Lab 3a: Classification\_ModelBuilding

Objectives

In this chapter you will build models using different Classify methods. After completing this lab, you will be able to:

* Have familiarity with the Insight Analytics Ribbon
* Understanding of several Classification modeling techniques
* **Approximate Hands-on time: 25 minutes; Overall time including reading: 60 min.**

**Prerequisites**

* Have materials folder available to navigate to the appropriate datasets.

# Classify Explained

The Classify function helps you build a classification model based on existing data that resides in Excel, PowerPivot, or from an existing Predixion Insight Dataset in the cloud.

A classification technique will extract patterns or profiles that predict and explain the value of a column you choose (the dependent variable) and then described by the independent values contained in other columns. Classification is best used for predicting and exploring discrete variables – in other words, to select the class a record belongs to. There are many uses for classification. For example, you might use it to group your customers by categories, and then use the preferences and tendencies in that category to make more detailed predictions, such as customers' preferred products. You can also use classification to find patterns in events that are related to an outcome: for example, you can classify events that are related to server downtime, or to customer satisfaction ratings.

Below you will see a list of the Predixion Insight Classify techniques:

## Decision Trees

The Decision Tree is a popular modeling technique that offers a high degree of accuracy with straightforward outputs. Models can be trained quickly and the patterns found can be interpreted and explained easily.

A decision tree consists of a hierarchical grouping of data characteristics and the relationships between the groupings. The tree begins with a single variable (the root node) that is split into two or more branches that represent the classes (if the input variable is discrete) or the range along the scale of a continuous input variable. Subsequent internal nodes are created from additional groupings and splits and eventually the paths terminate at a leaf node. These paths make up a set of rules regarding the target variable.

When making predictions using a decision tree model, each case moves through the tree following a path according to its characteristics from root to a leaf. The eventual outcome on a leaf will determine the predicted value and probability. These types of predictions are very efficient and the Predixion Insight query tool will output the prediction results along with the rules applied to the case.

## Naïve Bayes

The Naïve Bayes algorithm is based on learning by evidence. The technique works by counting the correlations between each input variable and the specified target outcome to design a prediction based on the evidence observed in the training set.

Models built using the Naïve Bayes technique provide a way to explore the relationships, similarities and differences among the attributes used.

## Neural Network

Neural Networks can be used for a variety of analytical tasks. Neural networks are well suited for finding correlations among complex sets of data, especially when rules cannot be easily derived by using other algorithms. Examples of uses include currency fluctuation and other highly fluid financial projections based on many inputs and relatively fewer outputs.

## Logistic Regression

Logistic Regression is a well-known statistical technique that is used for modeling the contribution of various factors to a set of outcomes, usually binary. Logistic regression is highly flexible, taking any kind of input, and supports analytical tasks such as using demographics to make predictions about diseases or financial risks.

**TIPS:**

* Choose a discrete target column with a small number of values to predict, such as Yes/No or Pass/Fail/Unknown.
* Use the Re-label tool before processing numeric values to create groupings. Otherwise the wizard will automatically group them and the results might not reflect more conventional groupings.
* Use Clean Data to reduce the number of unique values in non-numeric columns.
* Don't include duplicate columns. For example, a date column, and its month too, the two values "echo" each other and can affect the results.
* Avoid using columns with many unique values, like ID columns.
* Target columns that have a small number of values such as Yes/No or Pass/Fail/Unknown, work best.

# Hands On Exercise – Model Building

### Background

You are working for a hospital who wants to identify patients that are most likely to be readmitted. To perform this task, the hospital collected patient data over several years that includes detail data on the following patient attributes:

* Outcomes
* Demographics
* Admission/Discharge Details
* Medications
* Findings and Labs
* Encounter History
* Comorbidity Scores and conditions

From a modeling perspective, the goal of this lab is to use the historic patient data to create four candidate models that will be used in the next lab to determine the champion model i.e. the best performing model from an accuracy perspective. Using this model we can determine if a patient will be admitted and the probability of that readmission.

|  |  |
| --- | --- |
| **Tasks** | **Action** |
| 1. Navigate to the data set. | 1. Select **File | Open | Libraries | Documents | Predixion** folder. 2. Open **ReadmissionsSanitizedTraining.xlsx** and open the **RawDataSet** data sheet to access the file.      1. Place cursor in the **Readmitted** Column |
| 1. Run the Classify Wizard for Decision Tree model. | 1. Select the **Classify** button from the Insight Analytics ribbon.   C:\Users\KSOUTH~1\AppData\Local\Temp\SNAGHTML9613b2c.PNG   1. Review the Introduction Wizard and then Select **Next**. 2. The wizard will default to the Excel table from step 1. Select **Next**.      1. Select the **~~Decision Tree~~ (All Methods)** method for analyzing the column **Readmitted** and select **Next**.     Check the checkbox to the left of the **Name** label under the “**Columns for the model**” dialog box label as shown in the dialog box above.  **Very important**: Remove proxy outcome columns! Scroll down and unselect the “**InpatCnt\_6m”**,”**PriorDC\_30dToHomeCare**”, “**PriorDC\_30dYesNo**”,”**PriorDC\_6mDaysSinceLastDC**”, checkboxes.   1. Select **Next**.   C:\Users\KSOUTH~1\AppData\Local\Temp\SNAGHTML961c66b.PNG  *The Specify Testing Data option that allows you to define the amount of testing or hold-out data you wish to use. This will be managed automatically for you to maintain a set of data for validation only along with the training data and models.*   1. Accept the dataset name and select **Finish**.   *Now you will see the system begin to work through the Classify operations. The window tells you the status of the job from uploading data through training and processing. Before the process completes, we will minimize this process to the Task Pane.*  C:\Users\KSOUTH~1\AppData\Local\Temp\SNAGHTML9625aae.PNG   1. Once the task completes running on the Predixion Insight server the **Results** button will appear. |
|  | 1. Click on the Results button *go to the newly created worksheet, and select the link :* ***Click here to view the report in the viewer****.)* Close the model viewer for the time being. |