = 10) +
  ggplot2::geom_vline(data = outliers[type == "LS", "ObsNum"],
                      ggplot2::aes(xintercept = outliers[
                        type == "LS"][["ObsNum"]]),
                      linetype = 8, colour = "red") +
  ggplot2::ggtitle("ResidualOutliers: Level Shift")

# Print plots
multiplot(plotlist = list(p1,p2,p3,p4), cols = 2)


# Create stable time series data
# Create more stable data and repeat
data <- data.table::data.table(
  a = seq(1,1000,1),
  predicted = sde::rcCIR(n=1000,
                         Dt=0.1,
                         x0=1,
                         theta=c(6,2,2)))

# Run function and collect results
stuff    <- ResidualOutliers(data = data, maxN = 5, cvar = 4)
data     <- stuff[[1]]
model    <- stuff[[2]]
resid    <- stuff[[3]]
outliers <- data[type != "<NA>"]

# Create Plots
p11 <- ggplot2::ggplot(data, ggplot2::aes(x = ObsNum)) +
  ggplot2::geom_line(ggplot2::aes(y = Preds),
                     color = "blue") +
  ChartTheme(Size = 10) +
  ggplot2::geom_vline(data = outliers[type == "AO", "ObsNum"],
                      ggplot2::aes(xintercept = outliers[
                        type == "AO"][["ObsNum"]]),
                      linetype = 8, colour = "red") +
  ggplot2::ggtitle("ResidualOutliers: Additive Outliers")
```

```r
p22 <- ggplot2::ggplot(data, ggplot2::aes(x = ObsNum)) +
  ggplot2::geom_line(ggplot2::aes( y = Residuals),
                     color = "blue") +
  ChartTheme(Size = 10) +
  ggplot2::geom_vline(data = outliers[type == "AO", "ObsNum"],
                      ggplot2::aes(xintercept = outliers[
                        type == "AO"][["ObsNum"]]),
                      linetype = 8, colour = "red") +
  ggplot2::ggtitle("ResidualOutliers: Additive Outliers")

p33 <- ggplot2::ggplot(data, ggplot2::aes(x = ObsNum)) +
  ggplot2::geom_line(ggplot2::aes(y = Preds),
                     color = "blue") +
  ChartTheme(Size = 10) +
  ggplot2::geom_vline(data = outliers[type == "LS", "ObsNum"],
                      ggplot2::aes(xintercept = outliers[
                        type == "LS"][["ObsNum"]]),
                      linetype = 8, colour = "red") +
  ggplot2::ggtitle("ResidualOutliers: Level Shift")

p44 <- ggplot2::ggplot(data, ggplot2::aes(x = ObsNum)) +
  ggplot2::geom_line(ggplot2::aes( y = Residuals),
                     color = "blue") +
  ChartTheme(Size = 10) +
  ggplot2::geom_vline(data = outliers[type == "LS", "ObsNum"],
                      ggplot2::aes(xintercept = outliers[
                        type == "LS"][["ObsNum"]]),
                      linetype = 8, colour = "red") +
  ggplot2::ggtitle("ResidualOutliers: Level Shift")

# Print plots
multiplot(plotlist = list(p11,p22,p33,p44), cols = 2)
```



ResidualOutliers: Additive Outliers
ResidualOutliers: Level Shift
Preds
Residuals
ObsNum
0
25
50
75
100
250
500
750
1000
−8
−4
4
8



Automated Model Evaluation, Feature Interpretation, and Cost Sensitive Optimization Functions

The model evaluation graphs are calibration plots or calibration boxplots. The calibration plots are used for regression (expected value and quantile regession), classification, and multinomial modeling problems. The calibration boxplots are used for regression (expected value and quantile regression). These graphs display both the actual target values and the predicted values, grouped by the number of bins that you specify. The calibration boxplots are useful to understand not only the model bias but also the model variance, across the range of predicted values.

Functions include:

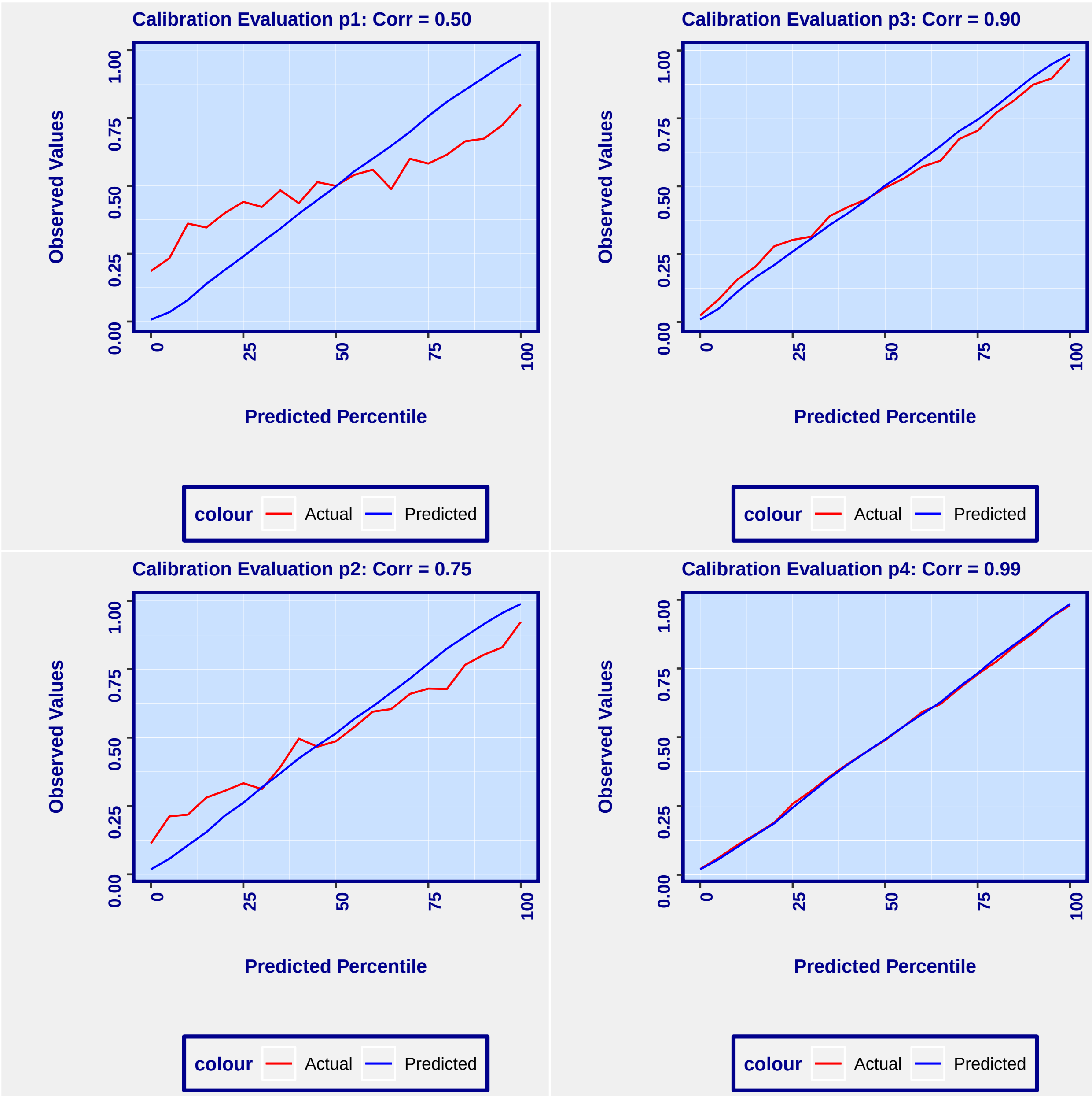
- `EvalPlot()`

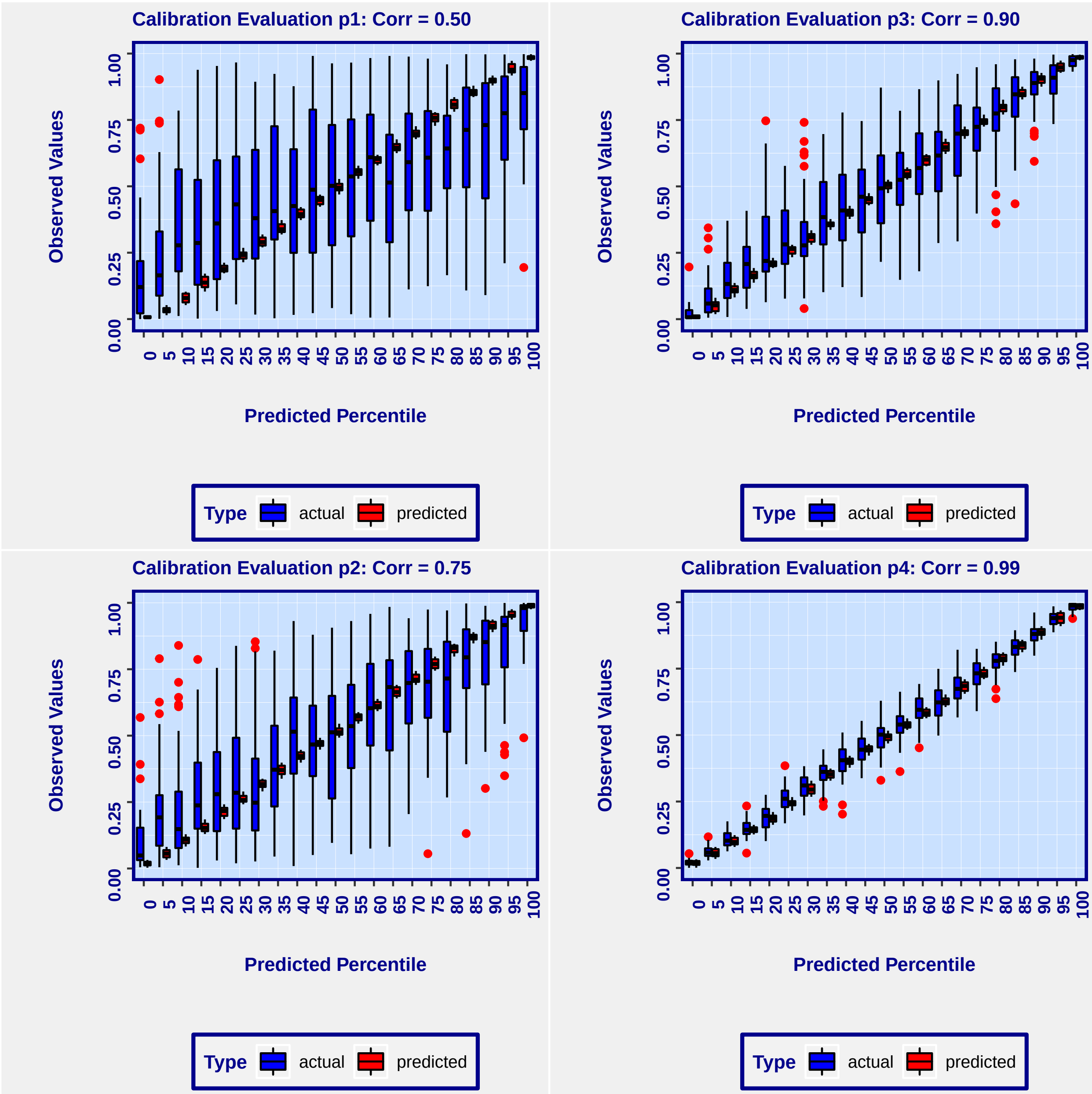
Demo of EvalPlot() for calibration and boxplots

Find more demos at **https://www.remyxcourses.com/course?courseid=modeling-tools-library-walkthrough**

```r
library(RemixAutoML)

# Data generator function
dataGen <- function(Correlation = 0.95) {
  Validation <- data.table::data.table(target = runif(1000))
  Validation[, x1 := qnorm(target)]
  Validation[, x2 := runif(1000)]
  Validation[, predict := pnorm(Correlation * x1 +
                                  sqrt(1 - Correlation ^2) * qnorm(x2))]
  return(Validation)
}

# Store data sets
data1 <- dataGen(Correlation = 0.50)
data2 <- dataGen(Correlation = 0.75)
data3 <- dataGen(Correlation = 0.90)
data4 <- dataGen(Correlation = 0.99)

# Generate EvalPlots (calibration)
p1 <- RemixAutoML::EvalPlot(data = data1,
                            PredColName = "predict",
                            ActColName = "target",
                            type = "calibration",
                            bucket = 0.05,
                            aggrfun = function(x) mean(x,
                                                       na.rm = TRUE))
p1 <- p1 + ggplot2::ggtitle("Calibration Evaluation p1: Corr = 0.50") +
  RemixAutoML::ChartTheme(Size = 10)

p2 <- RemixAutoML::EvalPlot(data = data2,
                            PredColName = "predict",
                            ActColName = "target",
                            type = "calibration",
                            bucket = 0.05,
                            aggrfun = function(x) mean(x,
                                                       na.rm = TRUE))
p2 <- p2 + ggplot2::ggtitle("Calibration Evaluation p2: Corr = 0.75") +
  RemixAutoML::ChartTheme(Size = 10)

p3 <- RemixAutoML::EvalPlot(data = data3,
                            PredColName = "predict",
                            ActColName = "target",
                            type = "calibration",
                            bucket = 0.05,
                            aggrfun = function(x) mean(x,
                                                       na.rm = TRUE))
p3 <- p3 + ggplot2::ggtitle("Calibration Evaluation p3: Corr = 0.90") +
  RemixAutoML::ChartTheme(Size = 10)

p4 <- RemixAutoML::EvalPlot(data = data4,
```

```r
                              PredColName = "predict",
                              ActColName = "target",
                              type = "calibration",
                              bucket = 0.05,
                              aggrfun = function(x) mean(x,
                                                         na.rm = TRUE))
p4 <- p4 + ggplot2::ggtitle("Calibration Evaluation p4: Corr = 0.99") +
  RemixAutoML::ChartTheme(Size = 10)
RemixAutoML::multiplot(plotlist = list(p1,p2,p3,p4), cols = 2)

# Generate EvalPlots (boxplots)
p1 <- RemixAutoML::EvalPlot(data = data1,
                              PredColName = "predict",
                              ActColName = "target",
                              type = "boxplot",
                              bucket = 0.05)
p1 <- p1 + ggplot2::ggtitle("Calibration Evaluation p1: Corr = 0.50") +
  RemixAutoML::ChartTheme(Size = 10)

p2 <- RemixAutoML::EvalPlot(data = data2,
                              PredColName = "predict",
                              ActColName = "target",
                              type = "boxplot",
                              bucket = 0.05)
p2 <- p2 + ggplot2::ggtitle("Calibration Evaluation p2: Corr = 0.75") +
  RemixAutoML::ChartTheme(Size = 10)

p3 <- RemixAutoML::EvalPlot(data = data3,
                              PredColName = "predict",
                              ActColName = "target",
                              type = "boxplot",
                              bucket = 0.05)
p3 <- p3 + ggplot2::ggtitle("Calibration Evaluation p3: Corr = 0.90") +
  RemixAutoML::ChartTheme(Size = 10)

p4 <- RemixAutoML::EvalPlot(data = data4,
                              PredColName = "predict",
                              ActColName = "target",
                              type = "boxplot",
                              bucket = 0.05)
p4 <- p4 + ggplot2::ggtitle("Calibration Evaluation p4: Corr = 0.99") +
  RemixAutoML::ChartTheme(Size = 10)

RemixAutoML::multiplot(plotlist = list(p1,p2,p3,p4), cols = 2)
```


Calibration Evaluation p1: Corr = 0.50
Calibration Evaluation p3: Corr = 0.90
Calibration Evaluation p2: Corr = 0.75
Calibration Evaluation p4: Corr = 0.99
Observed Values
Predicted Percentile
0.00
0.25
0.50
0.75
1.00
0
25
50
75
100
colour
Actual
Predicted



The feature interpretation function graphs are very similar in nature to the model evaluation graphs. They display partial dependence calibration line plots, partial dependence calibration boxplots, and partial dependence calibration bar plots (for factor variables with the ability to limit the number of factors shown with the remainder grouped into "other"). The line graph version is for numerical features and have the ability to aggregate by quantile for quantile regression.

Functions include:

- `ParDepCalPlots()`

The cost sensitive optimizaition functions provide the user the ability to generate utility-optimized thresholds for classification tasks. There are two of these functions: one for generating a single threshold based on the values supplied to your cost confusion matrix outcomes and the second one provides two thresholds, where your final predicted classification could be (0|1) and "do something else". With the latter function, you would also need to supply a cost to the "do something else" option.

Functions include:

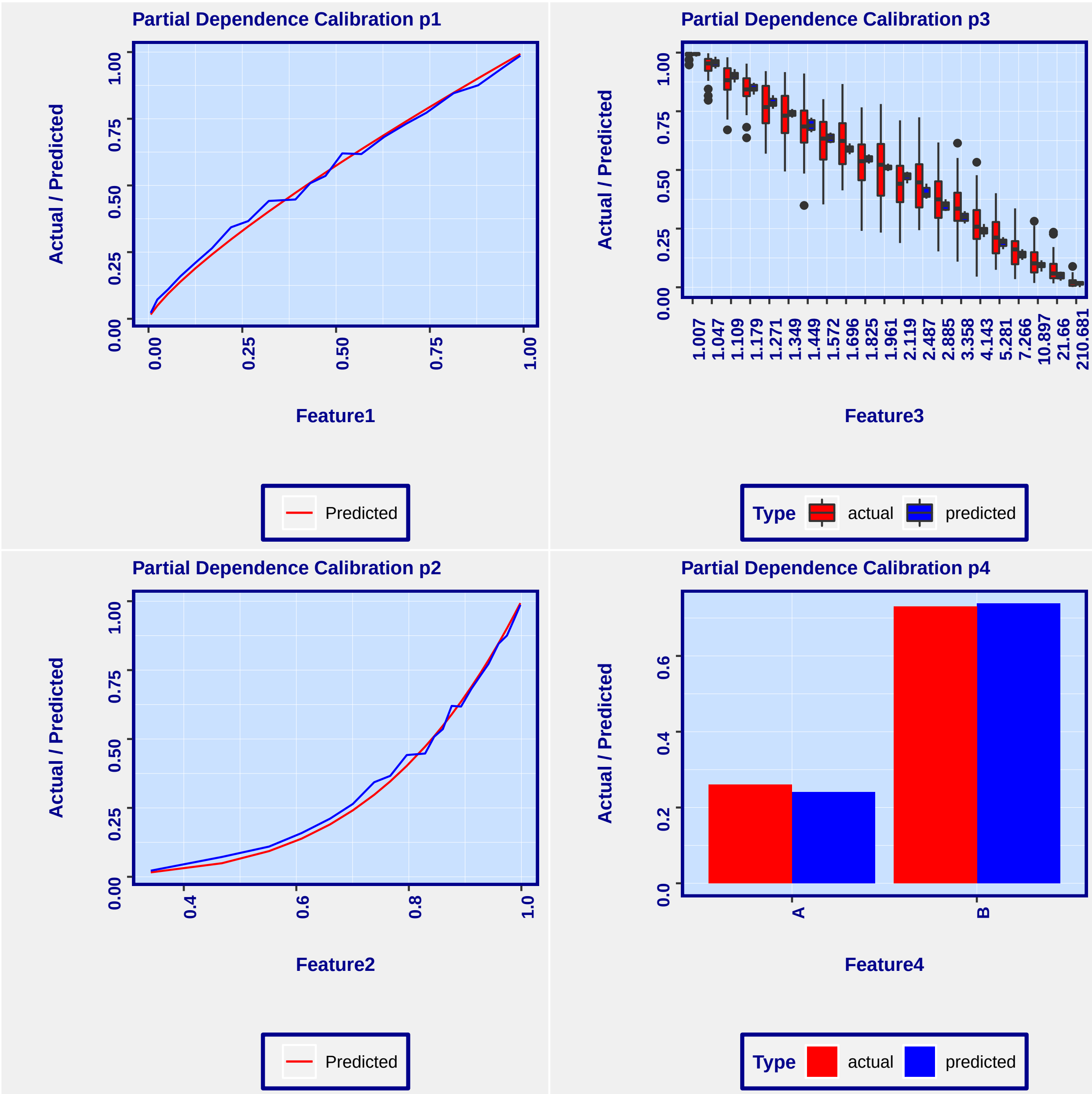
- `ParDepCalPlots()`

Demo of ParDepCalPlots() for calibration, boxplots, and barplots

Find more demos at **https://www.remyxcourses.com/course?courseid=modeling-tools-library-walkthrough**

```
library(RemixAutoML)

# Data generator function
dataGen <- function(Correlation = 0.95) {
  Validation <- data.table::data.table(target = runif(1000))
  Validation[, x1 := qnorm(target)]
  Validation[, x2 := runif(1000)]
  Validation[, predict := pnorm(Correlation * x1 +
                                  sqrt(1 - Correlation ^2) * qnorm(x2))]
  Validation[, Feature1 := (pnorm(Correlation * x1 +
                                    sqrt(1 - Correlation ^2) * qnorm(x2)))^1.25]
  Validation[, Feature2 := (pnorm(Correlation * x1 +
                                    sqrt(1 - Correlation ^2) * qnorm(x2)))^0.25]
  Validation[, Feature3 := (pnorm(Correlation * x1 +
                                    sqrt(1 - Correlation ^2) * qnorm(x2)))^(-1)]
  Validation[, Feature4 := pnorm(Correlation * x1 +
                                    sqrt(1 - Correlation ^2) * qnorm(x2))]
  Validation[, Feature4 := ifelse(Feature4 < 0.5, "A",
                                  ifelse(Feature4 < 1, "B",
                                         ifelse(Feature4 < 1.5, "C","D")))]
  return(Validation)
}

# Store data sets
data1 <- dataGen(Correlation = 0.95)

# Generate EvalPlots (calibration)
p1 <- RemixAutoML::ParDepCalPlots(data = data1,
                                  PredColName = "predict",
                                  ActColName = "target",
                                  IndepVar = "Feature1",
                                  type = "calibration",
                                  bucket = 0.05,
                                  Function = function(x) mean(x,
                                                              na.rm = TRUE),
                                  FactLevels = 10)
p1 <- p1 + ggplot2::ggtitle("Partial Dependence Calibration p1") +
  RemixAutoML::ChartTheme(Size = 10)

p2 <- RemixAutoML::ParDepCalPlots(data = data1,
                                  PredColName = "predict",
                                  ActColName = "target",
```

```r
                                  IndepVar = "Feature2",
                                  type = "calibration",
                                  bucket = 0.05,
                                  Function = function(x) mean(x,
                                                              na.rm = TRUE),
                                  FactLevels = 10)
p2 <- p2 + ggplot2::ggtitle("Partial Dependence Calibration p2") +
  RemixAutoML::ChartTheme(Size = 10)

p3 <- RemixAutoML::ParDepCalPlots(data = data1,
                                  PredColName = "predict",
                                  ActColName = "target",
                                  IndepVar = "Feature3",
                                  type = "boxplot",
                                  bucket = 0.05,
                                  Function = function(x) mean(x,
                                                              na.rm = TRUE),
                                  FactLevels = 10)
p3 <- p3 + ggplot2::ggtitle("Partial Dependence Calibration p3") +
  RemixAutoML::ChartTheme(Size = 10)

p4 <- RemixAutoML::ParDepCalPlots(data = data1,
                                  PredColName = "predict",
                                  ActColName = "target",
                                  IndepVar = "Feature4",
                                  type = "calibration",
                                  bucket = 0.05,
                                  Function = function(x) mean(x,
                                                              na.rm = TRUE),
                                  FactLevels = 10)
p4 <- p4 + ggplot2::ggtitle("Partial Dependence Calibration p4") +
  RemixAutoML::ChartTheme(Size = 10)
RemixAutoML::multiplot(plotlist = list(p1,p2,p3,p4), cols = 2)
```



Functions include:

- `threshOptim()`
- `RedYellowGreen()`

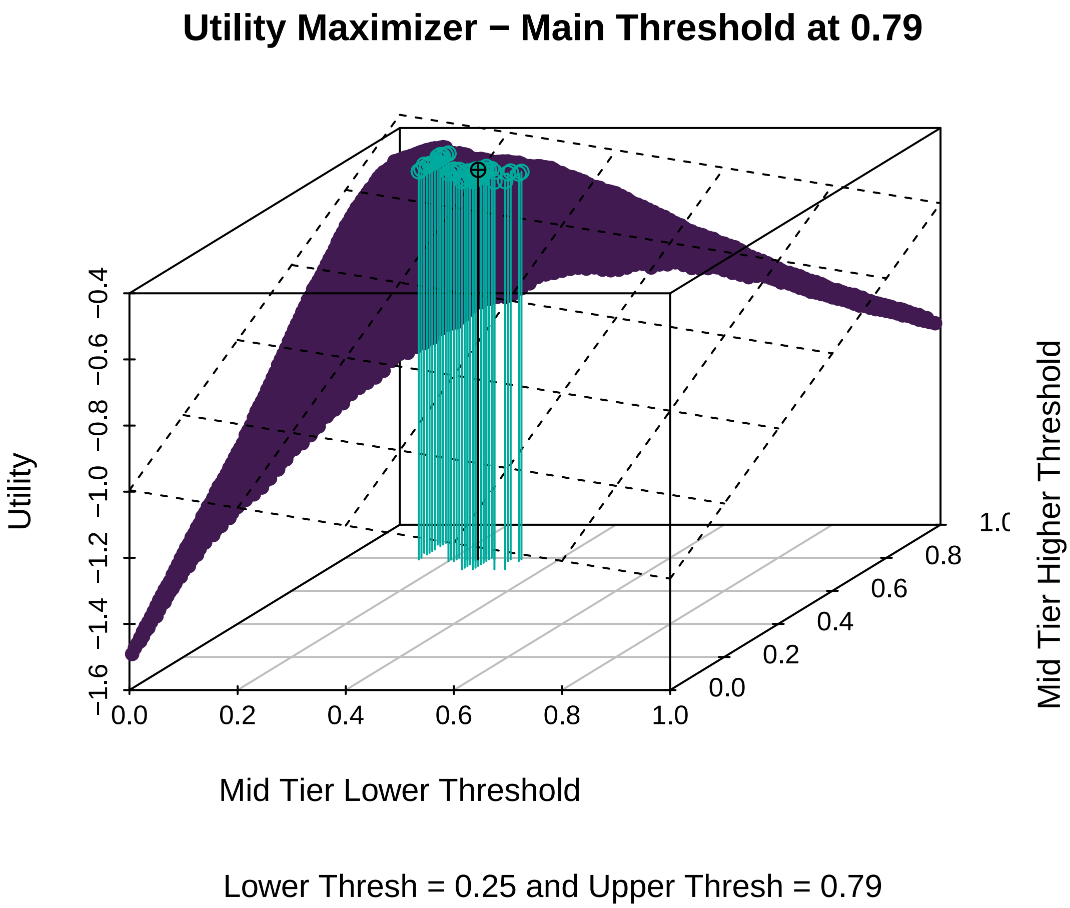
RedYellowGreen Output (simulated data)

Find more demos at **https://www.remyxcourses.com/course?courseid=modeling-tools-library-walkthrough**

```
library(RemixAutoML)
Correl <- 0.70
  aa <- data.table::data.table(target = runif(1000))
  aa[, x1 := qnorm(target)]
```

```r
aa[, x2 := runif(1000)]
aa[, predict := pnorm(Correl * x1 +
                        sqrt(1 - Correl ^2) *
                        qnorm(x2))]
aa[, target := as.numeric(ifelse(target < 0.5, 0, 1))]
data <- RemixAutoML::RedYellowGreen(
  aa,
  PredictColNumber  = 4,
  ActualColNumber   = 1,
  TruePositiveCost  = 0,
  TrueNegativeCost  = 0,
  FalsePositiveCost = -3,
  FalseNegativeCost = -2,
  MidTierCost       = -0.5,
  Cores = 1
)

knitr::kable(data[order(-Utility)][1:10])
```

TPP	TNP	FPP	FNP	MTDN	MTC	Threshold	MTLT	MTHT	Utility
0	0	-3	-2	TRUE	-0.5	0.79	0.25	0.79	-0.4215843
0	0	-3	-2	TRUE	-0.5	0.79	0.27	0.79	-0.4221797
0	0	-3	-2	TRUE	-0.5	0.78	0.25	0.78	-0.4224990
0	0	-3	-2	TRUE	-0.5	0.78	0.27	0.78	-0.4228998
0	0	-3	-2	TRUE	-0.5	0.80	0.25	0.80	-0.4233939
0	0	-3	-2	TRUE	-0.5	0.79	0.26	0.79	-0.4237080
0	0	-3	-2	TRUE	-0.5	0.83	0.14	0.83	-0.4240261
0	0	-3	-2	TRUE	-0.5	0.79	0.24	0.79	-0.4241313
0	0	-3	-2	TRUE	-0.5	0.80	0.27	0.80	-0.4243068
0	0	-3	-2	TRUE	-0.5	0.78	0.26	0.78	-0.4245630



Automated Feature Engineering Functions

This suite of functions are what will take your models to the next level. The core functions are the generalized distributed lag and rolling statistics functions. I have four of them.

Functions include:

- `GDL_Feature_Engineering()`
- `DT_GDL_Feature_Engineering()`
- `FAST_GDL_Feature_Engineering()`
- `Scoring_GDL_Feature_Engineering()`

The first three are used for building out lags and rolling statistics from target variables (numeric type; including classification models (0|1) and multinomial models with a little bit of work) and numeric features over your entire data set (no aggregation is done) with the option for creating the rolling statistics on the main variable or the lag1 version of the main variable. You can also compute time between records (by group) and add their lags and rolling statistics as well (really useful for transactional data). They can be generated using a single grouping variable (for multiple grouping variables you can concatenate them) and you can feed in a list of grouping variables to generate them by. The first function (**GDL__**) has the largest variety of rolling statics options but runs the slowest. The second function (**DT__GDL__**) runs the fastest but only generates moving averages. The third function (**FAST__GDL__**) is used for cases where you don't need to generate the features across the entire data set. Suppose you have a limited number of target variable instance but a rich history of data. You can use the FAST_GDL_ version to create lags and rolling statistics for N number of records previous to each target instance (i.e. not the entire historical data set). The fourth function (**Scoring_GDL_**) is for use in a production setting where you need to generate single instances of the feature set quickly. You basically feed in the same arguments as you used for the other versions and out the other end is the same set of features, identically named.

DT_GDL_Feature_Engineering and Scoring_GDL_Feature_Engineering Demo (simulated data)

Find more demos atFind many more at **https://www.remyxcourses.com/course?courseid=modeling-tools-library-walkthrough**

```
library(RemixAutoML)

# Build data for feature engineering for modeling
N <- 25116
ModelData <-
  data.table::data.table(GroupVariable = sample(
    x = c(letters,LETTERS,paste0(letters, letters),
          paste0(LETTERS, LETTERS),
          paste0(letters, LETTERS),
          paste0(LETTERS, letters))),
    DateTime = base::as.Date(Sys.time()),
    Target = stats::filter(rnorm(N,
                                 mean = 50,
                                 sd = 20),
                           filter = rep(1, 10),
                           circular = TRUE))
ModelData[, temp := seq(1:161), by = "GroupVariable"][
       , DateTime := DateTime - temp][
         , temp := NULL]
ModelData <- ModelData[order(DateTime)]
ModelData <- RemixAutoML::DT_GDL_Feature_Engineering(
  ModelData,
  lags           = c(seq(1, 5, 1)),
  periods        = c(3, 5, 10, 15, 20, 25),
  statsNames     = c("MA"),
  targets        = c("Target"),
  groupingVars   = "GroupVariable",
  sortDateName   = "DateTime",
  timeDiffTarget = c("Time_Gap"),
  timeAgg        = c("days"),
  WindowingLag   = 1,
  Type           = "Lag",
  Timer          = FALSE,
  SkipCols       = FALSE,
  SimpleImpute   = TRUE)
#> [1] 22

# Build data for feature engineering for scoring
N <- 25116
ScoringData <-
```

```r
  data.table::data.table(GroupVariable = sample(
    x = c(letters,LETTERS,paste0(letters, letters),
          paste0(LETTERS, LETTERS),
          paste0(letters, LETTERS),
          paste0(LETTERS, letters))),
    DateTime = base::as.Date(Sys.time()),
    Target = stats::filter(rnorm(N,
                                 mean = 50,
                                 sd = 20),
                           filter = rep(1, 10),
                           circular = TRUE))
ScoringData[, temp := seq(1:161),
      by = "GroupVariable"][, DateTime := DateTime - temp]
ScoringData <- ScoringData[order(DateTime)]

# Use WindowingLag = 1 to build moving averages off of the lag1 Target Variable to eliminate forward le
ScoringData <- RemixAutoML::Scoring_GDL_Feature_Engineering(
  ScoringData,
  lags           = c(seq(1, 5, 1)),
  periods        = c(3, 5, 10, 15, 20, 25),
  statsFUNs      = c(function(x) mean(x, na.rm = TRUE)),
  statsNames     = c("MA"),
  targets        = c("Target"),
  groupingVars   = c("GroupVariable"),
  sortDateName   = c("DateTime"),
  timeDiffTarget = c("Time_Gap"),
  timeAgg        = "days",
  WindowingLag   = 1,
  Type           = "Lag",
  Timer          = FALSE,
  SkipCols       = FALSE,
  SimpleImpute   = TRUE,
  AscRowByGroup  = "temp",
  RecordsKeep    = 1
)

# View some of new features
knitr::kable(ModelData[order(GroupVariable,-DateTime)][1:10,c(3,4,14)])
```

Target	GroupVariable_LAG_1_Target	GroupVariableMA_3_GroupVariable_LAG_1_Target
526.2521	486.5613	470.4853
486.5613	519.1806	498.2886
519.1806	405.7139	501.1943
405.7139	569.9713	499.9855
569.9713	527.8978	417.6583
527.8978	402.0875	406.4550
402.0875	322.9894	442.2419
322.9894	494.2880	498.4700
494.2880	509.4483	482.5530
509.4483	491.6735	509.3251

```r
# Ensure names equal
knitr::kable(
  data.table::as.data.table(
    cbind(ModelData_Names = sort(names(ModelData)),
          ScoringData_Names = sort(names(ScoringData[, temp := NULL])))))
```

ModelData_Names	ScoringData_Names
DateTime	DateTime
GroupVariable	GroupVariable
GroupVariable_LAG_1_Target	GroupVariable_LAG_1_Target
GroupVariable_LAG_2_Target	GroupVariable_LAG_2_Target
GroupVariable_LAG_3_Target	GroupVariable_LAG_3_Target
GroupVariable_LAG_4_Target	GroupVariable_LAG_4_Target
GroupVariable_LAG_5_Target	GroupVariable_LAG_5_Target
GroupVariableMA_10_GroupVariable_LAG_1_Target	GroupVariableMA_10_GroupVariable_LAG_1_Target
GroupVariableMA_10_GroupVariableTime_Gap1	GroupVariableMA_10_GroupVariableTime_Gap1
GroupVariableMA_15_GroupVariable_LAG_1_Target	GroupVariableMA_15_GroupVariable_LAG_1_Target
GroupVariableMA_15_GroupVariableTime_Gap1	GroupVariableMA_15_GroupVariableTime_Gap1
GroupVariableMA_20_GroupVariable_LAG_1_Target	GroupVariableMA_20_GroupVariable_LAG_1_Target
GroupVariableMA_20_GroupVariableTime_Gap1	GroupVariableMA_20_GroupVariableTime_Gap1
GroupVariableMA_25_GroupVariable_LAG_1_Target	GroupVariableMA_25_GroupVariable_LAG_1_Target
GroupVariableMA_25_GroupVariableTime_Gap1	GroupVariableMA_25_GroupVariableTime_Gap1
GroupVariableMA_3_GroupVariable_LAG_1_Target	GroupVariableMA_3_GroupVariable_LAG_1_Target
GroupVariableMA_3_GroupVariableTime_Gap1	GroupVariableMA_3_GroupVariableTime_Gap1
GroupVariableMA_5_GroupVariable_LAG_1_Target	GroupVariableMA_5_GroupVariable_LAG_1_Target
GroupVariableMA_5_GroupVariableTime_Gap1	GroupVariableMA_5_GroupVariableTime_Gap1
GroupVariableTime_Gap1	GroupVariableTime_Gap1
GroupVariableTime_Gap2	GroupVariableTime_Gap2
GroupVariableTime_Gap3	GroupVariableTime_Gap3
GroupVariableTime_Gap4	GroupVariableTime_Gap4
GroupVariableTime_Gap5	GroupVariableTime_Gap5
Target	Target

Functions include:

- `AutoWord2VecModeler()`
- `ModelDataPrep()`
- `DummifyDT()`

The **AutoWord2VecModeler** function converts your text features into numerical vector representations. You supply the function with your data set and all the text column names you want converted, and out the other end you have a data set with all the features merged on. The models can be saved to file and metadata saves their paths for scoring purposes in a production setting. The models built are based on H2O's word2vec algorithm and has done an execellent job at extracting high quality information out of those text columns. The **ModelDataPrep** function is used to prepare your data for modeling with the **AutoH20Modeler** function. It will convert character columns to factors, replace inf values to NA, and impute missing values (both numeric and factor based on supplied values). The **DummifyDT** function will turn your character (or factor) columns into dummy variable columns. You can specify one-hot encoding or not in which you will get N+1 columns for one-hot or N columns otherwise.

Miscellaneous Functions

Functions include:

- `AutoWordFreq()`
- `AutoH20TextPrepScoring()`
- `RecomDataCreate()`
- `ChartTheme()`
- `RemixTheme()`
- `multiplot()`
- `PrintObjectsSize()`
- `percRank()`

The **AutoWordFreq** function will go through a process of cleaning your text column, doing some other text operations, and ouput a table with word frequencies and a word cloud plot. The **AutoH20TextPrepScoring** will automatically prepare your text data for scoring. This function is run internally in the **AutoH20Scoring** function but you can utilize it outside for other purposes. The **ChartTheme** and **RemixTheme** functions will turn your ggplots into nicely formatted and colored charts, worthy of presentation. The **multiplot** function are for those who have had a terrible time plotting multiple graphs onto a single image. The **PrintObjectsSize** function is more of a debugging function for inspecting the size of variables in your environment (useful in looping functions). The **percRank** is simply a function to compute the percentile rank of every value in a column of data. **RecomDataCreate** will turn your transactional data set into a binary ratings matrix fast.