



Microsoft SQL Server 2008 Data Mining Addins for Microsoft Office 2010

October 2013

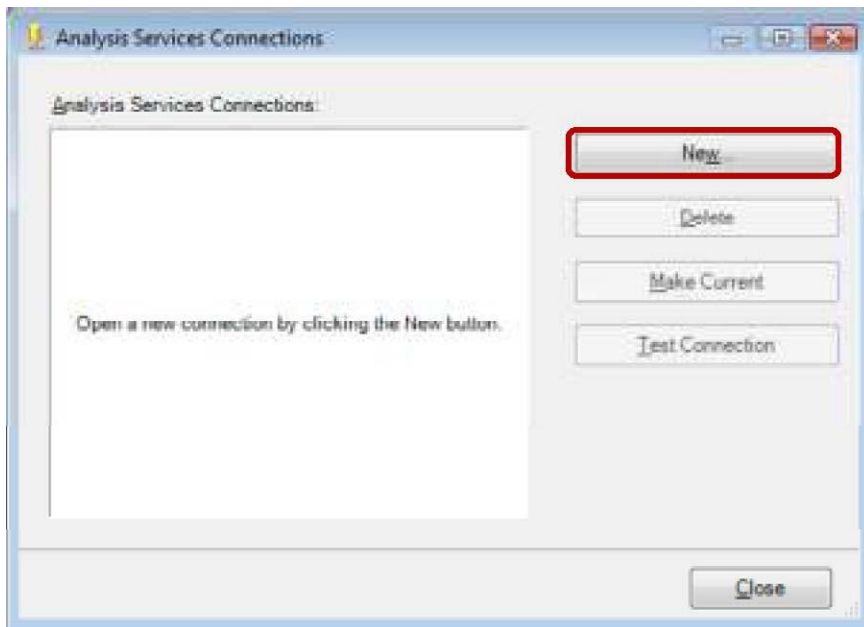
Data Mining Client for Excel 2010

The Data Mining Client addin enables you to go through the full data mining lifecycle within Excel by using your spreadsheet data. If you have the Data Mining Client addin installed, you should see the Data Mining ribbon when you launch MS Office Excel 2010 in your desktop. If you don't have Data Mining Client addin installed, you can download it for free and install it from Microsoft's website at: <http://www.microsoft.com/downloads/details.aspx?FamilyId=896A493A2502479594AE-E00632BA6DE7&displaylang=en> The addins installation wizard will walk you through the process.

The ribbon has Data Preparation, Data Modeling, Accuracy and Validation, Model Usage and Management tools, as shown below. In this document, we will go through all the tools in the Data Mining ribbon and see their functionality.

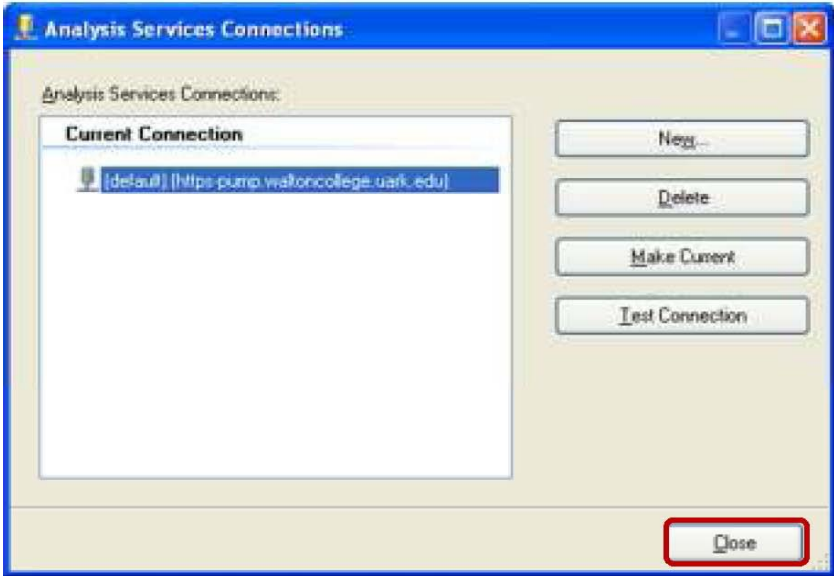
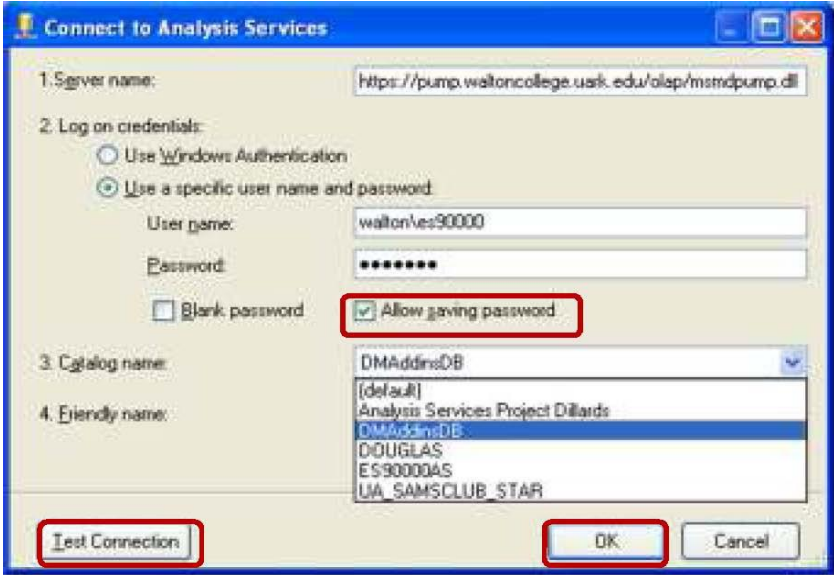


Now, before you can use these Data Mining tools, you need a connection to Analysis Services server. To connect to Analysis Services server, click on the No Connection button (shown above) in the ribbon under Data Mining. Analysis Services Connections screen comes up.

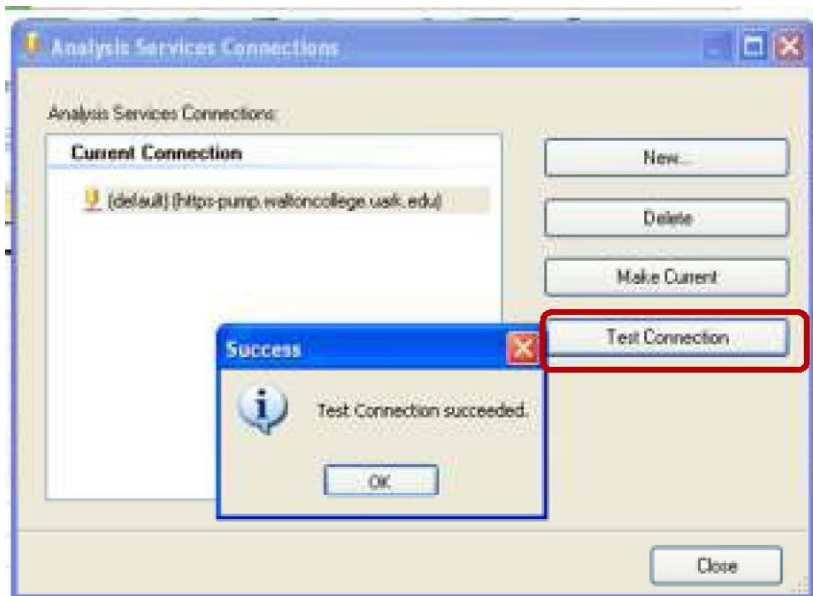


Open a new connection by clicking the New... button. Enter Server name as <http://ent-asrs.waltoncollege.uark.edu/olap/msmdpump.dll> (http connection to the Analysis Services server), select a specific user name and password as Log on credentials. Enter the credentials supplied to you by the University of Arkansas – (WALTON\ES#### and password. NOTE: If you are using Remote Desktop, simply select "Use Windows Authentication") and check the check box for Allow saving password. Then, select DMAddinsDB database from the dropdown list of Catalog names.

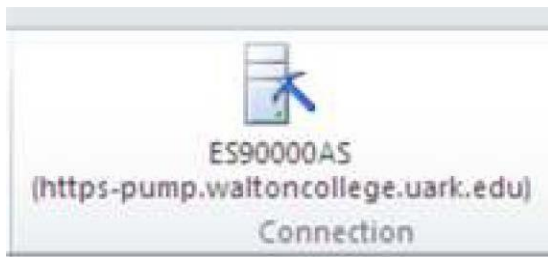
NOTE: The database named DMAddinsDB is created in the process of 'preparing' the server. DMAddinsDB database acts as a container for the mining models created by the Addins. It is a shared database that is supposed to be there to hold data **temporarily** while users connect to Analysis Services for the tools from Excel.



Then, click Test Connection to make sure a connection exists and click OK. Close the last screen that shows your Current Connection. Once the connection is made, you're all set to start using the Data Mining tools.



As a side note, once you make the connection, it will remain in your excel 2010 settings until you Delete it. In other words, next time you open an excel workbook, you will see the connection already made in the Connection section of the Data Mining ribbon in Excel (shown on the left). You just need to click the Connection and click Test Connection to ensure it is still valid. If not valid, click New... to establish a connection again.

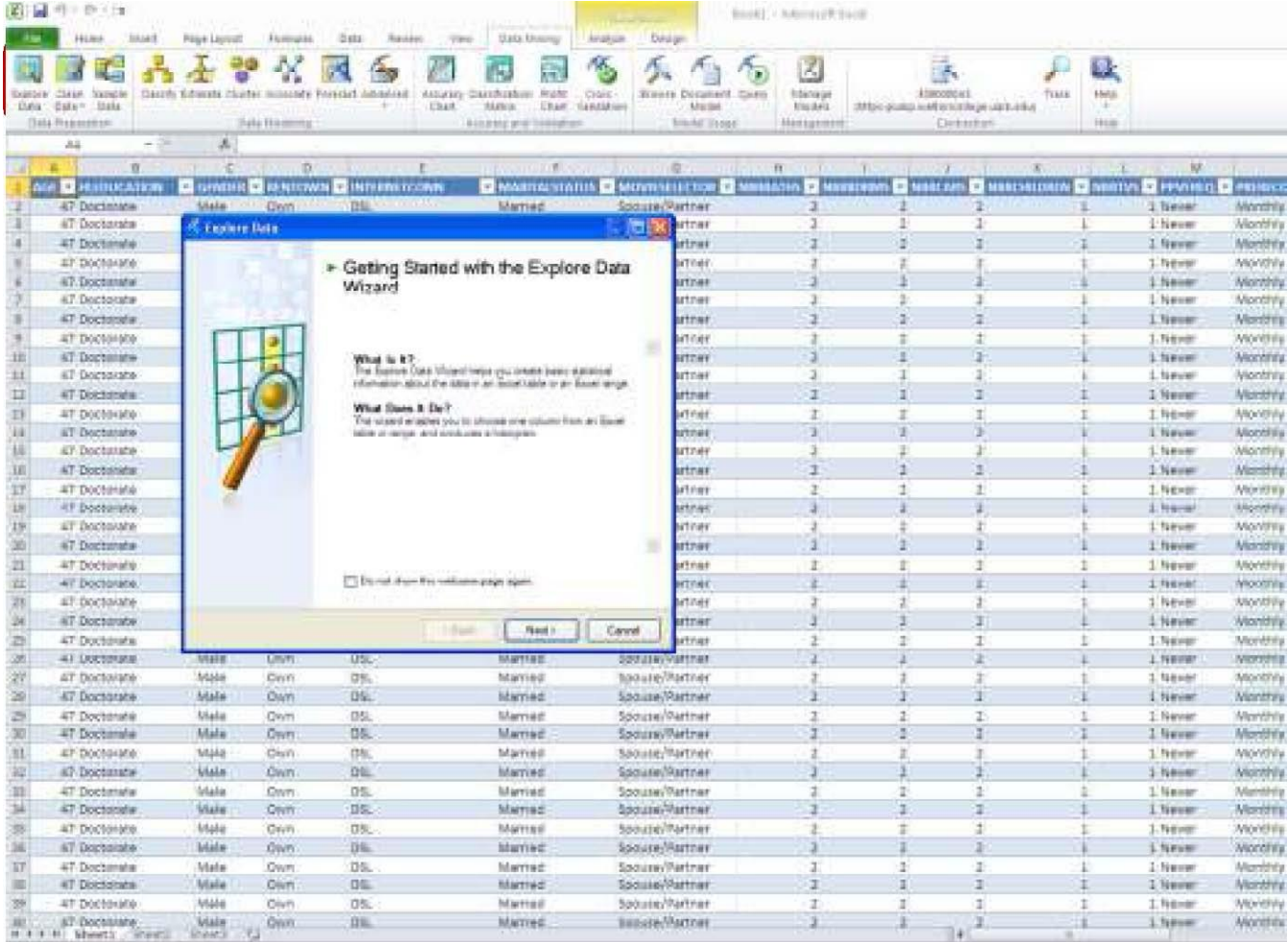


Data Preparation

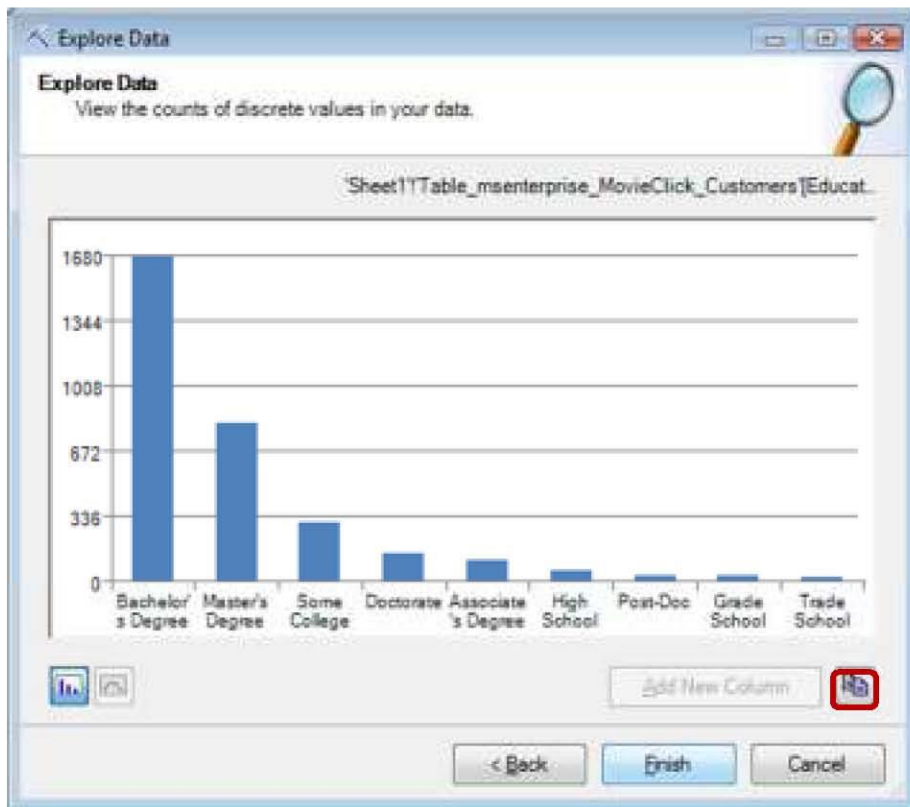
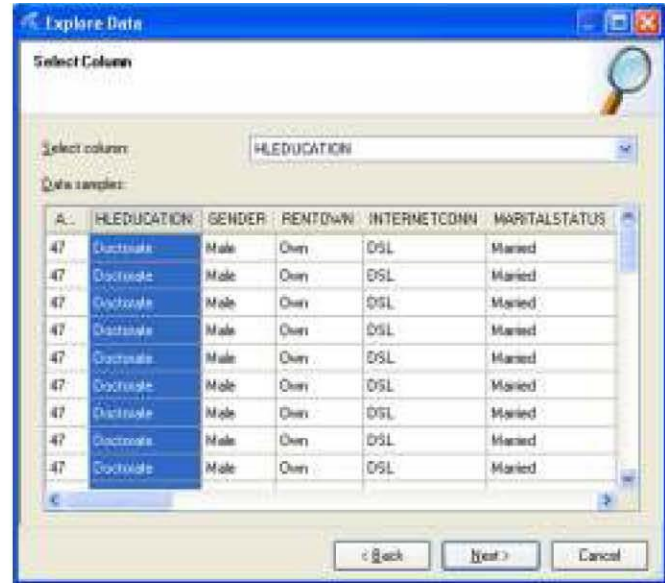
Selecting the right attributes from the source data and getting them into the right format for mining typically takes up a large percentage of the time in a typical data mining project. This section provides tools that address common data preparation needs for data mining: explore data, clean data and partition data.


Explore Data Visually plots the distribution of discrete and continuous values and possibly add groupings back to the source data. It is designed to help users identify and resolve imperfections in a data set. The Explore Data tool allows view histograms of numeric and nonnumeric data in your worksheet and group numeric data into equal size buckets. First you will need to load the data to excel workbook. This tool works only on data present in Excel and not on external data. To demonstrate how this tool works, I have imported the Customers table in the MovieClick database in MSEnterprise server to Excel workbook (refer to page 28 of this document on how to import data from SQL server to Excel workbook).

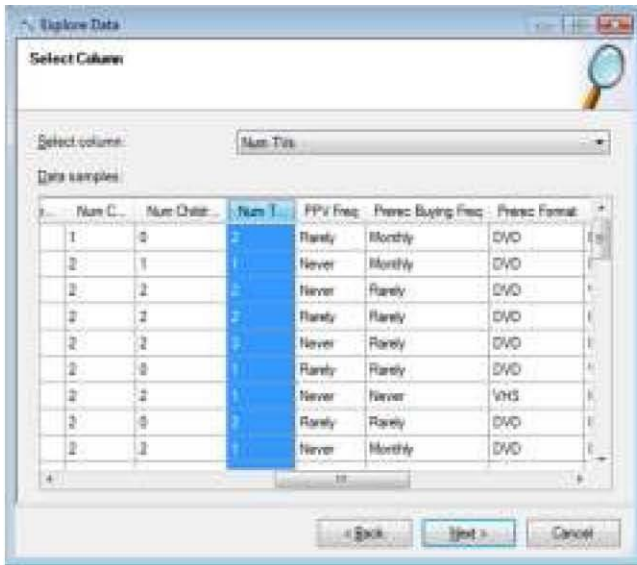
Click Explore Data to get started. Click Next in the getting started page.



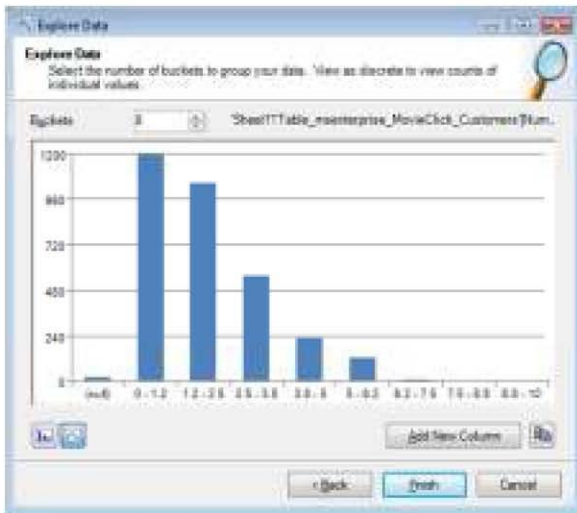
Select either the Table or Data range and click Next. Then, Select Column page comes up. You can select any numeric or nonnumeric column. In this example, select the nonnumeric Education Level column as shown in the screen below and Click Next to see its histogram.



At this point, you can click  to copy a bitmap of the chart to a clipboard. You can also click Finish to close the wizard or click Back to explore other columns. Let's click Back to explore a numeric column.

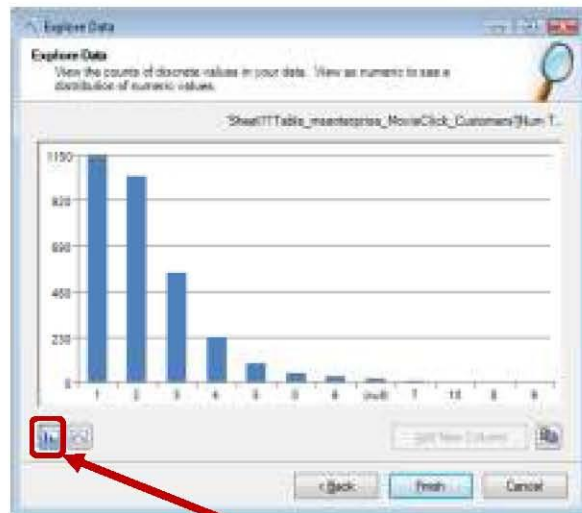
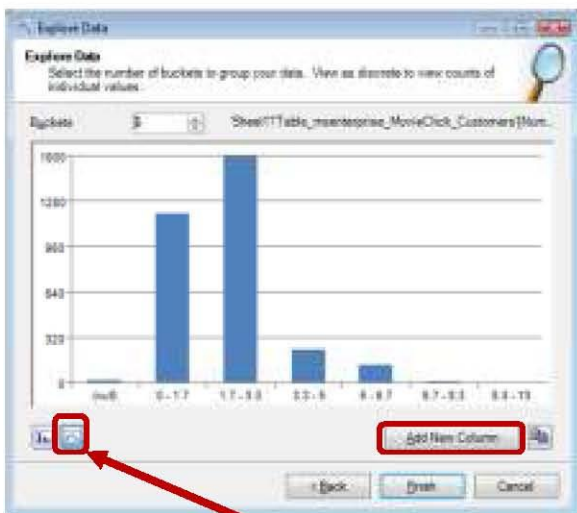


Now, let's explore the Num of TVs column. Here, on the left, a numeric column (Num of TVs) is selected.



When viewing numeric data, the Explore Data wizard shows the data grouped to buckets of equal ranges.

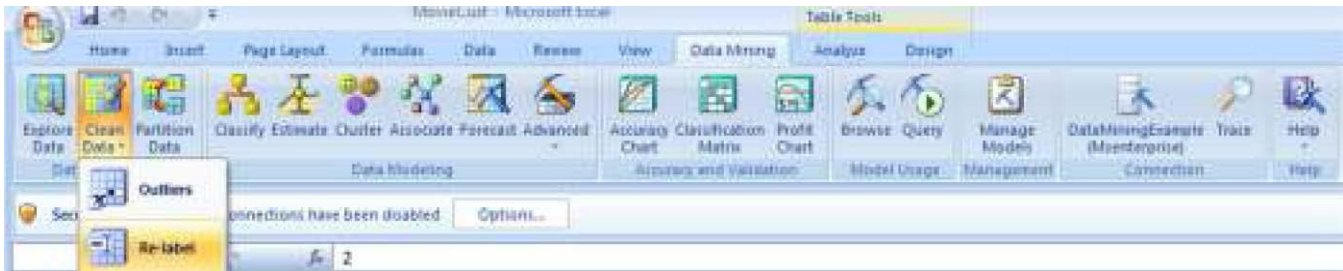
Note that in the screen on the left, the number of Buckets selected is 8. You can increase or decrease that number by changing the number in the buckets control. You can also toggle between View as Discrete and View as Numeric to change how the data is viewed.



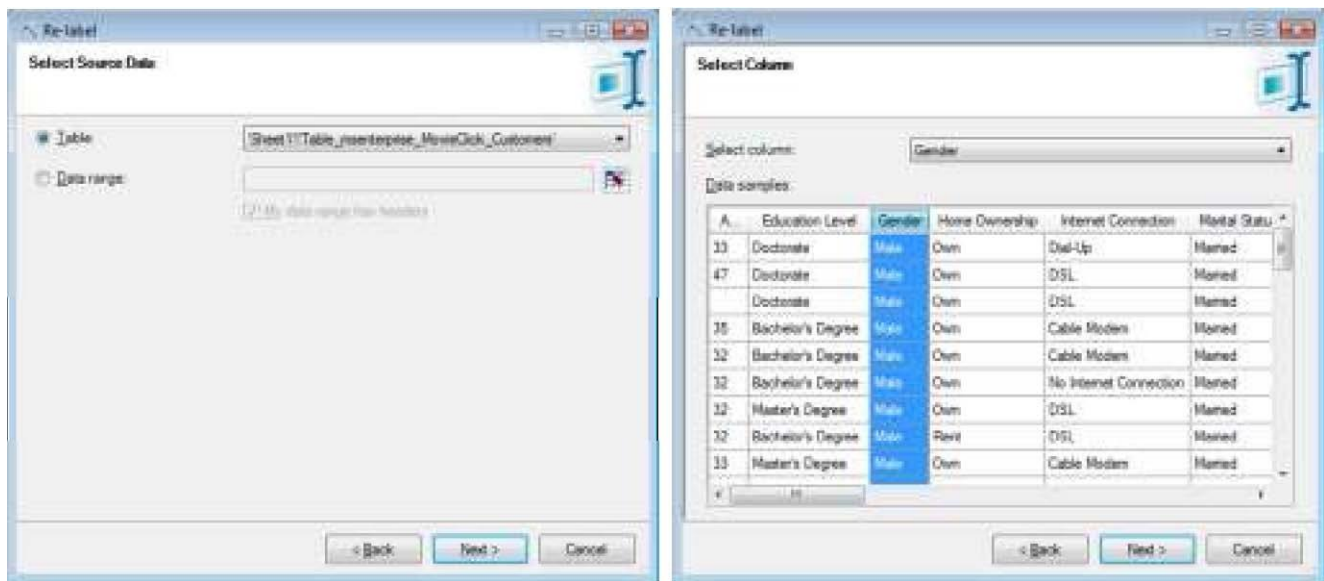
Notice that 'View as Numeric' is selected in the first screen, and 'View as Discrete' is selected in the second screen above. At this point, as explained above, you can click the copy button to copy a

bitmap of the chart to a clipboard, click Finish to close the wizard or click Back to explore other columns. In addition, if you want to use the bucketed data as a column in your worksheet for further analysis, you can click the Add New Column button which inserts a column of the appropriate ranges to your source data.

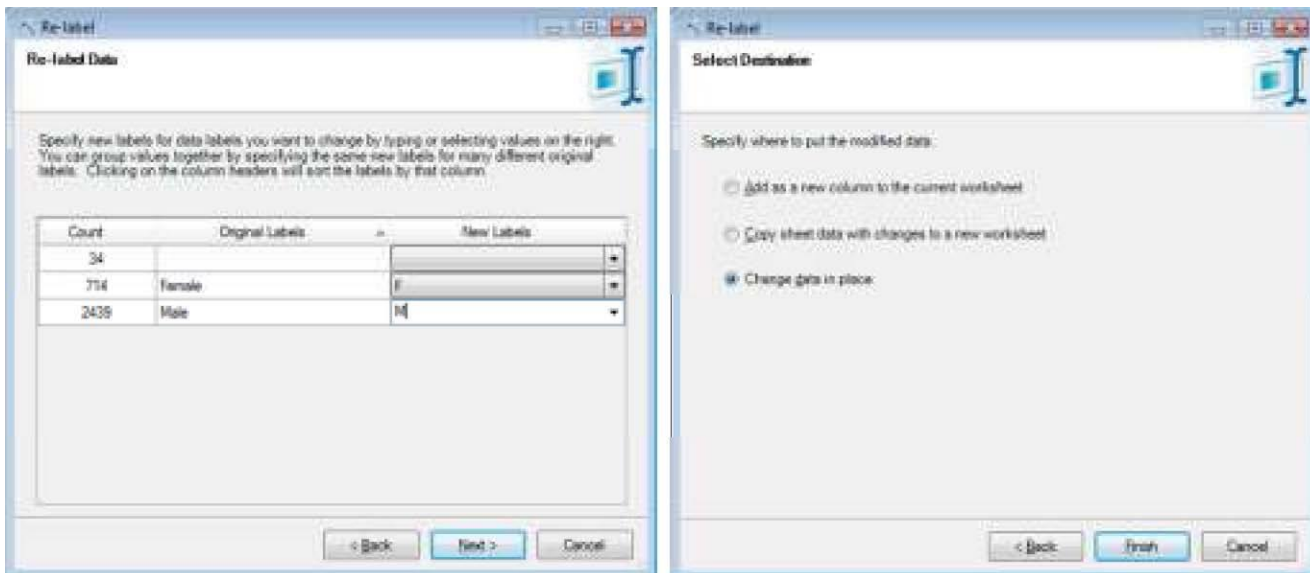
Clean Data Removes outliers and relabels discrete state values. For example, your source data may contain "Male" and "Female" for the Gender column and you prefer to use "M" and "F" when presenting the model results. For this example, let's use the MovieClickCustomers table from MSEnterprise we used for the above example. To start the relabel tool, click the Clean Data button and select Relabel.



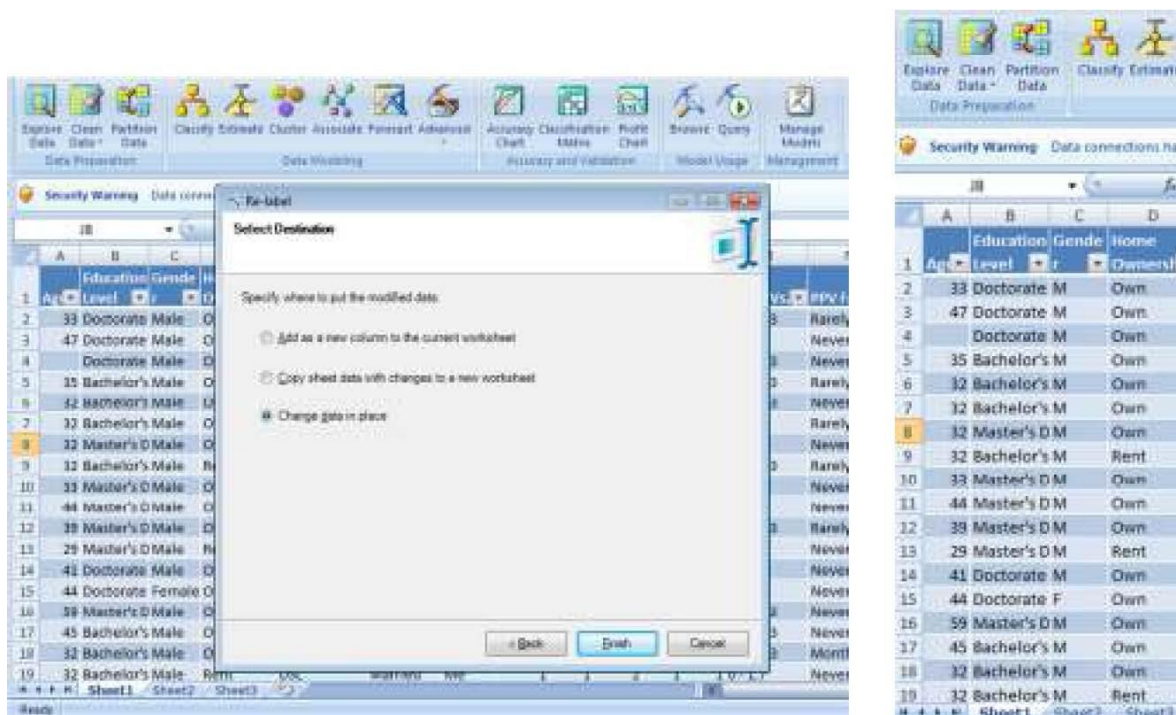
The getting started with the Relabel data wizard screen (not shown here) comes up; click Next. Select the data source and the column in the next screens.



You will see the original labels of your data for the column selected and you enter the new labels.



Select your destination and you will see your column relabeled as shown below.



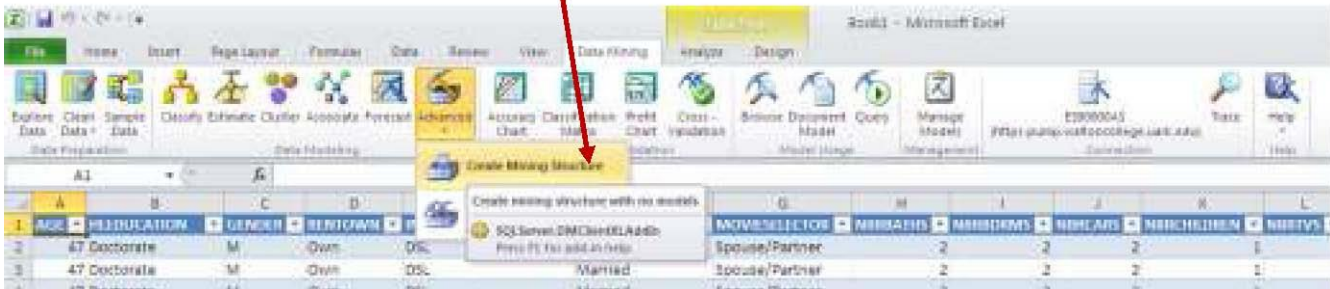
The Outliers tool detects and cleans up outliers from your data. To start the remove Outliers tool, click the Clean Data button, select Outliers and follow the instructions on the screen.

Partition Data Splits the source data into training and test sets, takes a random sample of the input data or perform oversampling to adjust for skewed distributions.

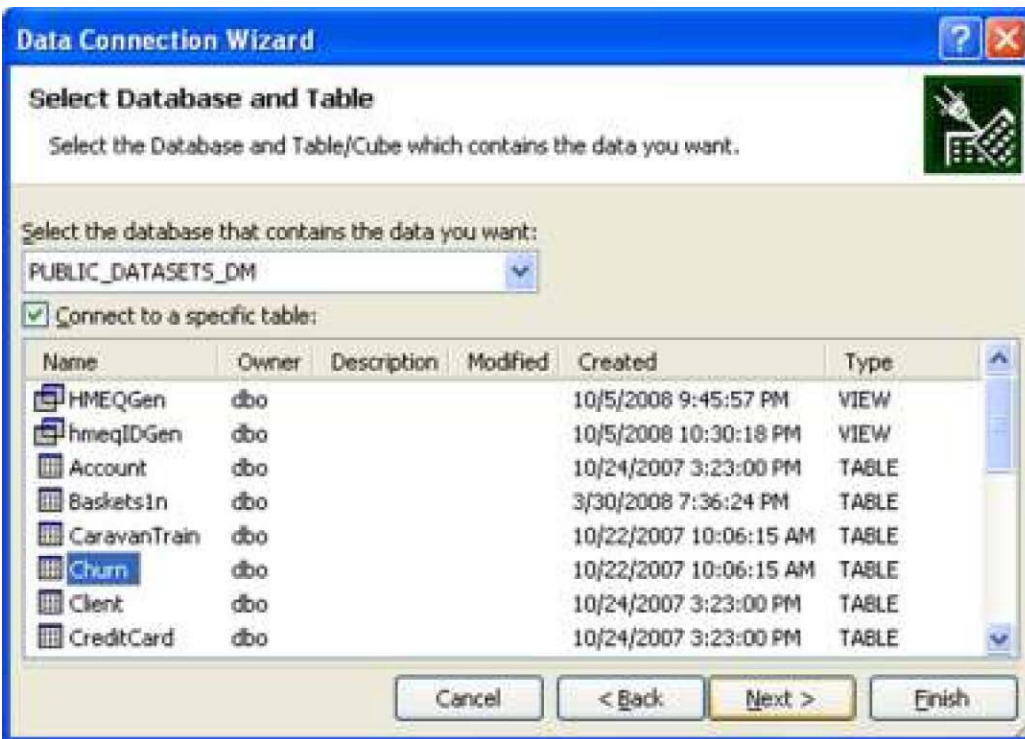
Data Modeling

This section covers the actual model definition and processing. It provides wizards that help you easily build common types of mining models without worrying about the actual mining algorithms and associated parameters supported on the server. Also included in this section are advanced options that allow the user to pick the exact mining algorithm and tweak additional parameters. To create a mining model, click Create Mining Structure under Advanced in the Data Modeling section of the Data Mining ribbon.

section of the Data Mining ribbon.

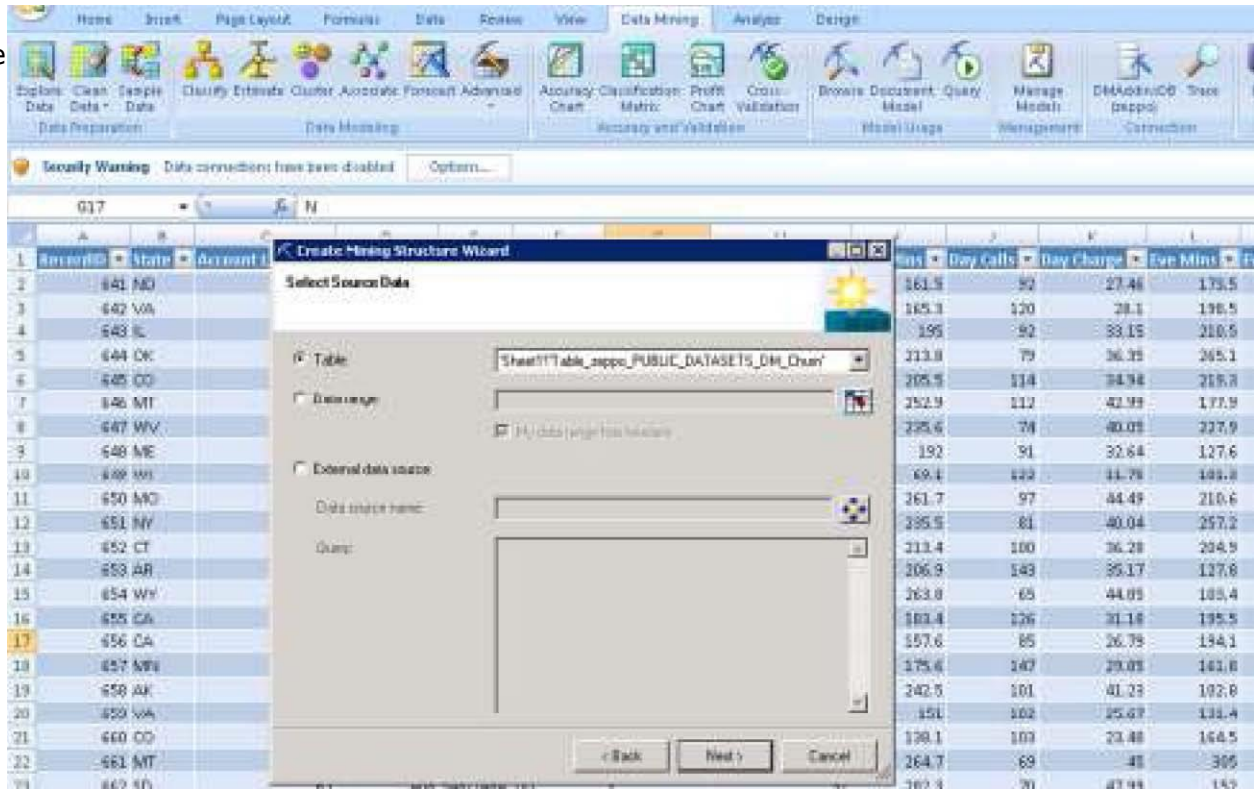


Click the Next button in the welcome screen (not shown here). We will use the churn table for this example. The churn table is imported to excel from SQL Server 2008. Refer to steps on Import Data on Page 5. Select PUBLIC_DATASETS_DM database and Churn table when asked for the database and table.

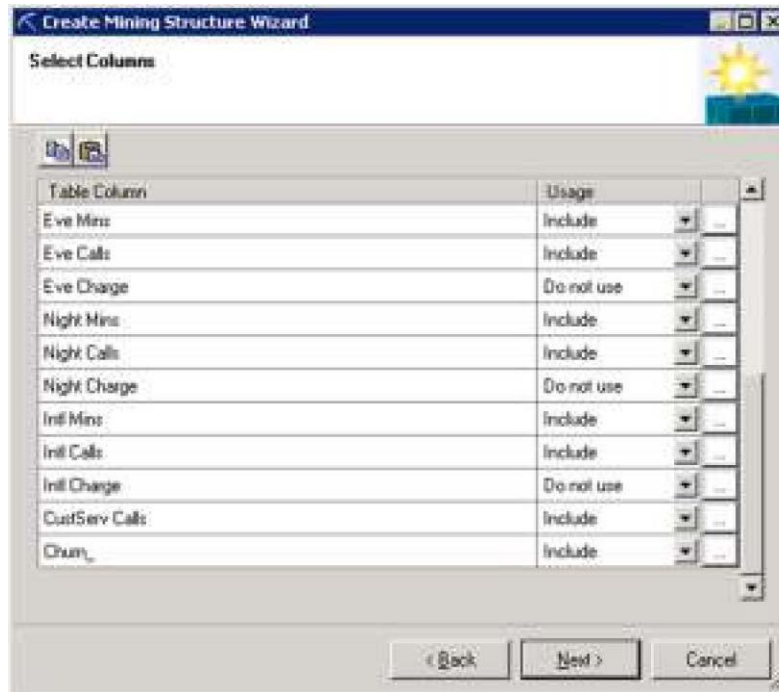


Select the table with the data to be create a Mining Structure in the Select Source Data page and click the Next button. Use the churn table.

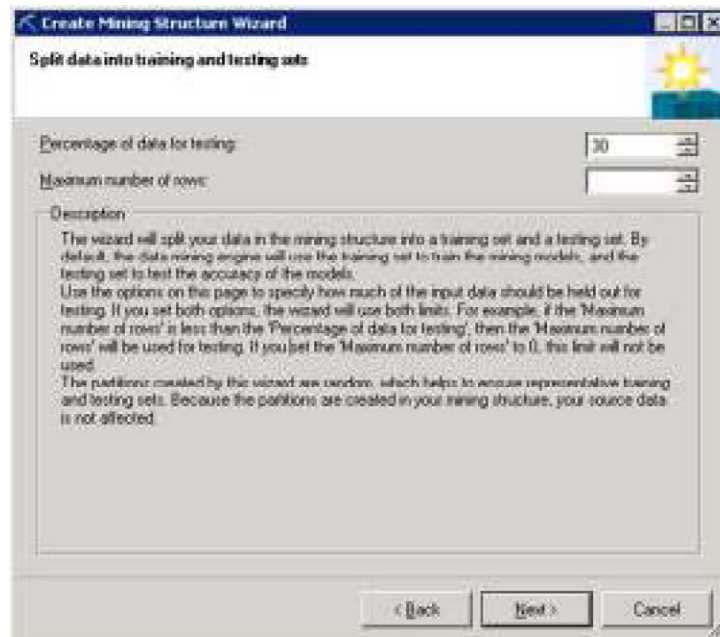
On the next



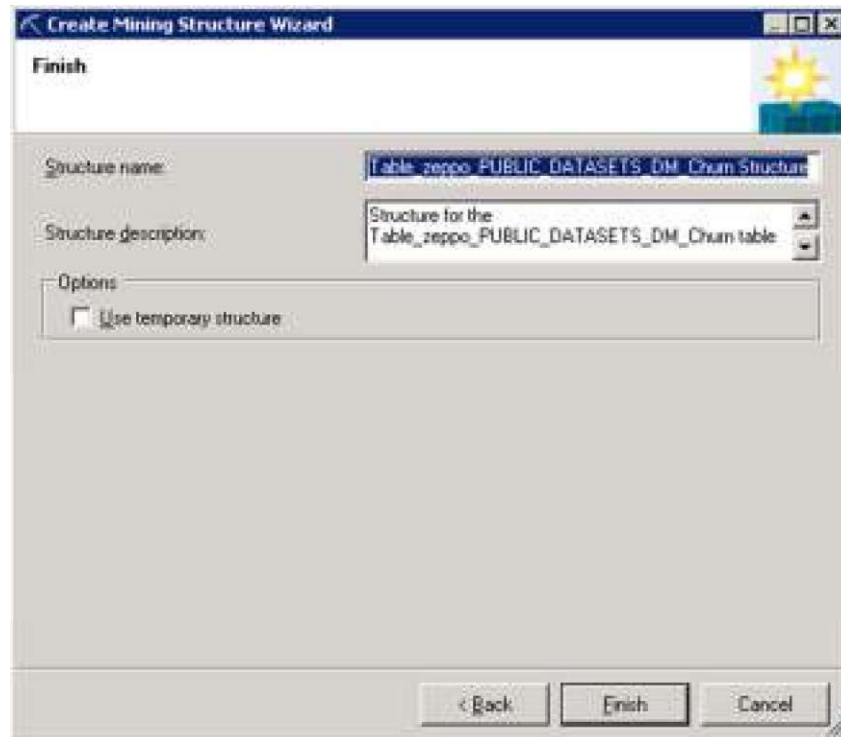
screen, select which columns to use and click Next. Mark RecordID as Key and do not use columns related to Charge, State, Area Code and Phone. From exploratory data analysis (not shown here), it was determined that the variables (columns) of **State**, **Area Code** and **Phone** contained bad data. Also, all the columns related to Charge were perfectly correlated to the corresponding Mins (Minutes) column. So, none of the Charge columns will be used in the analysis. Mark all these columns as Do not use. Mark the remaining columns as Include. Click Next after selections are made.



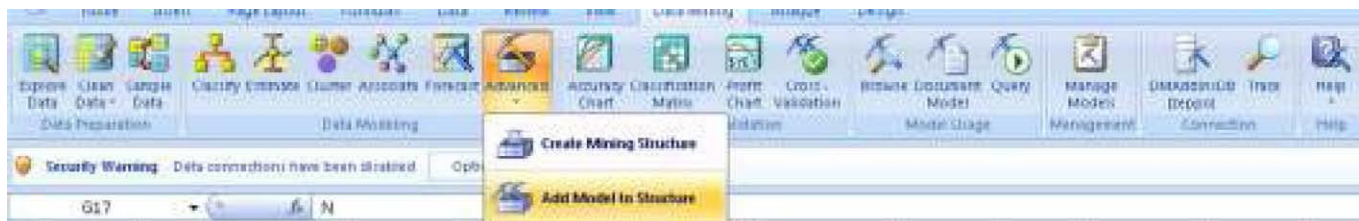
On the next screen you are given the option of selecting what percentage of the data should be used for testing. Keep the default settings and click Next.



Click Finish on the next screen to create the Mining structure.

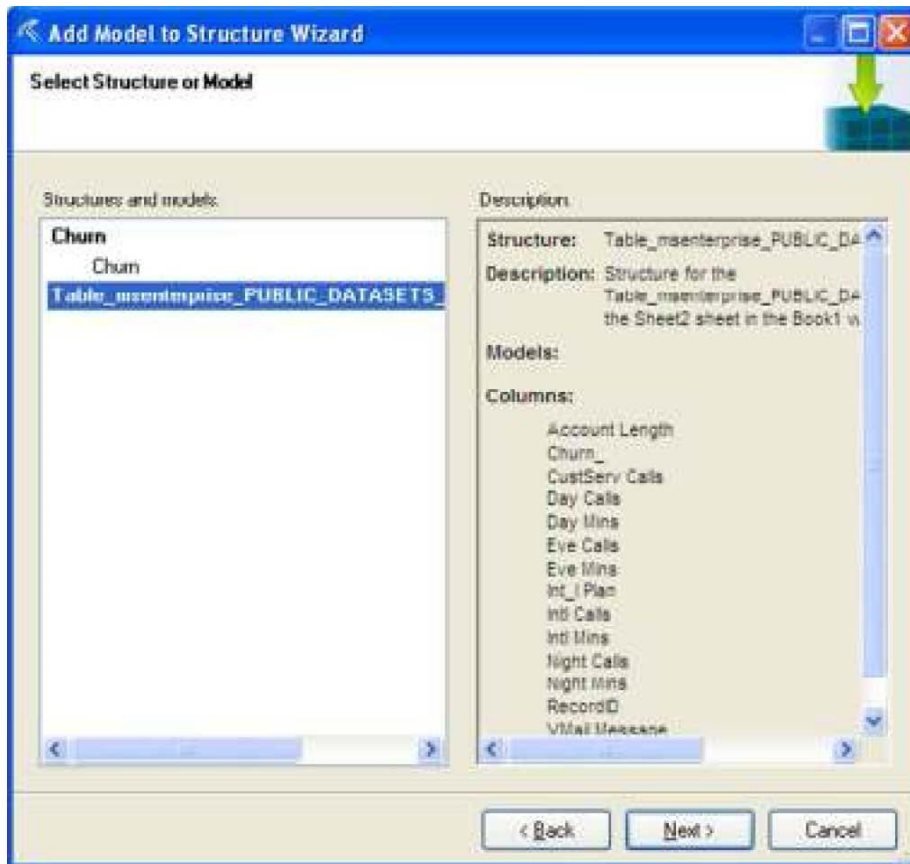


To add a mining model to the structure that was previously created, select the Add Model to Structure option under Advanced in the Data Modeling section of the Data Mining ribbon.

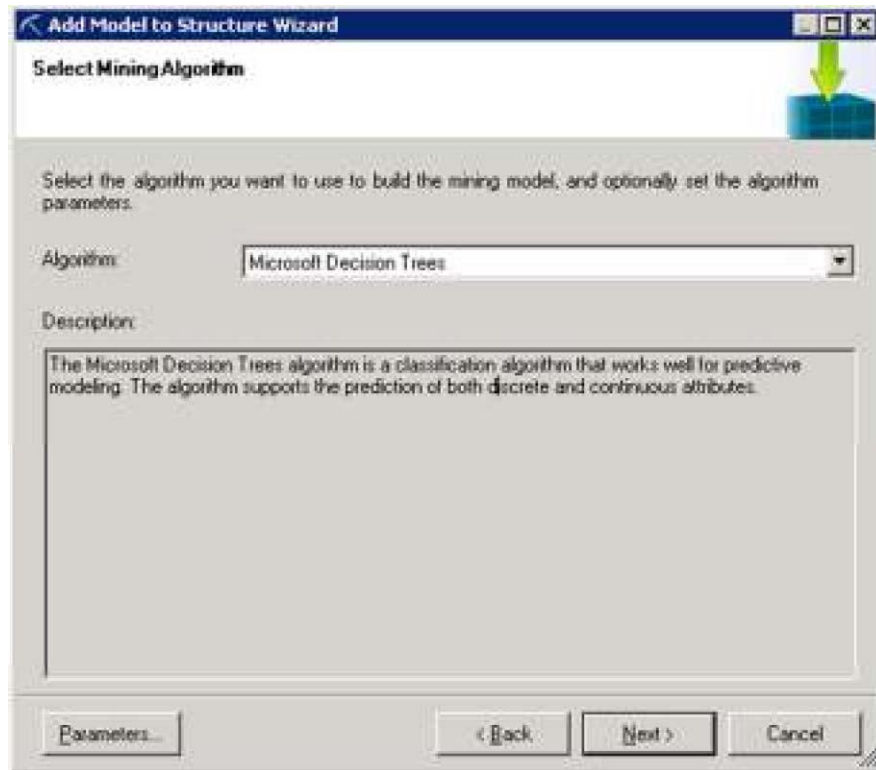


Select the structure which was created above as shown below and click Next.

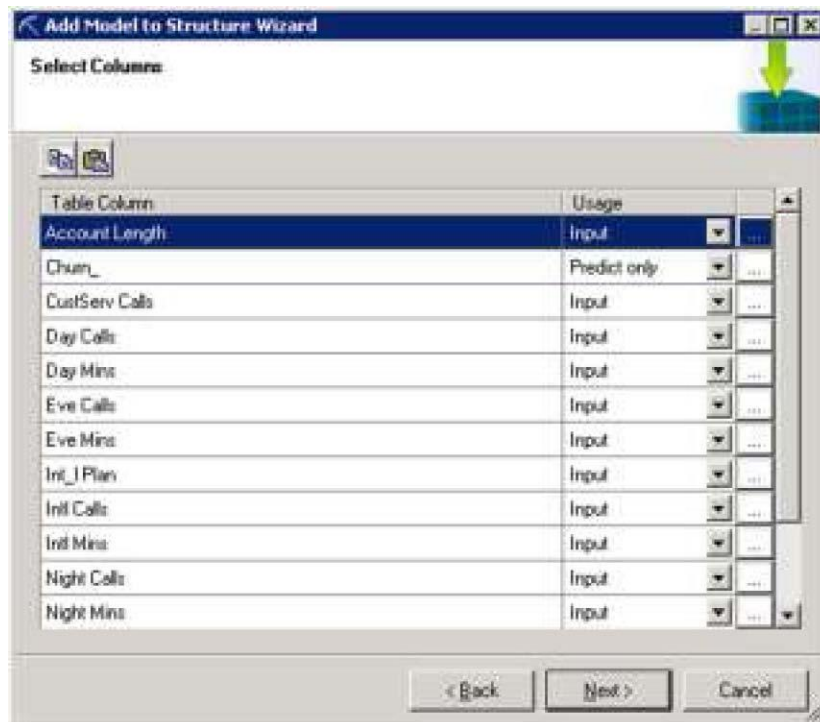
Select Decision Trees from the list of Algorithms Next.



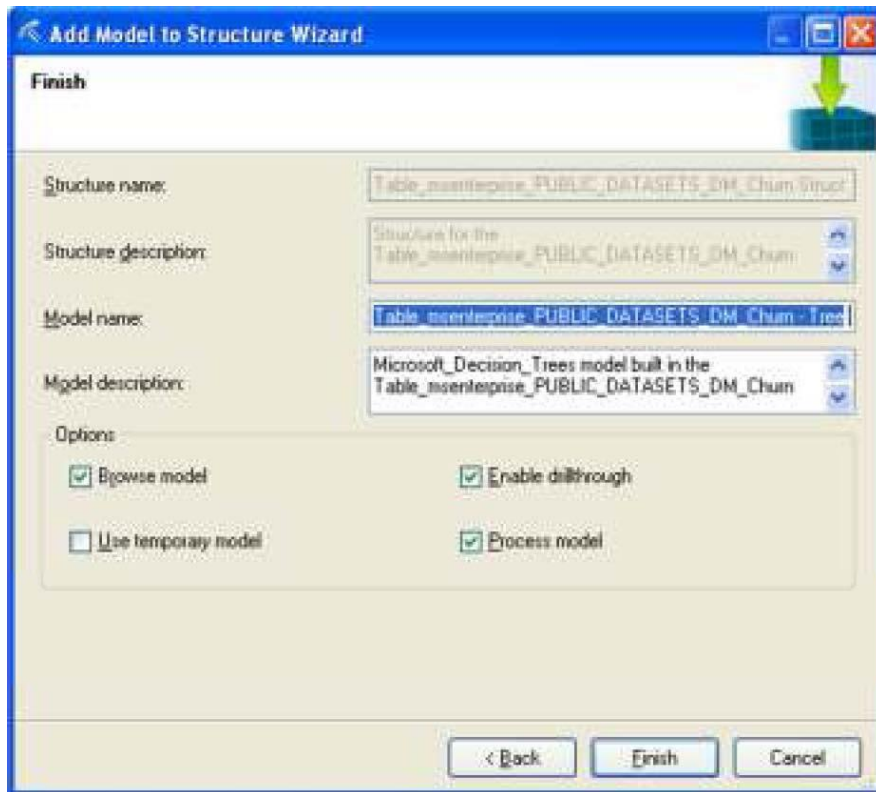
Microsoft Trees from dropdown Mining and click



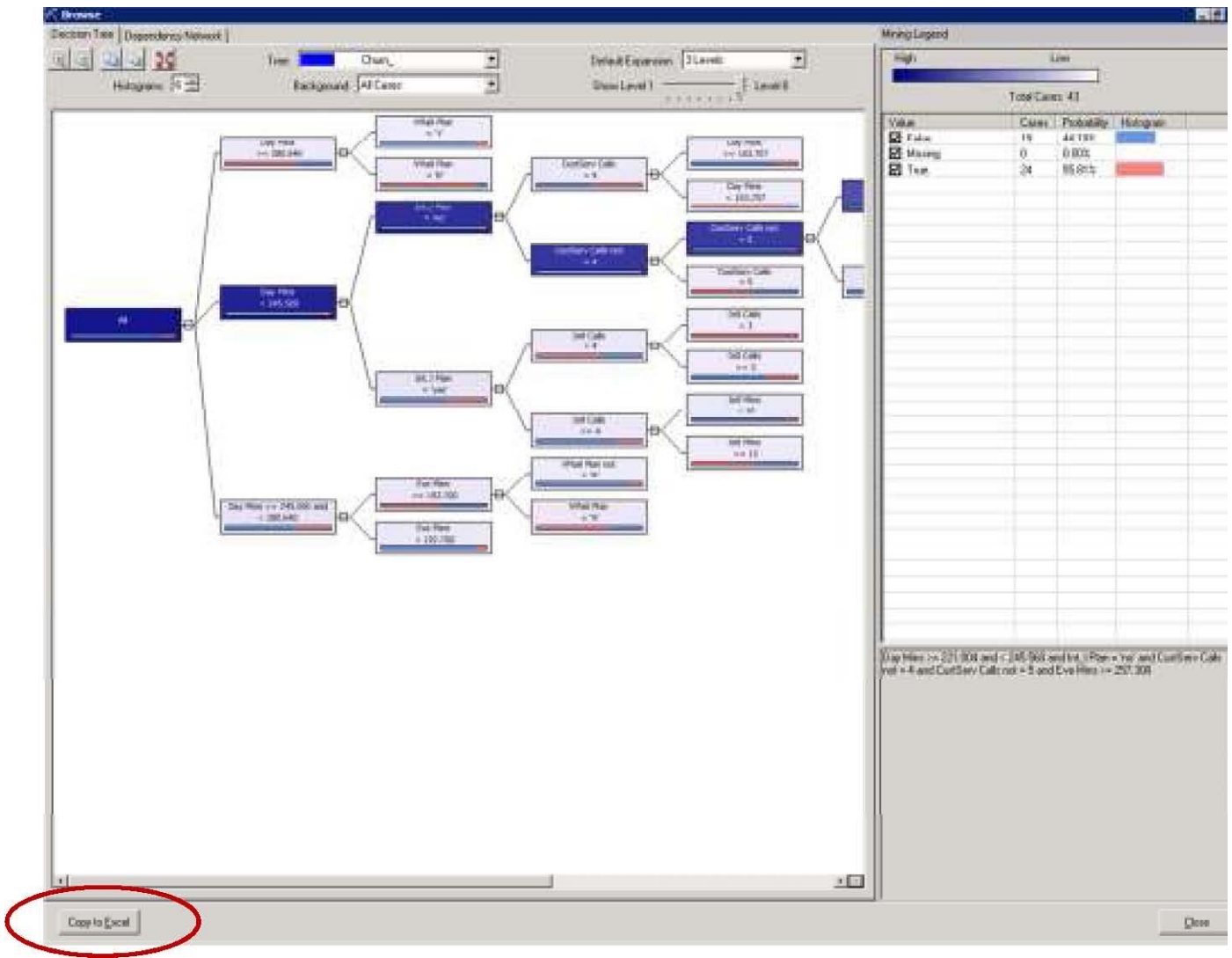
On the next screen, select the type of usage for the columns that were included. Mark RecordID as Key and Churn? column as Predict Only. Notice that the columns Charge, State, Area Code and Phone which were marked as "Do not use" no longer appear on the select columns screen. Leave the remaining columns as Input and click Next.



Click Finish on the next screen to add the model to the mining structure. Optionally, you can give the model any name or leave the default name for the model.



The Decision Tree Page comes up. You can copy the diagram to Excel by clicking the Copy to Excel button in the left bottom section of the screen.



You can also add another model to existing mining structure by clicking Add Model to Structure, under under the Advanced toolbar.



the Advanced toolbar.

Accuracy and Validation

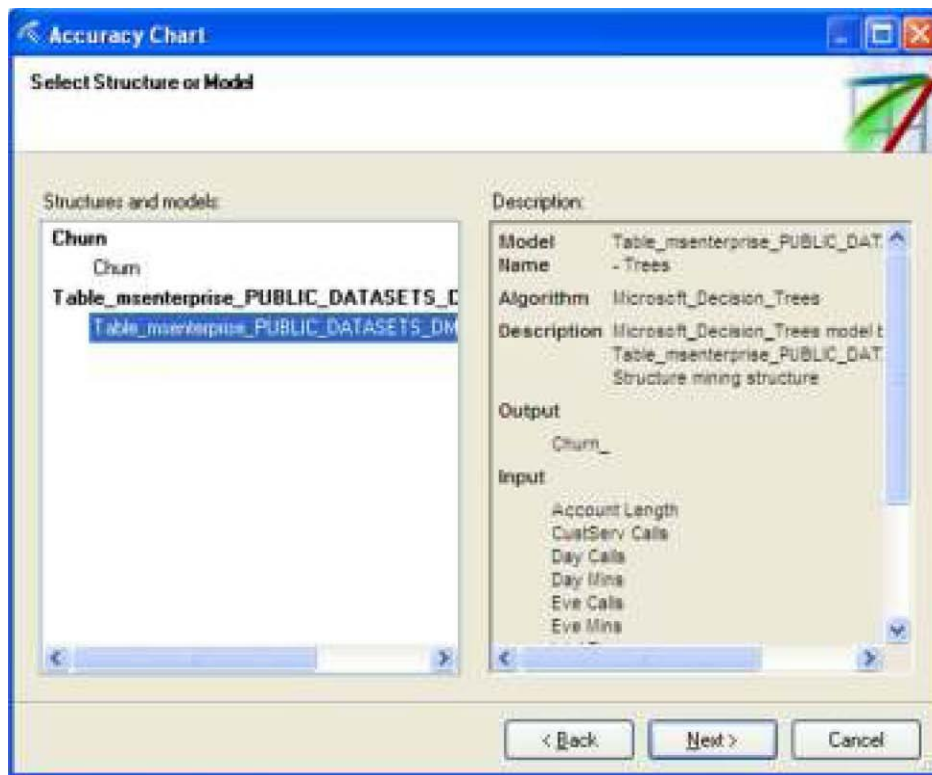
This section includes charts for validating and verifying the accuracy of the mining models. The three charts provided are: Accuracy Chart, Classification Matrix and Profit Chart.

Accuracy Chart – Evaluates the performance of the model against test data by drawing a lift chart for classification models and a scatter plot for estimation models. To see the Accuracy Chart, click the Accuracy Chart in the Accuracy and Validation Section of the Data Mining ribbon.

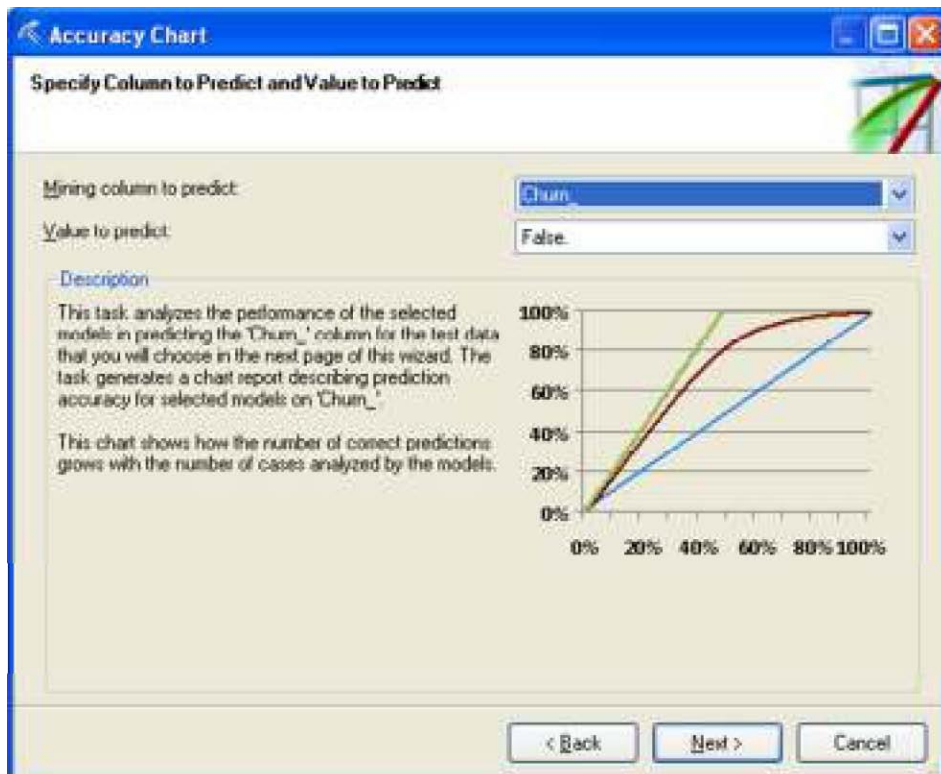


The getting Started page comes up. Click Next to select the existing model. Select the Trees model created above and click Next.

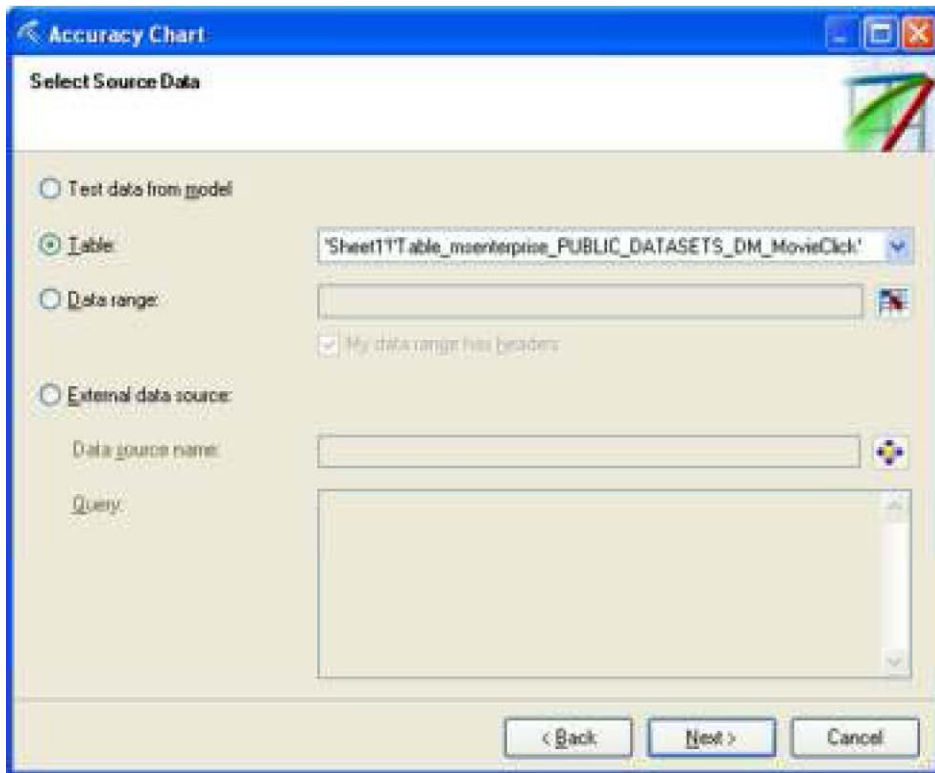




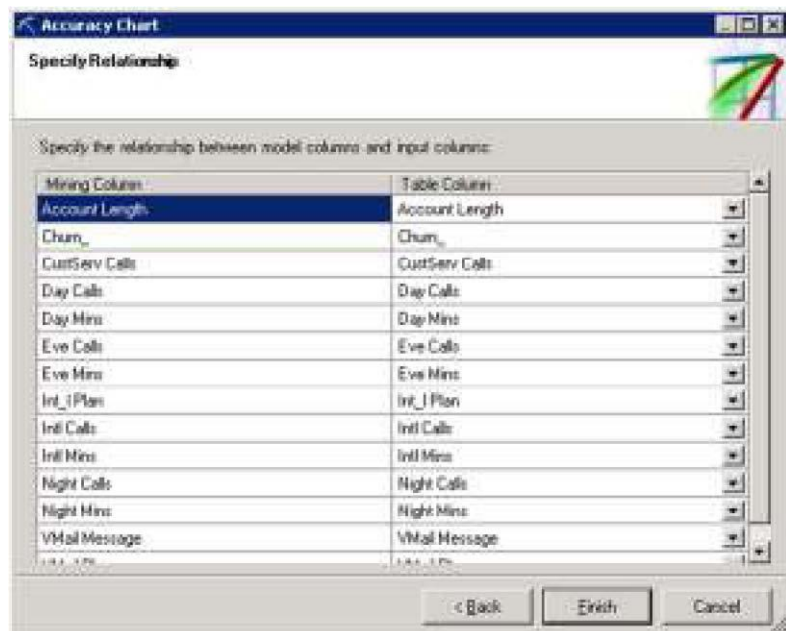
Specify Column to Predict (in this example, churn_) and Value to Predict (False) and click Next.



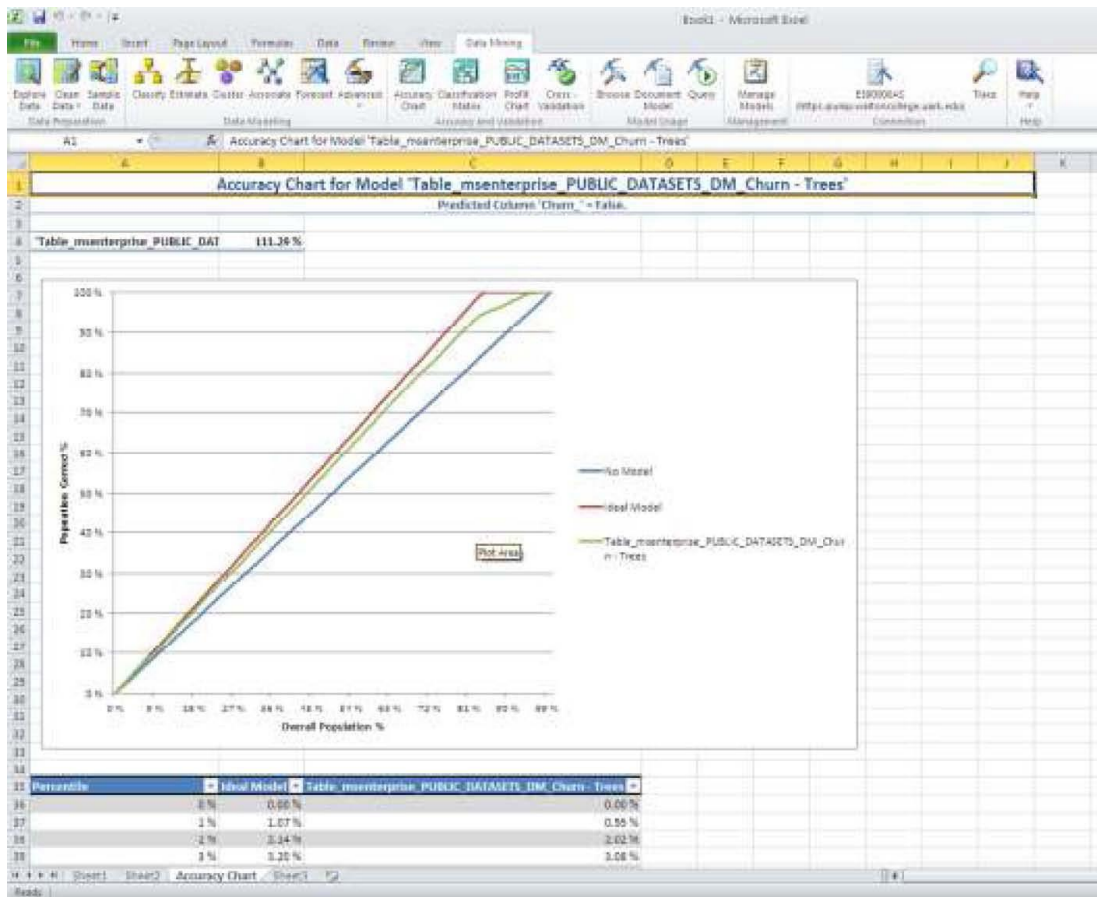
Select Source Data and click Next.



Accept the defaults in the Specify Relationship page and click Finish.



The chart will be in a new tab called 'Accuracy Chart' as shown below.

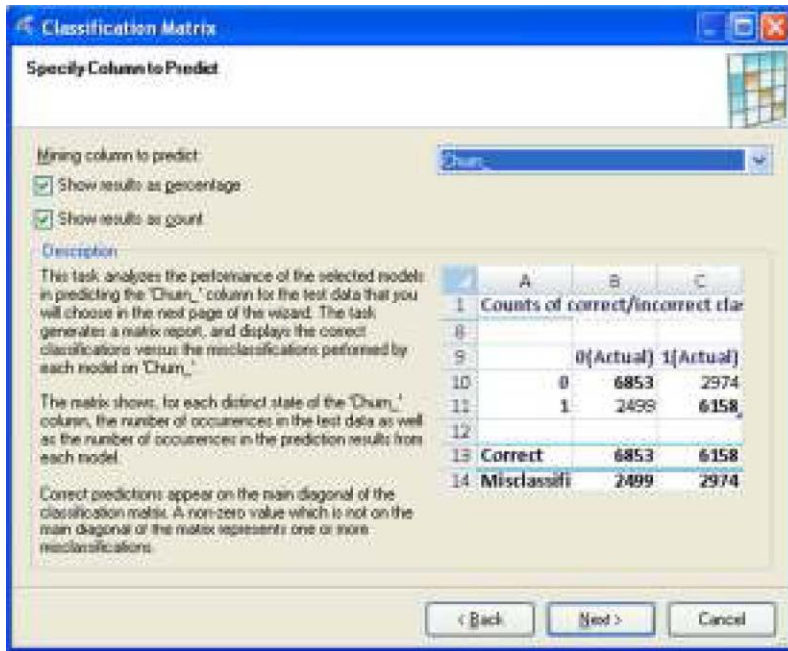


Classification Matrix Displays a matrix of correct and incorrect classifications by evaluating your model against test data. To create Classification Matrix, click the Classification Matrix button in the Accuracy and Validation Section of the Data Mining ribbon.

Accuracy and Validation Section of the Data Mining ribbon.



Click Next in the getting Started page (not shown) and select the ...Trees model (the model created in the process before) from the Select Model screen and click Next.

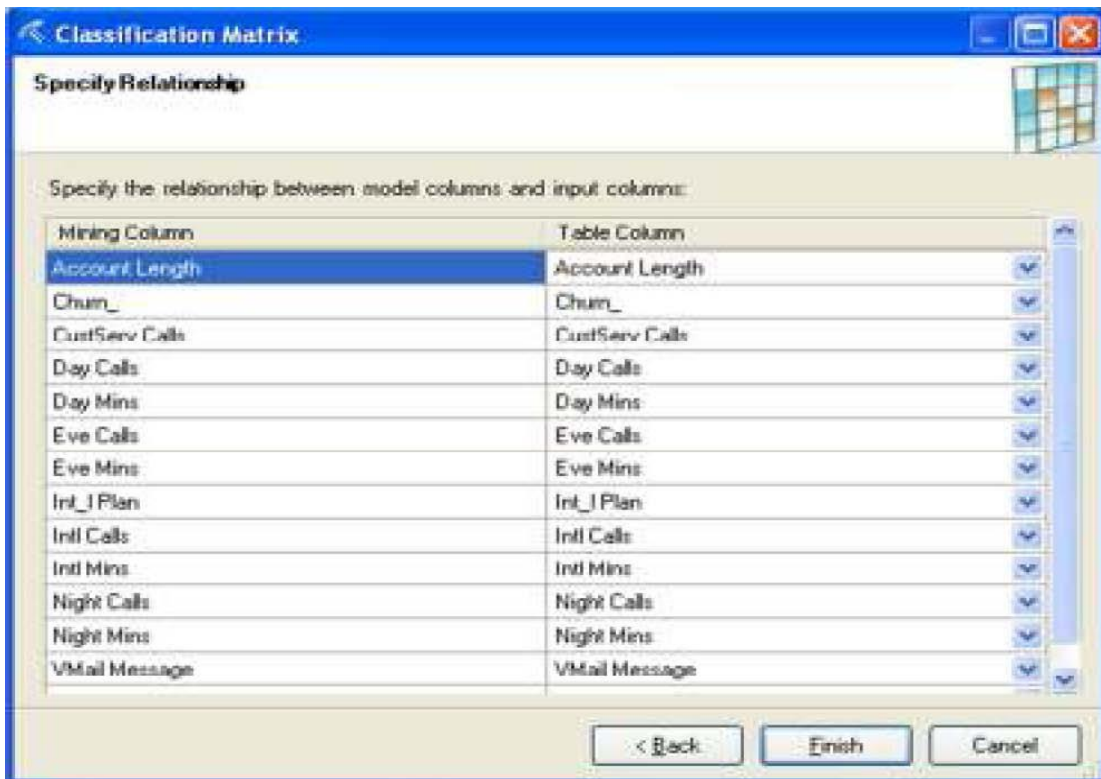


Specify Column to predict (churn_) in the next page and click Next.

Choose the churn table in the Select Source Data wizard page and click Next.



Accept the default values in the Specify Relationship page and click Finish.



The Classification Matrix output will open a new tab in the same spreadsheet and will look like the screenshot below.

Book1 - Microsoft Excel

File Home Insert Page Layout Formulas Data Review View Data Mining

Options Data Tables Data Tables Data Tables
 Data Presentation Data Presentation Data Presentation
 Data Mining
 Accuracy Chart Accuracy Chart Accuracy Chart
 Accuracy and Validation Accuracy and Validation Accuracy and Validation
 Model Usage Model Usage Model Usage
 Manage Models Manage Models Manage Models
 Conversion Conversion Conversion
 Tools Help

A1 | A | B | C | D | E | F

1 **Counts of correct/incorrect classification for model 'Table_mseenterprise_PUBLIC_DATASETS_DM_Churn - Trees'**

2 Predicted Column 'Churn'

3 Columns correspond to actual values

4 Rows correspond to predicted values

5

6 Model name:	Table_mseenterprise_PUBLIC_DATASETS_DM_Churn - Trees	Table_mseenterprise_PUBLIC_DATASETS_DM_Churn - Trees
7 Total correct:	92.14%	8071
8 Total misclassified:	7.86%	282

9

10 Results as Percentages for Model 'Table_mseenterprise_PUBLIC_DATASETS_DM_Churn - Trees'

11	← Table (Actual)	← True (Actual)	
12 False	94.67%		22.77%
13 True	5.33%		77.23%
14			
15 Correct	94.67%		77.23%
16 Misclassified	5.33%		22.77%

17

18 Results as Counts for Model 'Table_mseenterprise_PUBLIC_DATASETS_DM_Churn - Trees'

19	← Table (Actual)	← True (Actual)	
20 False	2698		110
21 True	152		871
22			
23 Correct	2698		871
24 Misclassified	152		110

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

43

44

45

46

47

48

49

50

51

52

53

54

55

56

57

58

59

60

61

62

63

64

65

66

67

68

69

70

71

72

73

74

75

76

77

78

79

80

81

82

83

84

85

86

87

88

89

90

91

92

93

94

95

96

97

98

99

100

101

102

103

104

105

106

107

108

109

110

111

112

113

114

115

116

117

118

119

120

121

122

123

124

125

126

127

128

129

130

131

132

133

134

135

136

137

138

139

140

141

142

143

144

145

146

147

148

149

150

151

152

153

154

155

156

157

158

159

160

161

162

163

164

165

166

167

168

169

170

171

172

173

174

175

176

177

178

179

180

181

182

183

184

185

186

187

188

189

190

191

192

193

194

195

196

197

198

199

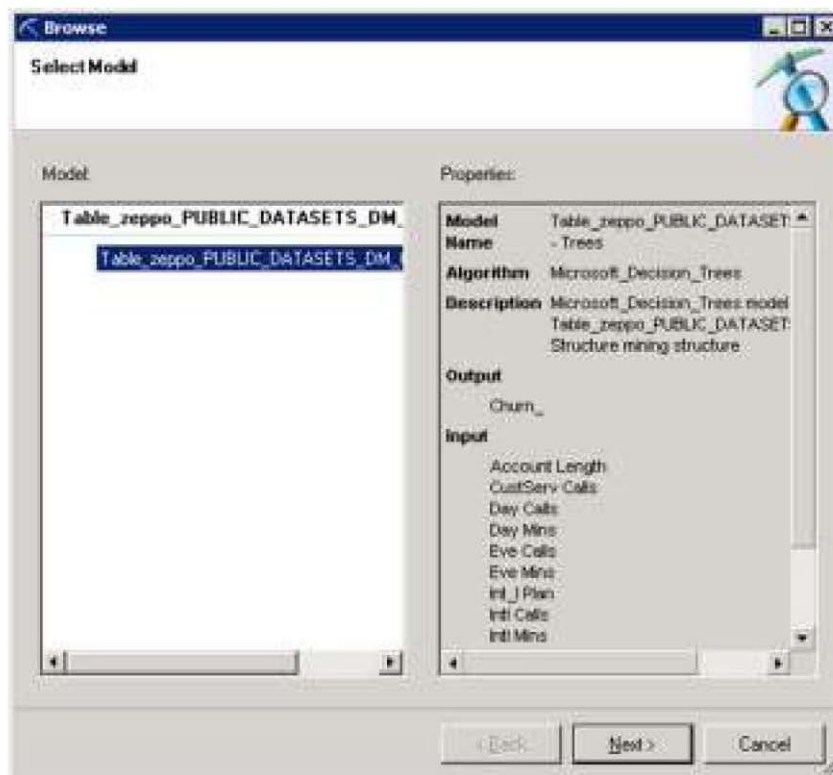
200

Profit Chart – Graphically models profit for targeted campaigns based on usersupplied parameters for cost of existing mining models.

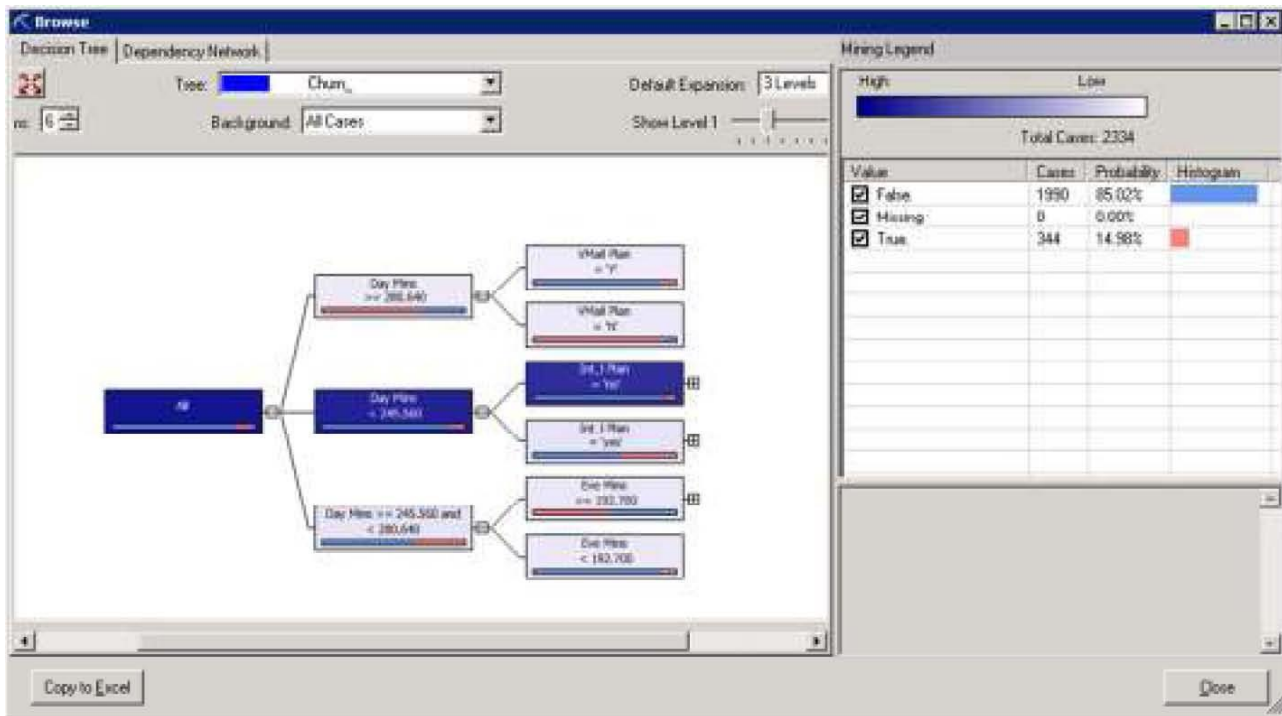
Model Usage

This section covers the two standard tasks that you would perform with trained mining models: Browse and Query.

Browse – Explores the patterns and rules learned by the mining algorithm from the training data. This “mining model content” is visualized in different ways depending on the type of model you’re browsing. To see the functionality of this tool, click **Browse** in the **Model Usage** section of the **Data Mining** ribbon and select the model you would like to browse.



Below is an example of the visualization of a decision tree model for the Trees model created before.



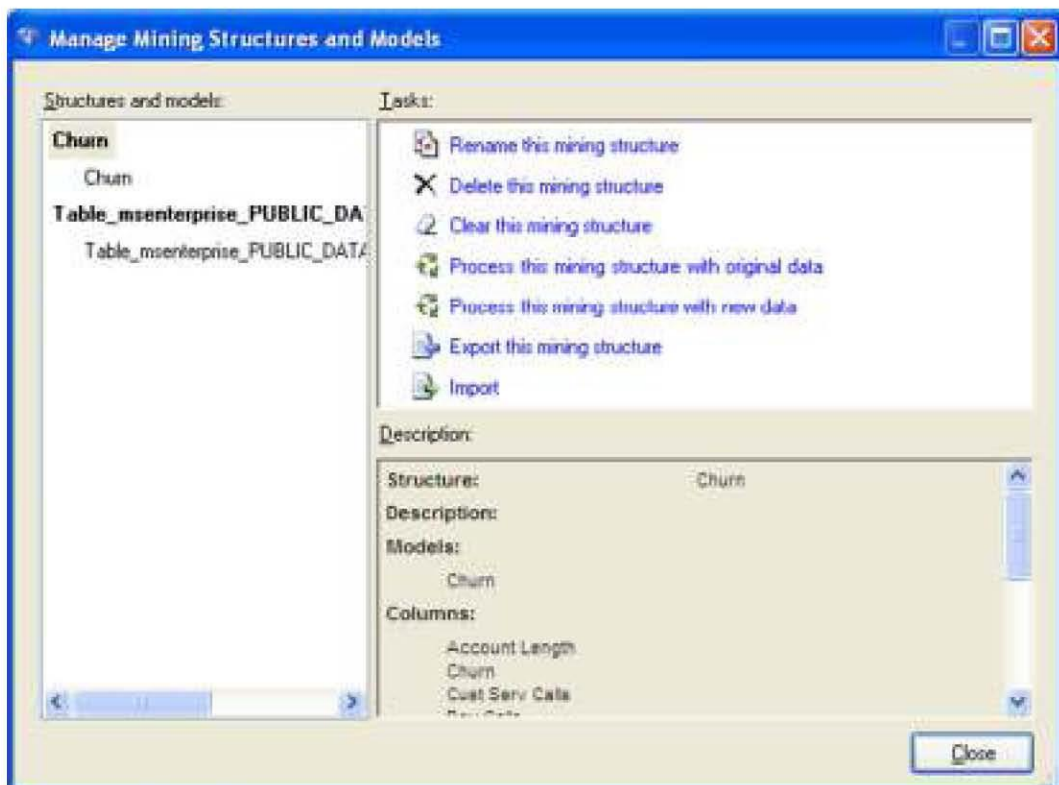
Query – Uses the trained model to make

predictions on new data. The Query task supports a wizard for building simple queries as well as an advanced editor where you can use DMX templates to build queries or manually type in the DMX statement.

Management

This gives the ability to manage existing models in the Analysis Services database you are connected to. You can rename, delete, clear, reprocess, export, or import mining structures and models as shown below.

Connection



As we saw at the beginning of this document, the Connection button allows to create and manage connections to Analysis Services databases – a connection to an Analysis Services database is required before you can run any of the analysis tools described above.

Trace

The Trace button allows you to trace the commands that are sent by the data mining addin(s) to the Analysis Services instance you're currently connected to.

Help

The Help button provides access to the documentation included with the addins as well as the Getting Started wizard and online tutorials.

Sources include:

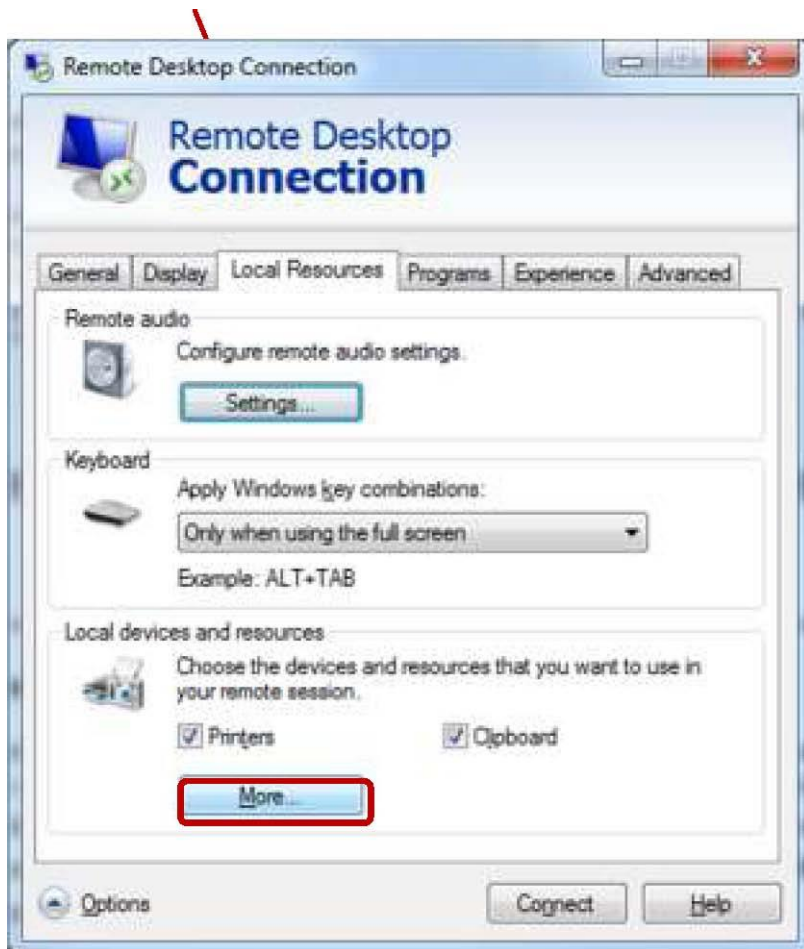
Microsoft's website
sqlserverdatamining.com

Importing data to Excel from SQL Server 2008

To demonstrate how to import data to excel workbook from SQL Server 2008, we will import the churn table from Public_Datasets_DM database in SQL Server 2008. You will need to log in to ts-mec.waltoncollege.uark.edu with your WALTON credentials to be able to do this. On your computer use Remote Desktop Connection to connect to **ts-mec.waltoncollege.uark.edu** and click Options.



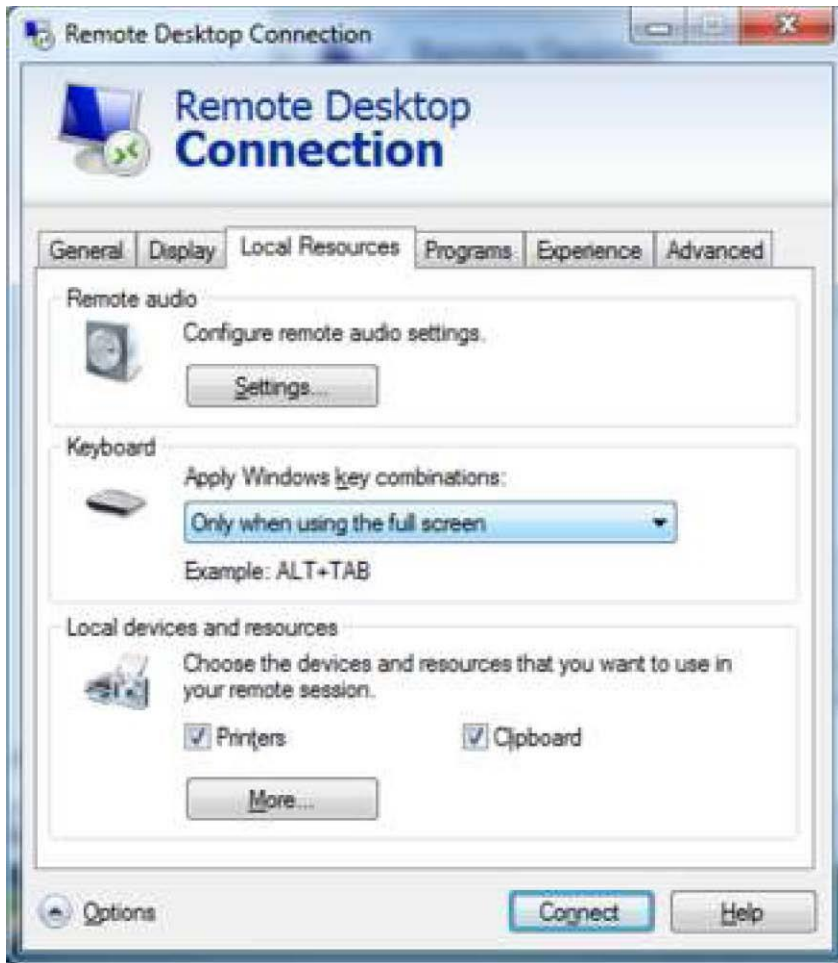
Click the Local Resources tab and click the More... button



In the Local devices and resources, expand Drives and check the check box for C: drive as shown in the screen below. Click OK.



Then, click Connect.



Click Connect.



Enter your credentials (Walton\Es##### and password provided by the University of Arkansas) and click OK.



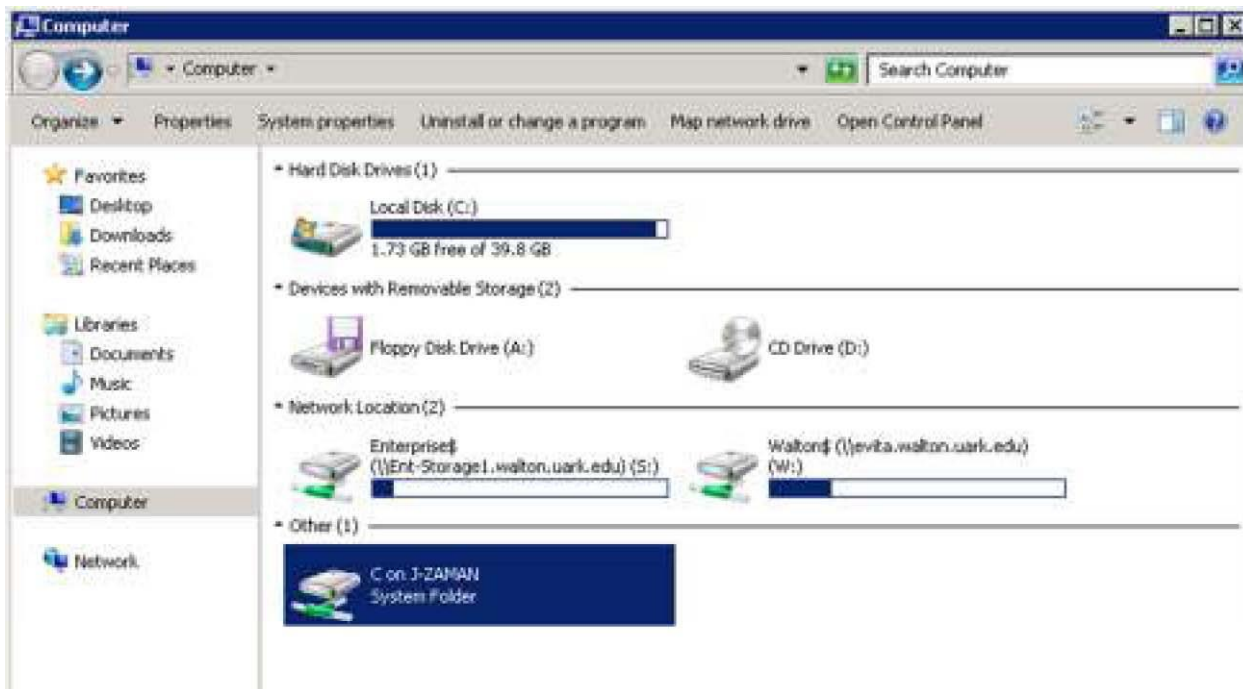
Click Yes.



Once you're logged in, open My Computer as shown in the screen below.



You will see YOUR C: drive connected to the ts-mec.waltoncollege.uark.edu server under Other drives. You can double click the drive to see the contents of YOUR C drive from REMOTE server.

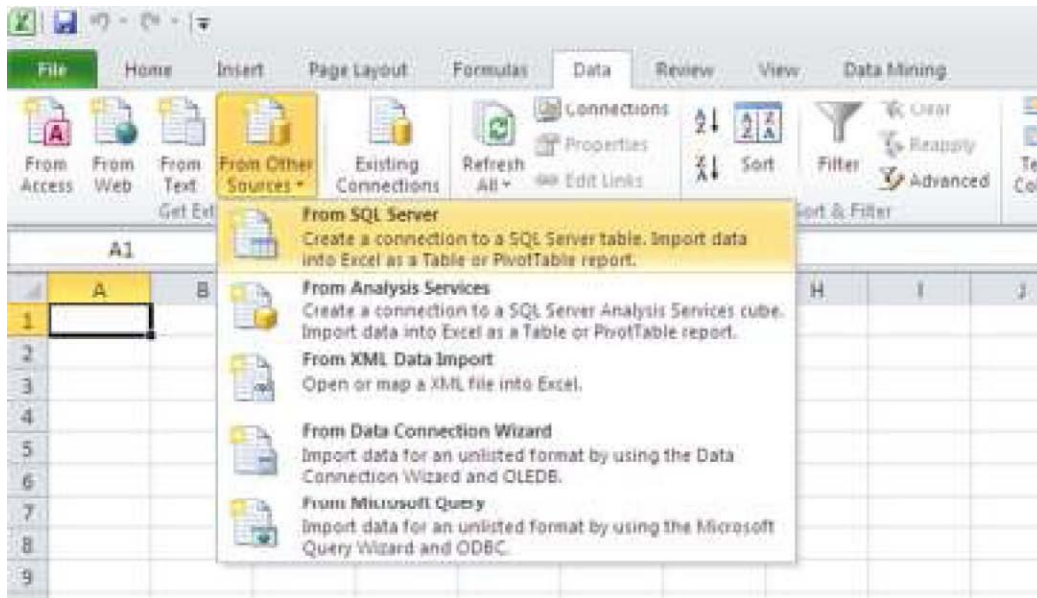


Then, let's start the process of importing data from SQL server to Excel workbook. On the Remote server, click Start and Open Excel.



Import Data

To import data from SQL server to Excel workbook, follow these steps. Go to the Data tab→From Other Sources→From SQL Server.



Type in `msenterprise.waltoncollege.uark.edu` for Server Name. Enter your ES account credentials (**without** the Walton domain). Then click Next.

Data Connection Wizard

Connect to Database Server

Enter the information required to connect to the database server.

1. Server name:

2. Log on credentials

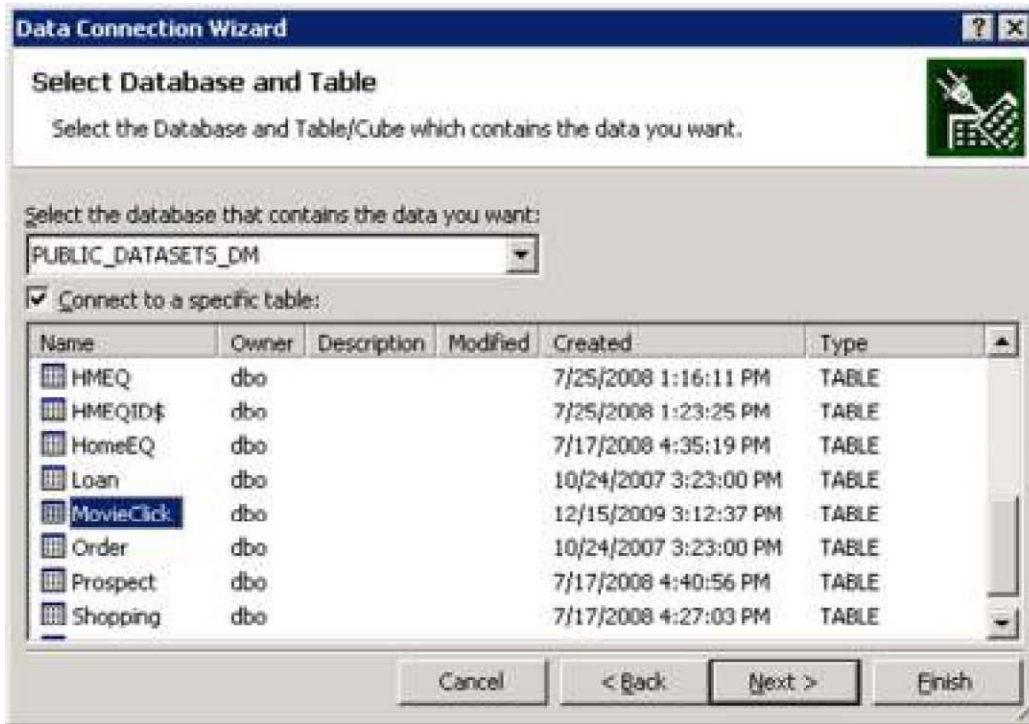
Use Windows Authentication

Use the following User Name and Password

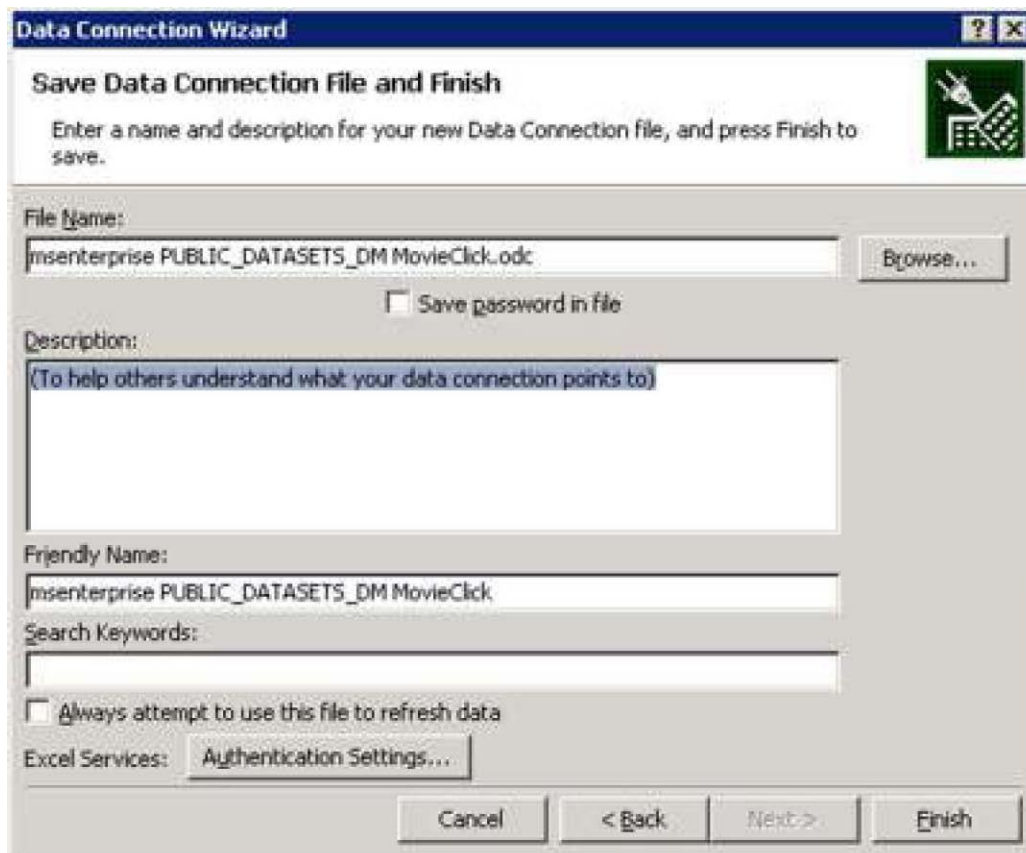
User Name:

Password:

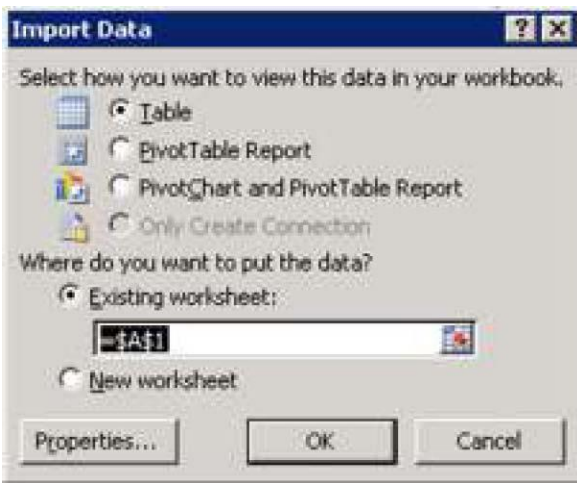
Select PUBLIC_DATASETS_DM from the drop down list and select MovieClick database. Click Next.



Now click Finish.



Click OK to insert the data in existing worksheet.



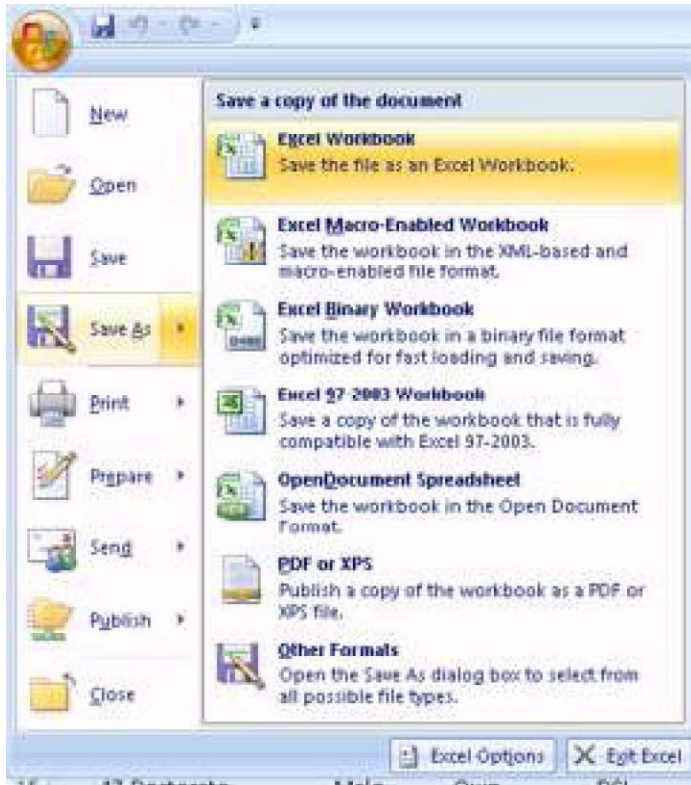
Enter your password again, if prompted.



Then, you will see the churn table data opened in your worksheet, as shown in the next screen.

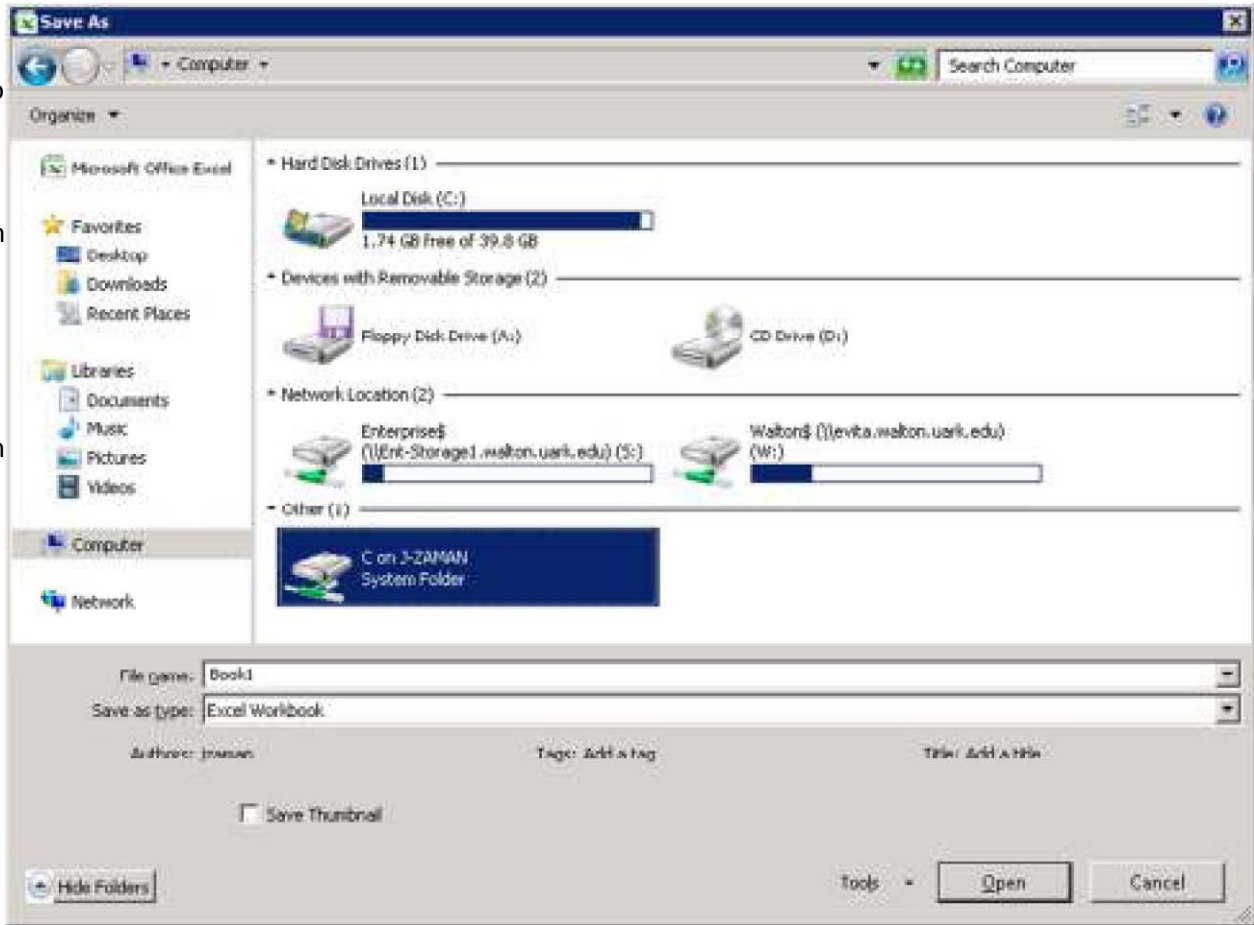
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	AGE	EDUCATION	GENDER	HORN OWN	REEMET COIN	MARITALSTATUS	MOVIESECTOR	NBRBRATHS	NBRDRIMS	NBRCAFS	NBRCHILDREN	DEBTVS	PPWAL
2	47	Doctorate	Male	Own	DSL	Married	Spouse/Partner	2	2	2	1	1	Never
3	47	Doctorate	Male	Own	DSL	Married	Spouse/Partner	2	2	2	1	1	Never
4	47	Doctorate	Male	Own	DSL	Married	Spouse/Partner	2	2	2	1	1	Never
5	47	Doctorate	Male	Own	DSL	Married	Spouse/Partner	2	2	2	1	1	Never
6	47	Doctorate	Male	Own	DSL	Married	Spouse/Partner	2	2	2	1	1	Never
7	47	Doctorate	Male	Own	DSL	Married	Spouse/Partner	2	2	2	1	1	Never
8	47	Doctorate	Male	Own	DSL	Married	Spouse/Partner	2	2	2	1	1	Never
9	47	Doctorate	Male	Own	DSL	Married	Spouse/Partner	2	2	2	1	1	Never
10	47	Doctorate	Male	Own	DSL	Married	Spouse/Partner	2	2	2	1	1	Never
11	47	Doctorate	Male	Own	DSL	Married	Spouse/Partner	2	2	2	1	1	Never
12	47	Doctorate	Male	Own	DSL	Married	Spouse/Partner	2	2	2	1	1	Never
13	47	Doctorate	Male	Own	DSL	Married	Spouse/Partner	2	2	2	1	1	Never
14	47	Doctorate	Male	Own	DSL	Married	Spouse/Partner	2	2	2	1	1	Never
15	47	Doctorate	Male	Own	DSL	Married	Spouse/Partner	2	2	2	1	1	Never
16	47	Doctorate	Male	Own	DSL	Married	Spouse/Partner	2	2	2	1	1	Never
17	47	Doctorate	Male	Own	DSL	Married	Spouse/Partner	2	2	2	1	1	Never
18	47	Doctorate	Male	Own	DSL	Married	Spouse/Partner	2	2	2	1	1	Never
19	47	Doctorate	Male	Own	DSL	Married	Spouse/Partner	2	2	2	1	1	Never
20	47	Doctorate	Male	Own	DSL	Married	Spouse/Partner	2	2	2	1	1	Never
21	47	Doctorate	Male	Own	DSL	Married	Spouse/Partner	2	2	2	1	1	Never

Now, Save this worksheet to YOUR desktop. (note that this worksheet is open in the Remote server). Click File > Save As...



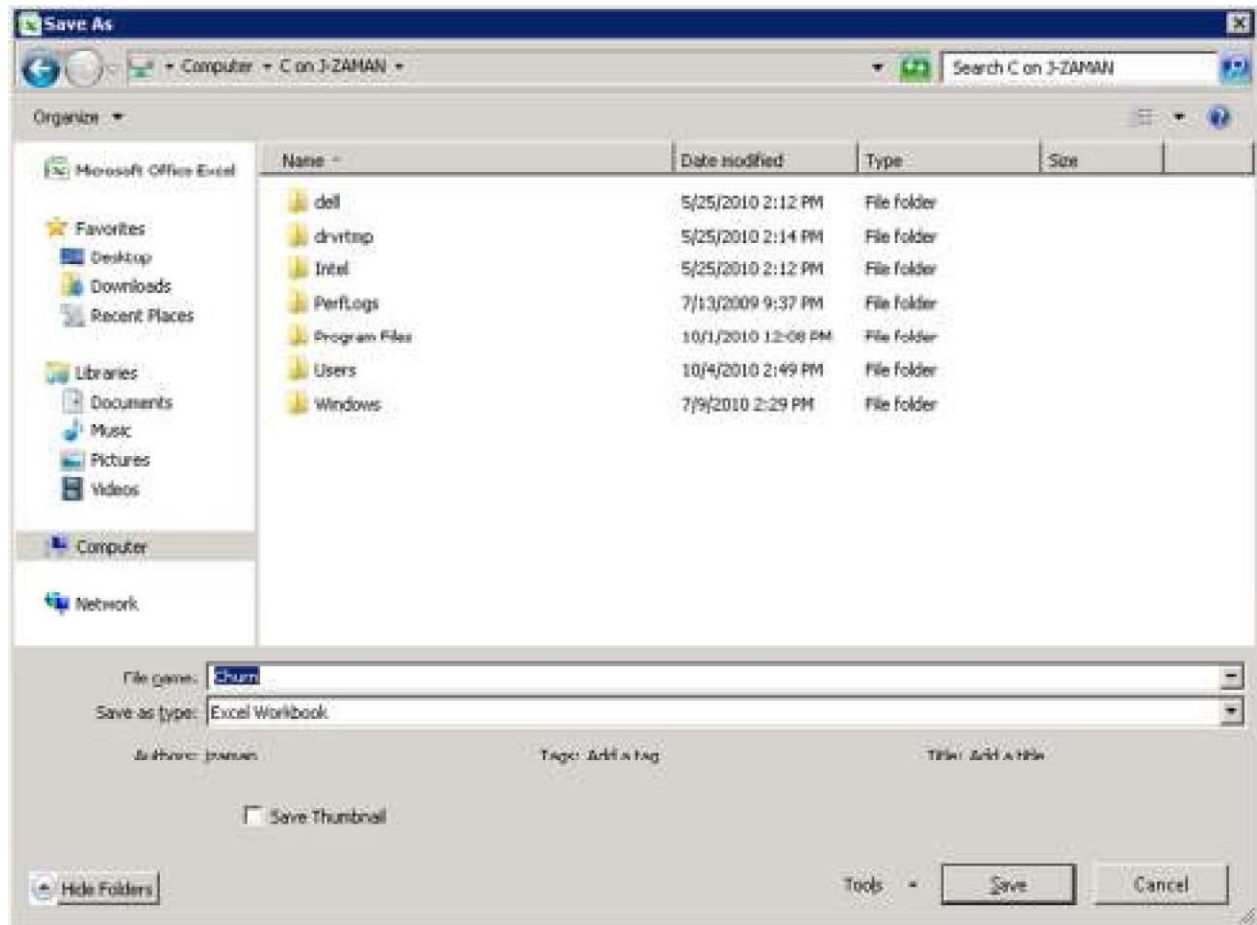
And make sure to select YOUR C: drive by double clicking it.

Click Save to save the churn table in that folder in YOUR C: drive in your local



computer.

Close
the
Excel



workbook in Remote server and log off ts-mec.waltoncollege.uark.edu

Now, on YOUR computer, open YOUR C: drive, and then open the Churn Excel file and you can proceed with Data Mining from Excel.