****

**Data Mining with IBM SPSS Modeler 14.2**

**University of Arkansas**

**David Douglas**

**Data mining tasks include classification (directed/supervised) models as well as (undirected/unsupervised) models of association analysis and clustering.**

**Data Mining—**

Data mining has many definitions and may be called by other names such as knowledge discovery. It is generally considered to be a part of the umbrella of tasks, tools, techniques etc. within business Intelligence (BI). Many corporate managers consider BI to be the heart of all the processes that support decision making at all levels. A definition of data mining typically includes large datasets, discovering previously unknown knowledge and patterns and that this knowledge is actionable. That what is discovered is not trivial but can be usefully applied. BI and its Data Mining component are receiving considerable attention and fanfare as companies utilize BI for competitive advantage.

Different authors may address the data mining tasks slightly different from each other but the following terminology provides a helpful and useful basis for discussing data mining. The data mining tasks are:

* Description
* Estimation
* Classification
* Prediction
* Association Analysis
* Clustering

**Description—**used descriptive statistics to better understand and profile areas of interest. Thus a variety of well known statistical tools and methods are used for this task—including frequency charts and other graphical output, measures of central tendency and variation.

**Data Mining Tasks with a Target or Dependent Variable**

**Estimation**, **classification** and **prediction** are data mining tasks that have a target (dependent) variable. Sometimes these are referred to as predictive analysis; however, many authors reserve the term Prediction to use of models for the future. The terms **supervised** and **directed** apply to these data mining tasks. **Estimation** data mining tasks have an interval level dependent target variable whereas **classification** data mining tasks have a categorical (symbolic) target variable. An example of an estimation data mining task would be estimating family income based on a number of attributes; whereas a model to place families into the three income brackets of Low, Medium or High would be an example of a classification data mining task. Thus, the difference between the two tasks is the type of target variable.

When either an estimation data mining task or classification task is used to predict future outcomes, the data mining task becomes one of **Prediction**. Again, estimation and classification are referred to as predictive models because that would be the typical application of models built for these data mining tasks.

In summary, the most important concept is that estimation and classification data mining tasks require a target variable. However, the difference lies in the data type of the target variable.

**Data Mining Algorithms for Directed/Supervised Data Mining Tasks—linear regression** models are the most common data mining algorithms for **estimation** data mining tasks; although **neural networks** are excellent for **estimation** data mining tasks. Of course, linear regression is a very well known and familiar technique. A number of data mining algorithms can be used for **classification** data mining tasks including **logistic regression**, **decision trees**, **neural networks**, **memory based reasoning (k-nearest neighbor)**, and **Naïve Bayes**.

**Data Mining Tasks without a Target or Dependent Variable**

**Association Analysis** and **Clustering** are data mining tasks that do not have a target (dependent) variable. Affinity analysis is another term that refers to association analysis and is typically used for market basket analysis (MBA) although association analysis can be used for other areas of study. MBA is essentially analyzing what purchases tend to be purchased together—that is what items tend to have an affinity with other items. **Clustering**, having no target variable, algorithms attempt to put records into groups based on the record’s attributes. The critical concept is that of similarity—those within a cluster are very similar to each other and not similar with those in another cluster.

**Note—**because these data mining tasks do not have a target variable, their corresponding models cannot be used for prediction. Thus, they are many times exploratory in nature and their results can be used downstream in predictive models.

**Data Mining Examples in this Tutorial**

The data mining tasks included in this tutorial are the directed/supervised data mining task of classification (Prediction) and the undirected/unsupervised data mining tasks of association analysis and clustering. **Simple** and **Multiple Linear Regression** will be used to illustrate estimation. Three data mining algorithms for the classification data mining tasks will be illustrated and compared: **Decision Trees, Logistic Regression,** and **Neural Networks**. Recall that classification has a categorical target variable.

Association analysis and clustering are the undirected/unsupervised data mining tasks illustrated in this tutorial. The clustering algorithm is ***k*-means**.

**Data mining overview summary**

|  |  |  |
| --- | --- | --- |
| ***Data mining tasks*** | ***Target Variable*** | ***Typical Data Mining Algorithm(s)*** |
| Description | No | Statistics, including descriptive, & visualization |
| Directed or supervised data mining | | |
| Estimation | Yes  Interval Numeric | Linear Regression, Neural Networks |
| Classification | Yes  Categorical | Logistic Regression, Decision Trees, Neural Networks, Memory Based Reasoning, Naïve Bayes |
| Prediction | Yes | Estimation and Classification models for future |
| Unsupervised (non directed) data mining | | |
| Association Analysis | No | Affinity Analysis (Market Basket Analysis) |
| Clustering | No | *k-*means, Kohonen Self Organizing Maps (SOM) |

**Data Mining using IBM SPSS Modeler 14.2**

**University of Arkansas**

There are multiple ways you can run IBM SPSS data mining models using the Enterprise Systems Resources at the University of Arkansas. IBM SPSS Modeler 14.2 is part of IBM’s Academic Initiative and the software if free to universities for academic use only. If you obtain your academic IBM SPSS Modeler 14.2, you can access the large DB2 tables on the University of Arkansas IBM z10 via a DSN created using ODBC software. If you use the University of Arkansas terminal server, you can use your own local data or access data in DB2 tables on the IBM z10 hosted at the University of Arkansas which contains a DSN. These options all require requesting an account from the University of Arkansas. Use the following link to request accounts.

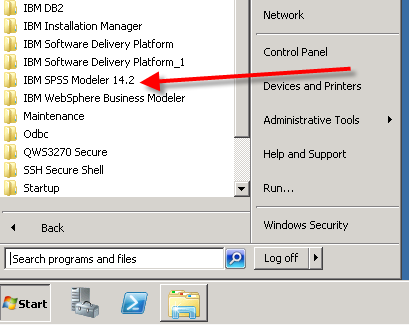
<http://enterprise.waltoncollege.uark.edu/IBM.asp>

All the examples in this document illustrate using IBM SPSS 14.2 via remote desktop. Once you receive your University of Arkansas IBM account, access will be via remote desktop connection. Remote access documentation is at the following link:

[http://enterprise.waltoncollege.uark.edu/Remote\_Desktop\_IBM\_GW.pdf](%20http:/enterprise.waltoncollege.uark.edu/Remote_Desktop_IBM_GW.pdf%20)

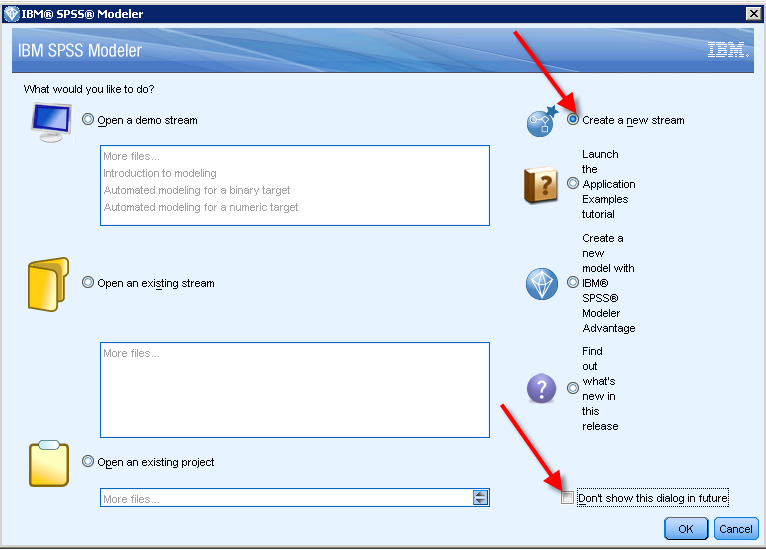
Once you’re logged in to REMOTE you can use several IBM software systems including DB2 and IBM SPSS Modeler 14.2. The following provides a data mining examples—the data mining models illustrating **classification** tasks use a table of 3333 telecommunications records. These historical records include the column, churn, which represents whether a customer left the telecommunications company or not. The idea is to build and select the best model so it can be used for predictive purposes with new customers.

Click either the IBM SPSS 14.2 icon on the Desktop or click Start 🡪 All Programs 🡪 IBM SPSS Modeler 14.2 as shown below.



**IBM SPSS Modeler 14.2 Basics**

When opeing IBM SPSS Modeler 14.2 , a prompt screen may appear, if so click the Create a new stream option. There is a check box that will prevent the prompt screen from appeariing that you may click. Click the OK button.



SPSS Modeler 14.2 uses the term Stream to represent the graphical flow of a data mining project. Streams can be saved independently of projects. The above dialog allows the user to select from a number of options includes tutorial examples, example streams, existing streams or existing projects.

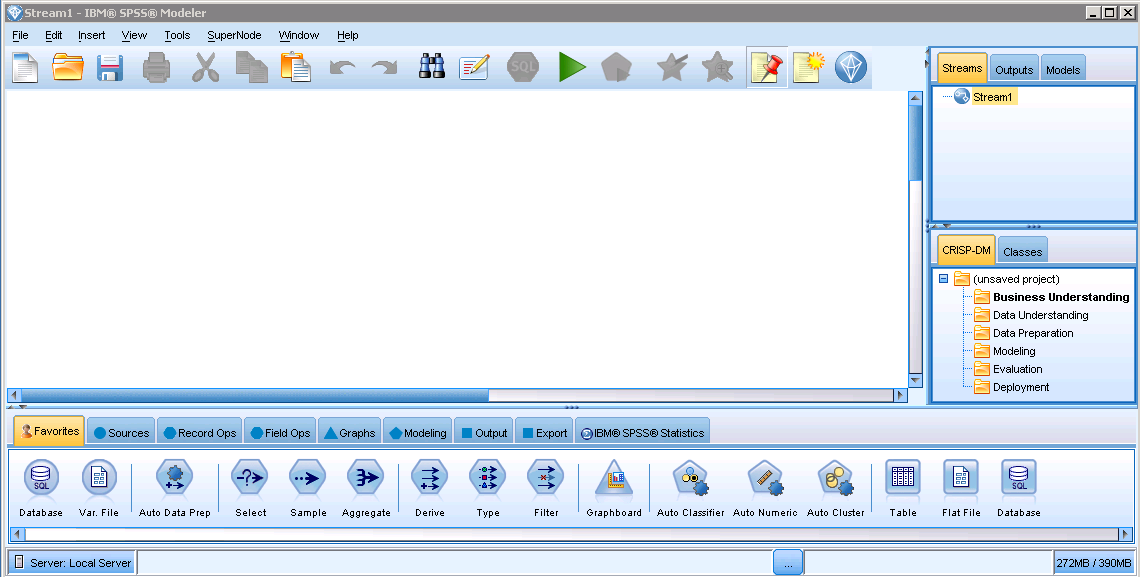
For this tutorial, the default selection of “Create a new stream” will be used. Clicking the OK button will open the following window. The opening window is shown below.

Shortcut bar

Default name

Main menu bar

Drawing area or Canvas



Steams, output and models manager

Project Manager

Nodes—arrows are not pointed to all node types

Tabs for different palettes

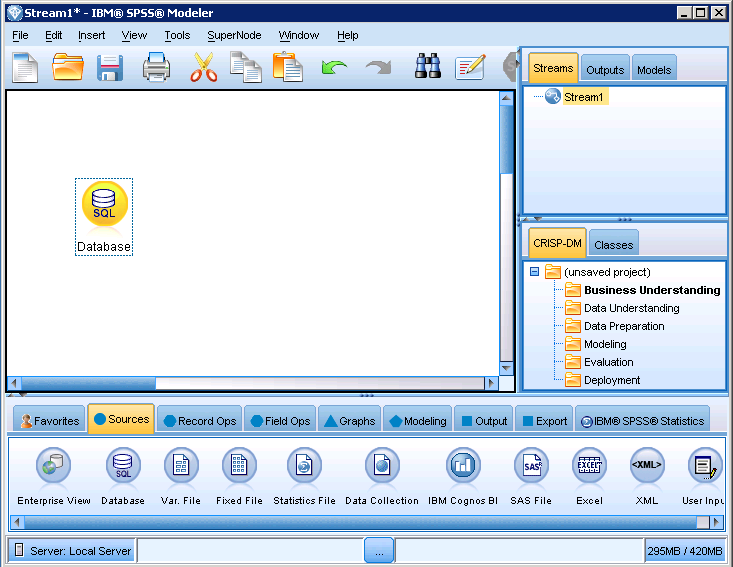
Again, IBM SPSS Modeler 14.2 refers to its models as streams—the default name being Stream1. Click on each of the tabs below the drawing canvas – Favorites, Sources, Record Ops, Field Ops, Graphs, Modeling, Output, Export and IBM SPSS Statistics. Also note that each of the icons in each category has the same geometric shape—for example, all the icons for graphics are triangular. Review the different tools available; the categorization makes it easy to locate the type of tool needed. For example, the Sources tab includes the various tools needed to use a data source; the Output tab includes a variety of tools for displaying data, etc. More later on the two Windows to the right of the drawing canvas.

To begin creating a simple model – stream if you will – do the following:

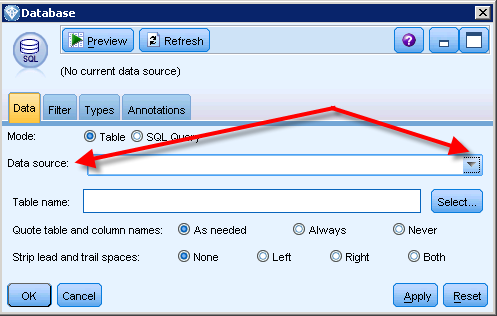
1. Click the Sources tab. As you move the mouse pointer over the Sources icons, the tool tip will indicate what type data source it handles—some are obvious like Excel and SAS. Note the tool tip for the Var. File icon indicates “Reads data from delimited column text files.”
2. This example will use a home equity sample of data contained in a DB2 Table—see metadata below

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Model Role | Measurement Level | Description |
| BAD | Target | Binary | 1=client defaulted on loan, 0=loan repaid |
| CLAGE | Input | Interval | Age of oldest trade line in months |
| CLNO | Input | Interval | Number of trade lines |
| DEBTINC | Input | Interval | Debt-to-income ration |
| DELINQ | Input | Interval | Number of delinquent trade lines |
| DEROG | Input | Interval | Number of derogatory reports |
| JOB | Input | Nominal | Six occupations categories |
| LOAN | Input | Interval | Amount of loan request |
| MORTDUE | Input | Interval | Amount due on existing mortgage |
| NINQ | Input | Interval | Number of recent credit inquires |
| REASON | Input | Binary | DebtCon=debt consolidation  HomeImp=home improvement |
| VALUE | Input | Interval | Value of current property |
| YOJ | Input | Interval | Years at present |

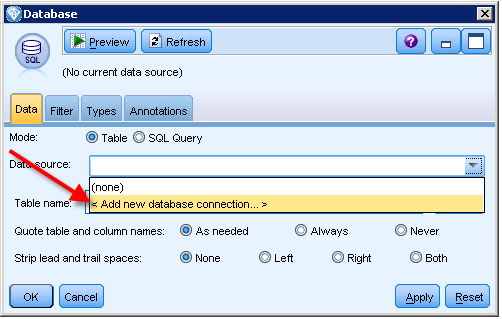
1. Click the **Sources** tab if needed and then double-click the database icon – this icon is placed on the drawing canvas.



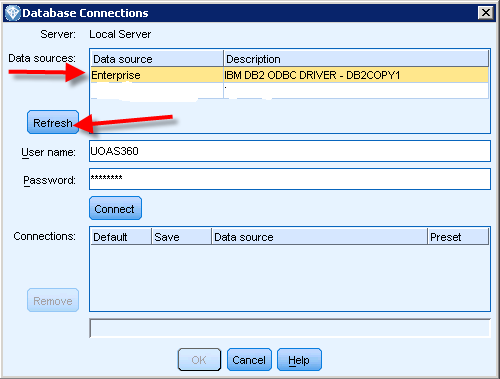
1. Right-click the Database node and select Edit. The following main dialog opens which allows you to create a new connection to a table via a DSN. If you just opened the terminal server, then you would click the drop-down list box for Data source: and select <Add new database connection…> as shown below.



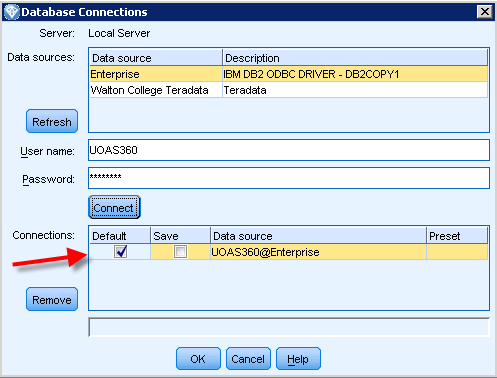
1. Click the drop down box and select Add new database connection.



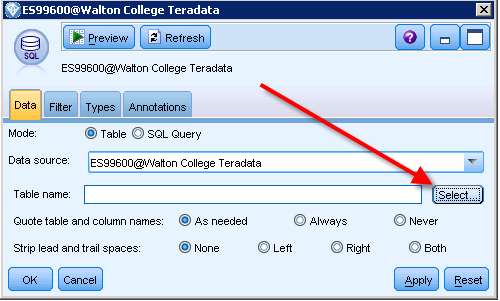
1. There is already a connection to the Teradata name Walton College Teradata.



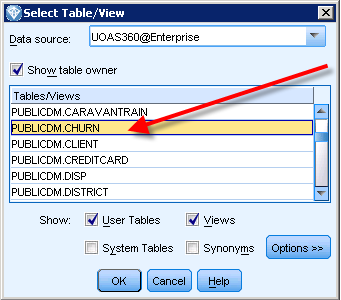
1. Click the Refresh button to show other DSN that are on the terminal server—if any. There should be one for DB 2 name **Enterprise** and one for Teradata named Walton College Teradata. Click the Enterprise data source, enter your login credentials and then click the Connect button. You should have something like below.



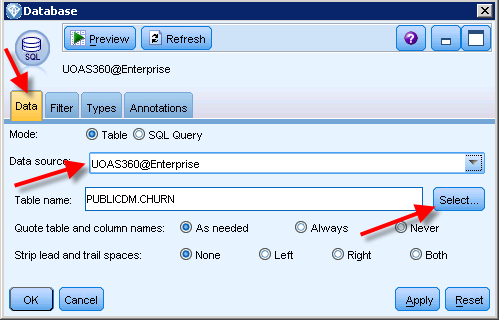
1. Click the OK button to return to the Database dialog. The DSN name is “UOAS360@Enterprise” is shown as the DSN. Click the Select button to search for the table you wish to use as shown below.



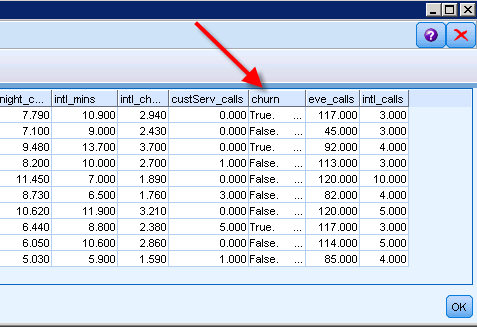
1. For this example, select the Churn table from the PUBLICDM database as shown below.



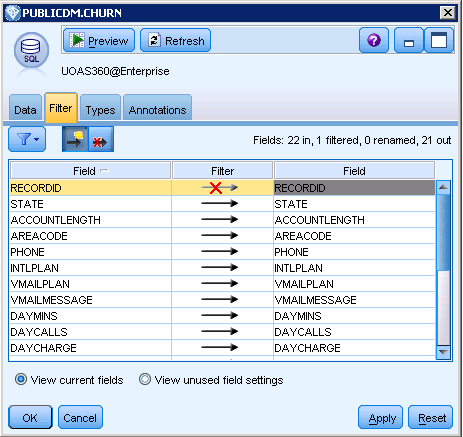
1. Click the OK button and you are now returned to the Database dialog with the connection and table selected as shown below. This data have values related to telecom company with the variable **churn** used to determine if a customer left the company or not. True means the custom left the company and false means the customer remained with the company.
2. IBM SPSS Modeler 14.2 has a number of tabs that help the modeler with typical data steps needed. The Edit dialog opens with the Data tab as its default setting—see red arrow toward upper left. Also note the **Filter** and **Types** tabs.



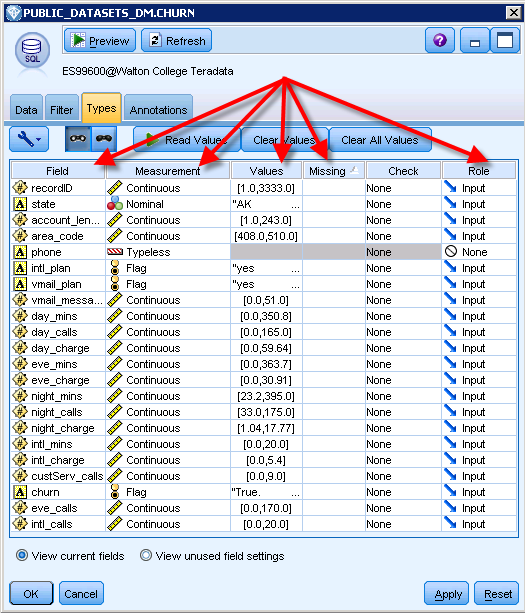
1. Click the **Preview button** toward the upper left—SPSS Modeler 14.2 will read ten rows of data for you to preview. ***If the file you are trying to review is open, you will receive an error when you try to Preview the data.*** A portion of the Preview window appears as shown below. Also, note that the column—Churn—is categorical as the entries are true/false. Churn represents whether a customer left the company or not.



1. Click the **Filter** tab. You can filter variables that you do not plan to be use in the model by clicking the arrow in the variable row—the arrow turn to a red **x** as shown below the variable RECORDID. Clicking the **x** will change the **x** back to an arrow which allows the variable to be used in the model. For this example, include all the variables in the model.



1. Click the **Types** tab and then click the **Read Values** button—accept by clicking the OK button on the ensuring prompt.



All the fields are listed in the left column (Field) and the icon to their right indicates the data type IBM SPSS MODELER 14.2 has determined it should be based on its reading of the data. Note that one usually has to change some of the data types because of numeric coding, etc. To change a data type for a variable, click in the variable’s cell in the Measurement column and select the correct data type. Note that SPSS Modeler 14.2 uses **Continuous** for numeric data and **Nominal** for a list of more than two categories that do not have a natural order. It uses **Flag** for binary variables and **Typeless** for nominal data types that too many unique values such as **Phone** and its role will be set to **None**.

In this model, the variable **Churn** indicates whether the person left the company or stayed with the company and will be the target variable. Measurement for BAD should be changed from **Continuous** to **Flag**.

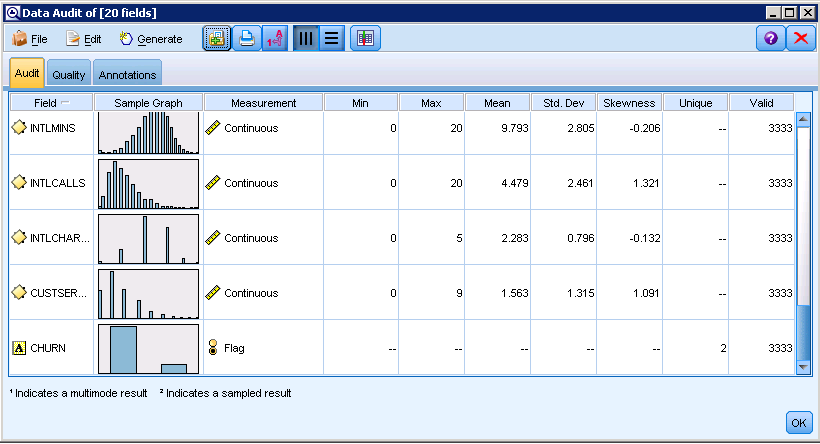
Click in the Measurements column to change the data type of a variable.

Also, note that the **Values** column shows a list of values it has read. For example, the two values of 0.0 and 1.0 are the only values read for the variable BAD. Also, note that the column **Missing** shows no missing values—that is because SPSS Modeler 14.2 changed the missing values to $null$ and does not automatically indicate they are missing values. Later, we will demonstrate how to have the $null$ be considered missing.

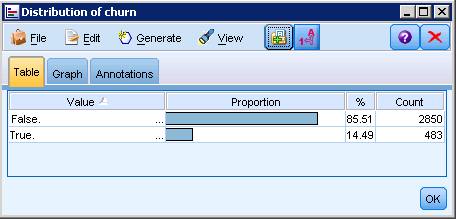
The last column, **Role** is exceptionally important. All the variables are initially set to the value of **Input** which means they will be input variables (or in statistically terms—independent variables). Recall from statistics that a regression model, for example, requires a dependent variable—which is designated as **Target** by IBM SPSS Modeler 14.2. To change a variable’s role, click in its cell and change to the desired value.

Be sure to set the variable **churn** to be the Target Variable and set the **recordID** to None.

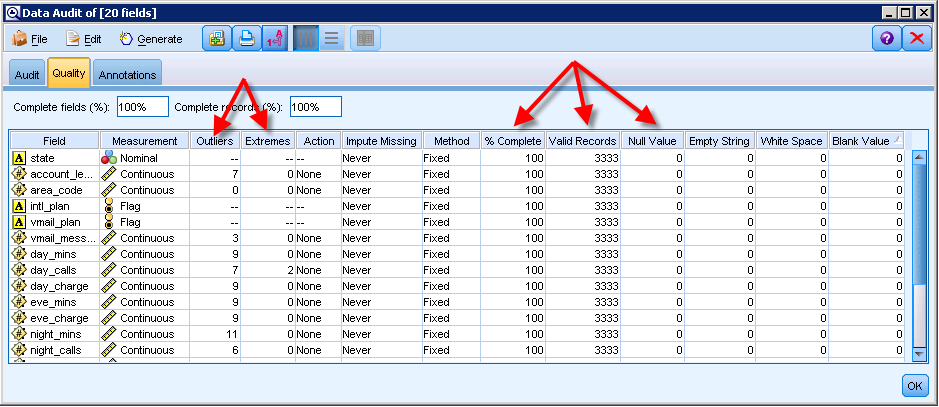
1. A good thing to do now is to get better acquainted with the data. Select the source node and then click the Output tab at the bottom. Locate and double-click the **Data Audit** icon which adds it to the right of the source node. Right-click the Data Audit node and select run. A portion of the output is shown below with the default tab **Audit** selected.



Review the row for the variable **churn.** Note that **churn** is a Flag data type with two unique values. Double-click the Sample Graph for the **churn** to see a display of the distribution. As shown, about 85.5% of the records have a False value for **churn** and 14.5% of the records have a value of True.

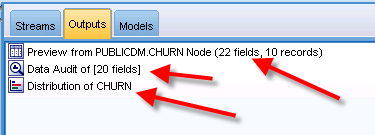


Close the Distribution of **churn** window to return to the Data Audit window. Click the Quality tab of the Data Audit Window which is shown below.



Note the table provides a lot of valuable information—outliers and extreme values; the % Complete for a variable for the all the records, the number of valid records for the variable and the Null Values for the record. Note there will be a horizontal slider at the bottom of the window to scroll to additional columns to the right. The Audit and Quality tabs provide a great way to get to know the data and provides guidance on data issues and/or creating indexes, etc.

Recall the upper right hand window of the Modeling window has three tabs – **Streams**, **Outputs** and **Models**. Click the **Outputs** tab in it is not already selected and you will find the outputs from the Preview, Data Audit and Distribution of bad. These will be available until you delete them or exit the program without saving the project (you may elect to only save the streams).

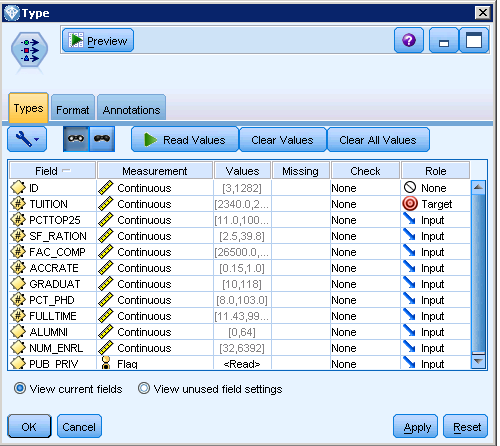


1. There are a number of tools you can use to explore the data. For example, under the Graph tab, the Plot, Distribution, Histogram, Pie chart, etc. provide a multitude of ways to view the data. Be sure to find the Statistics tool in the Output tab. These tools are relative easy to use and you are left to experiment with them.

**Simple Linear Regression model**

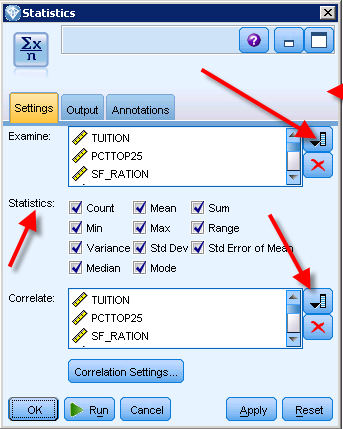
Using the output and graph tools to explore the data is relatively straight forward. This part of the introduction will illustrate using a simple model—a simple linear regression model (one dependent variable and one independent variable) using the data in the table **TUITION**. The dependent variable (target variable in data mining terminology) will be **TUITION** and we will seek the strongest numerical predictor before building a simple regression model.

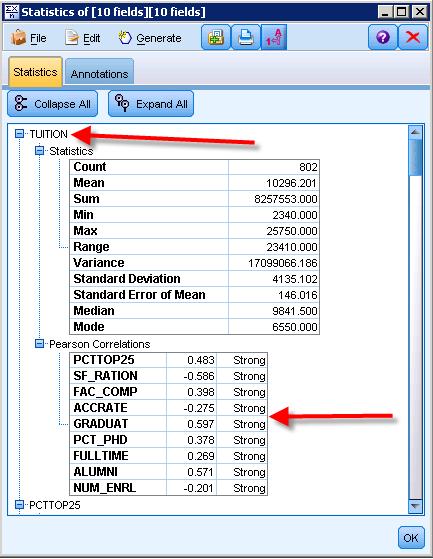
Add a Database source node, edit and select the PUBLICDM.**TUITION** Table. Connect the source node to a Type node and Edit the Type node to be as follows. Note that the Role for the predictor variable **PUB\_PRIV** is set to flag and **tuition** has been set as the Target variable.



Select the Type node and double-click the Statistics node in the Output tab. Double-click the **Statistics** node and add the variables for which you wish to have statistics produced. To the right is a drop down box to select the variables for the Examine: pane. Select all the variables except the **PUB\_PRIV** flag variable.

Click the desired statistics check boxes for the Statistics: section of the Dialog. For the Correlate: pane, select all the variables except the **ID** and **PUB\_PRIV** variables.

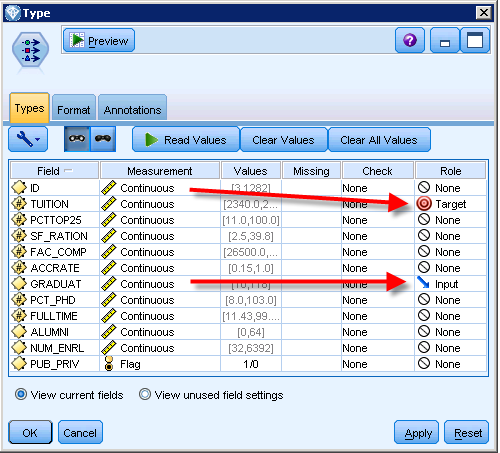




Run the Statistics node. A part of the output is shown to the right. Statistics are produced for the target variable, **TUITION**, and the Pearson Correlations for each of the predictor variables are shown in a table below the Statistics table. In this case, the strongest correlation is the variable, **GRADUAT**, and the correlation is positive.

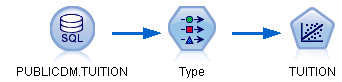
Based on this information, it makes sense to create a simple linear regression model with tuition as the target variable and **GRADUAT** as the single predictor variable.

1. Move the Statistics nodes out of the way or just delete it.
2. Double-click the Type node to open it – or right-click it and select Edit. The resulting dialog box is shown below.

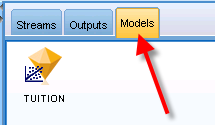


The task that has to be done here is to set **TUITION** as the target variable and **GRADUAT** the predictor variable. Click the **TUITION** in the **Field** column and then set its Role to Target and set all the remaining variable to None except **GRADUAT** which will be the single predictor variable. This is accomplished by clicking in the **Role** column and selecting the appropriate values—above screen shot is after these changes have been made to the Type node. Note that multiple variable may be set at one time by selecting them and changing the Role.

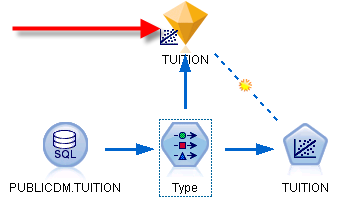
1. Add a Regression node to the right of the Type node. The Regression node defaults to using the variable specifications of the Type node so nothing else needs to be changed—but note that there are a number of options one can set on the Regression node.



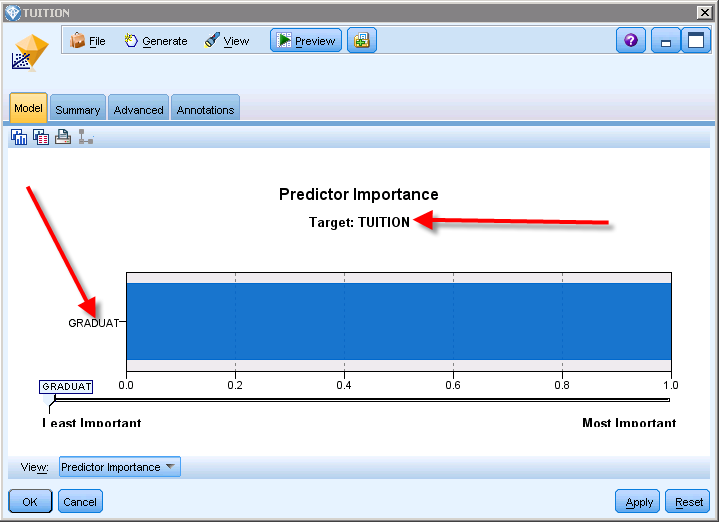
1. Run the Regression node accepting default values for all the remaining options. Note that a model node MORTDUE takes on the name of the target variable—however you can change the name. Successful execution of the model does two things—a model nugget is placed in the Models tab in the Window to the upper right of the Canvas as shown below. To view the results of the model, right-click and select Browse.



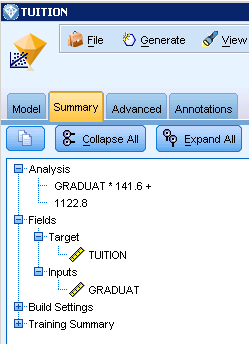
Starting with version 14, a model nugget is also placed on the canvas as shown below with a dashed line from the model that created the model nugget. The model nugget is in gold.



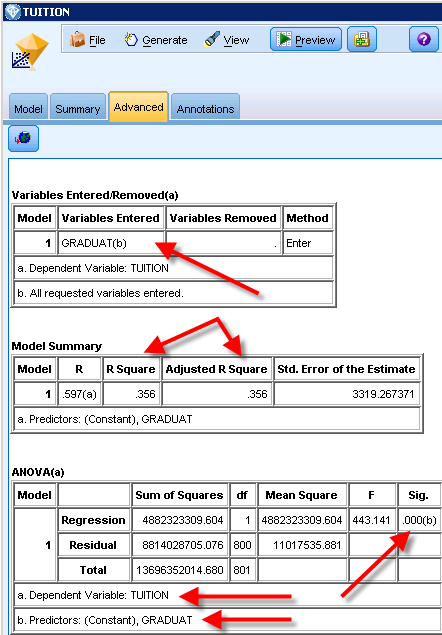
Double-clicking the model nugget on the canvas opens the results window in the default Model tab. Note that the graph provides predictor importance in the model—not helpful in this case as there is only one predictor variable-**GRADUAT**. However, it is very helpful in models with multiple predictors. Click the **Summary** tab.



1. This Window displays the TUITION regression model shown to the right. The **Summary** tab provides considerable information about the model including the fields and analysis setting. What is the regression equation?

The resulting model is:

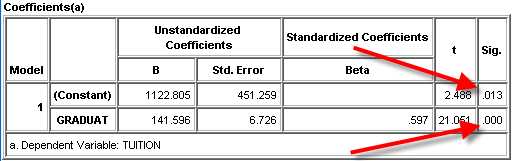
TUITION = 1122.8 + 141.6\*GRADUAT

1. Click the **Advanced** tab to get a table output of the statistics for the run—a partial screen listing is below.

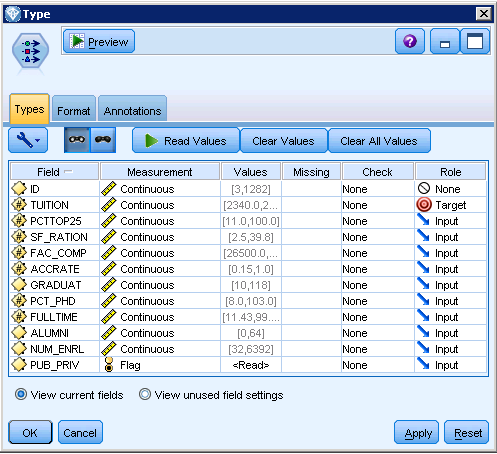
Note that the model is significant based on the F statistic in ANOVA analysis – WITH A significance value = .000. If using a significance level of .05, the model is significant because the computed significance level is less than the specified significance level.

The R-Square and Adjusted R Square values are equal and indicate the model explains quite a bit of the variation, 35.6%, in the target variable.

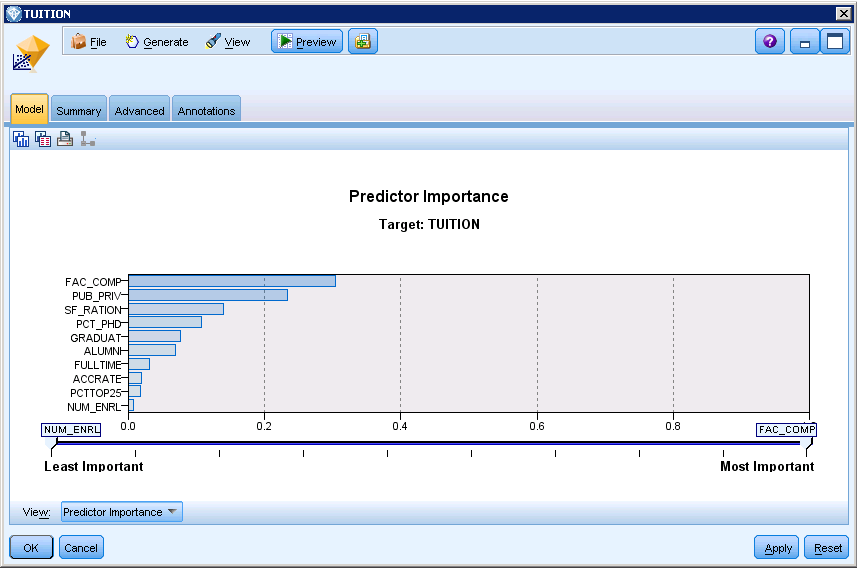
The Coefficients output table shown below indicates that both the intercept (Constant) and predictor variable, **GRADUAT**, are significant.

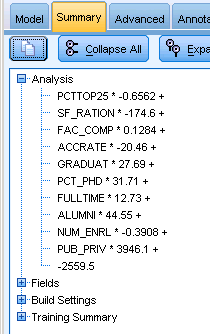


Return to the canvas and add a new Type node or change the current Type node to add predictor variables (set to Input in the Role column) as shown below. If you add a new type node, then also connect a new Regression node to it.



Run the new model view the output. The model out with predictor importance indicates the FAC\_COMP variable is the most important predictor variable followed closely by several other variables. GRADUAT is fifth in importance when considered with other predictor variables.

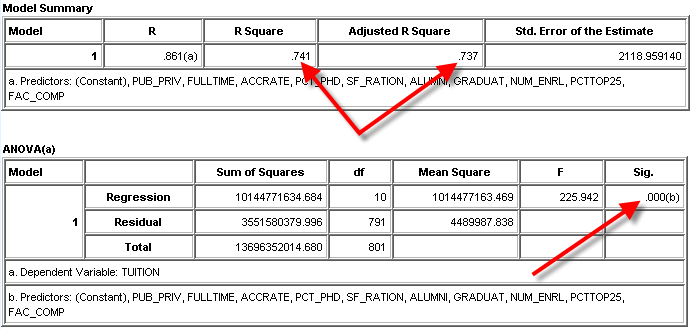


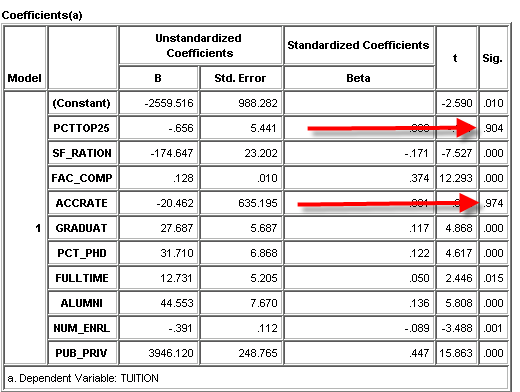


Click the Summary tab to find the regression model for TUTITION as shown to the right.

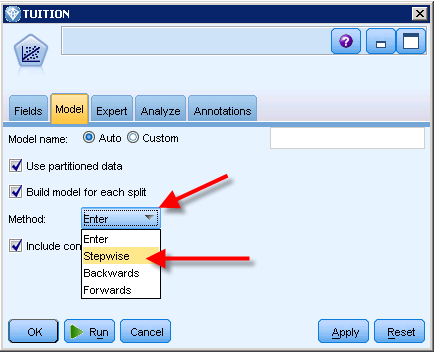
Clicking the Advanced tab will show the model is significant and the R-Square and Adjustred R Square values are high at .741 and .737 respectively. Adding the additonal predictor variaables more than double the amount of variation in the target variable explained by the predictor varaibles.

However, note that in the coefficients table, two variables are not signigicant—**PCTTOP25** and **ACCRATE**. If you were to select a method other than the default method of **Enter**, such as **Stepwise**, then these two varialbes my not have been included in the model and the model would probably be a better model.

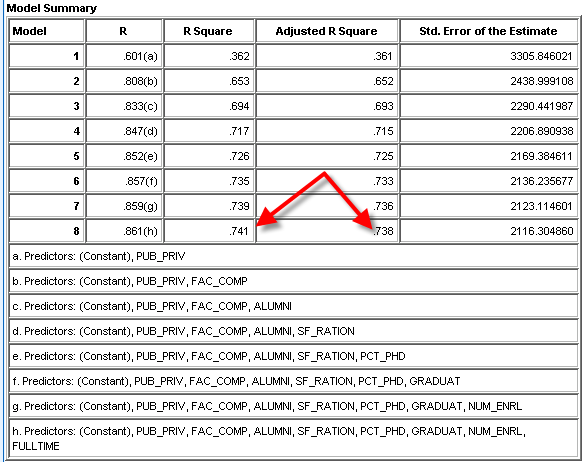




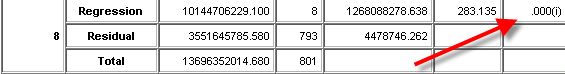
Edit the Regression node and select **Stepwise** as the method as shown below. Then run the Regression.

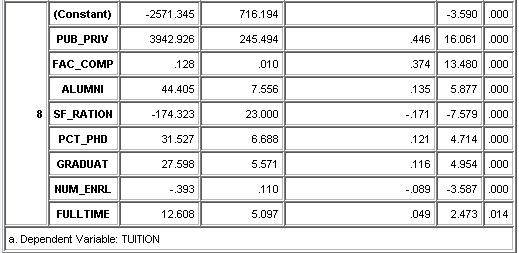


Click the Advanced tab to review the model results. Notice that the model has only 8 predictor variables instead of 10—the 2 non-significant variables, PCTTOP25 and ACCRATE are not in the model. The Model Summary shows the sequence of the predictors as they were added to the model—PUB\_PRIV then FAC\_COMP, and so on.



The R Square and Adjusted R Square values for the 8th and final model are .741 and .738 respectively—values as good as the model that had the 2 non-significant variables. The model is significant as well and all 8 variables are significant.





**How to save and exit**

IBM SPSS MODELER 14.2 allows many options for saving. For now, the stream is the most important part to save. Note that you can always tell is a stream has not been saved—it will have an asterisk to the right of its name. A stream is a file that can be saved and retrieved similar to most files and will run properly when retrieved —assuming the data sources remains in the same location.

Exit IBM SPSS MODELER 14.2 similar to most any other program, File🡪Exit.

**Recommendations:**

*Data and streams on your own computer:* My recommendation is to create a folder for data mining datasets – C:\DataMiningDataSets and a folder for SPSS MODELER 14.2 streams—C:\IBMSPSSStreams. Generally, develop all the SPSS MODELER 14.2 models referencing data in the DataMiningDataSets folder and save all the SPSS MODELER 14.2streams in the C:\SPSSStreams folder. This makes retrieving and running the models easy.

**Summary**

This short tutorial is designed to get you familiar with using SPSS MODELER 14.2and does not include discussion of the significant of the inputs, target variables and outputs. However, content on these values will be added as needed during the course. The most important learning objective at this time is to feel comfortable attaching to a data source, exploring the data, and building simple models.