CMLgroup

Define Measure Manage

Business Performance Management

Advance Data Mining Demo

14 Wall Street, 20th Floor – New York, NY 10005 – Phone 1 646 827-2291 – http://www.cmlgroup.com

Advance Data Mining – Demo Summary

The purpose of this demo is to provide an additional deep-dive into the functionality of the Data Mining Client for Excel, which is part of the new Microsoft SQL Server 2005 Data Mining Add-ins for Office 2007. This add-in allows users to go through the full data mining model development lifecycle within Excel 2007 using either your spreadsheet data or external data accessible through your SQL Server 2005 Analysis Services instance.

In this demo, you will see how to use the Data Mining Client for Excel to:

- Clean a data source and prepare it for use in data modeling
- Partition a data source into training and testing data
- Build a data mining model
- Verify the overall accuracy of the data mining model
- Create and model a scenario using the data mining model
- Apply the scenario results to a separate data set

Advance Data Mining – Demo Summary

As companies accumulate large amounts of data in data warehouses, one of their key challenges becomes how to sort through all of their data to uncover useful information. Typically, analyzing all this data is the realm of specialized data analysts who use complex data-modeling programs to sift through a data warehouse to uncover actionable business intelligence.

In order to make this process easier, and allow more knowledge workers in an organization to perform data mining activities, Microsoft is introducing the Microsoft SQL Server 2005 Data Mining Add-ins for Office 2007. These add-ins allow workers to take advantage of SQL Server 2005 predictive analytics in Office Excel 2007 and Office Visio 2007.

The Data Mining Add-ins for Excel consist of two parts:

1. Table Analysis Tools for Excel: This add-in provides you with easy-to-use tasks that leverage SQL Server 2005 Data Mining under the covers to perform powerful analytics on your spreadsheet data.
2. Data Mining Client for Excel: This add-in allows you to go through the full data mining model development lifecycle within Excel 2007 using either your spreadsheet data or external data accessible through your SQL Server 2005 Analysis Services instance.

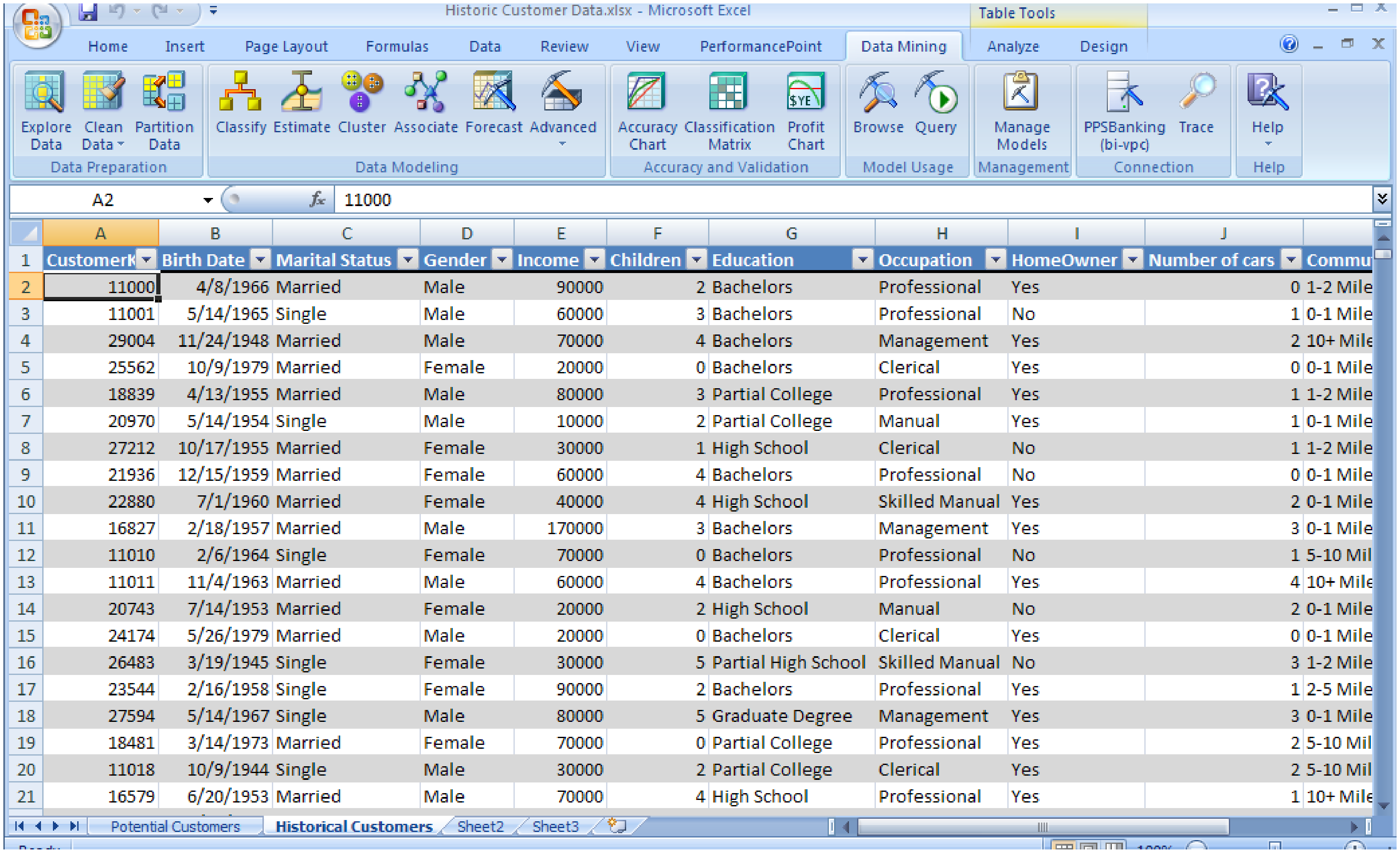
In this demonstration, let's see how we can use the full Data Mining Client for Excel to analyze historic customer purchasing data to prepare a targeted direct mail marketing campaign for potential customers.

In this SharePoint Report Library, we'll go ahead and open up this workbook containing historic customer data. This spreadsheet contains two separate sheets of information. The first tab contains a list of potential customers leads. The second tab, which we currently have open, is a list of historic customers. Most importantly, this list indicates whether a customer has previously purchased a bike.

By using this information, along with the Data Mining Client for Excel, we'll be able to create a new data model, verify it's accuracy, and then use it to model the marketing campaign. Finally, we'll apply these results to the list of potential customers in order to identify those customers that will be most likely to respond to the marketing and purchase a bike.

Before we begin building the data model, it's usually a good idea to spend some time cleaning up the data. Most data sets contain a small number of extremely uncommon values, known as outliers. These outliers have the potential to skew the data model. By removing them, we can improve the accuracy of the entire model.

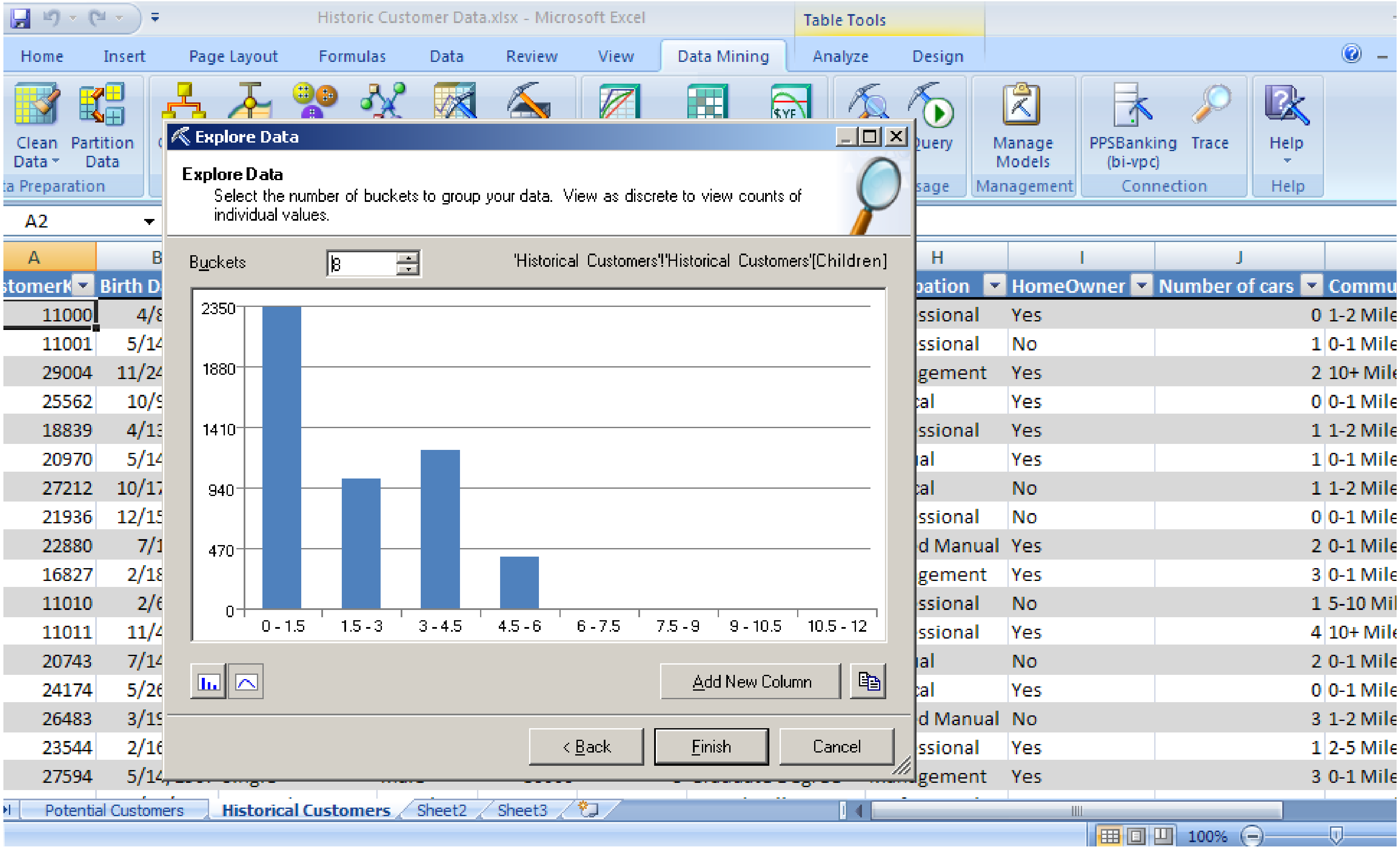
To begin this process, let's explore the data and see the current distribution for some of the values looks like.



	A	B	C	D	E	F	G	H	I	J	
1	CustomerK	Birth Date	Marital Status	Gender	Income	Children	Education	Occupation	HomeOwner	Number of cars	Commu
2	11000	4/8/1966	Married	Male	90000	2	Bachelors	Professional	Yes	0	1-2 Mile
3	11001	5/14/1965	Single	Male	60000	3	Bachelors	Professional	No	1	0-1 Mile
4	29004	11/24/1948	Married	Male	70000	4	Bachelors	Management	Yes	2	10+ Mile
5	25562	10/9/1979	Married	Female	20000	0	Bachelors	Clerical	Yes	0	0-1 Mile
6	18839	4/13/1955	Married	Male	80000	3	Partial College	Professional	Yes	1	1-2 Mile
7	20970	5/14/1954	Single	Male	10000	2	Partial College	Manual	Yes	1	0-1 Mile
8	27212	10/17/1955	Married	Female	30000	1	High School	Clerical	No	1	1-2 Mile
9	21936	12/15/1959	Married	Female	60000	4	Bachelors	Professional	No	0	0-1 Mile
10	22880	7/1/1960	Married	Female	40000	4	High School	Skilled Manual	Yes	2	0-1 Mile
11	16827	2/18/1957	Married	Male	170000	3	Bachelors	Management	Yes	3	0-1 Mile
12	11010	2/6/1964	Single	Female	70000	0	Bachelors	Professional	No	1	5-10 Mil
13	11011	11/4/1963	Married	Male	60000	4	Bachelors	Professional	Yes	4	10+ Mile
14	20743	7/14/1953	Married	Female	20000	2	High School	Manual	No	2	0-1 Mile
15	24174	5/26/1979	Married	Male	20000	0	Bachelors	Clerical	Yes	0	0-1 Mile
16	26483	3/19/1945	Single	Female	30000	5	Partial High School	Skilled Manual	No	3	1-2 Mile
17	23544	2/16/1958	Single	Female	90000	2	Bachelors	Professional	Yes	1	2-5 Mile
18	27594	5/14/1967	Single	Male	80000	5	Graduate Degree	Management	Yes	3	0-1 Mile
19	18481	3/14/1973	Married	Female	70000	0	Partial College	Professional	Yes	2	5-10 Mil
20	11018	10/9/1944	Single	Male	30000	2	Partial College	Clerical	Yes	2	5-10 Mil
21	16579	6/20/1953	Married	Male	70000	4	High School	Professional	Yes	1	10+ Mile

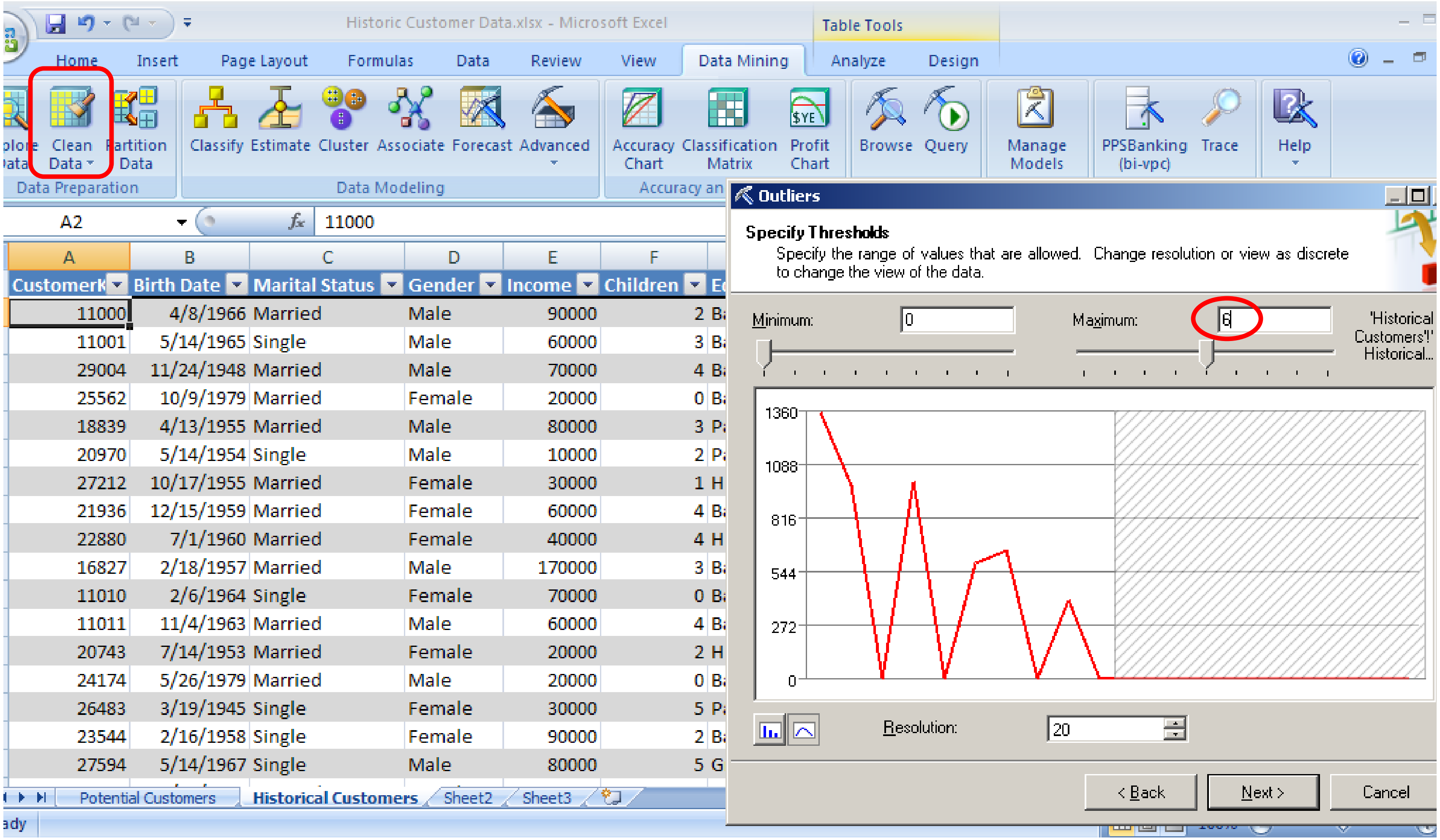
Data Mining – Database Analysis

Using this Explore Data wizard, we can see a list of all the columns of data contained in the spreadsheet. If we look at the Customer Key, we can see that it's evenly distributed. This is because the Customer Key is nothing more than an identifier field, and contains no relevant demographic information about a customer. This is important to know, because we should remove this column when we build the actual data model.



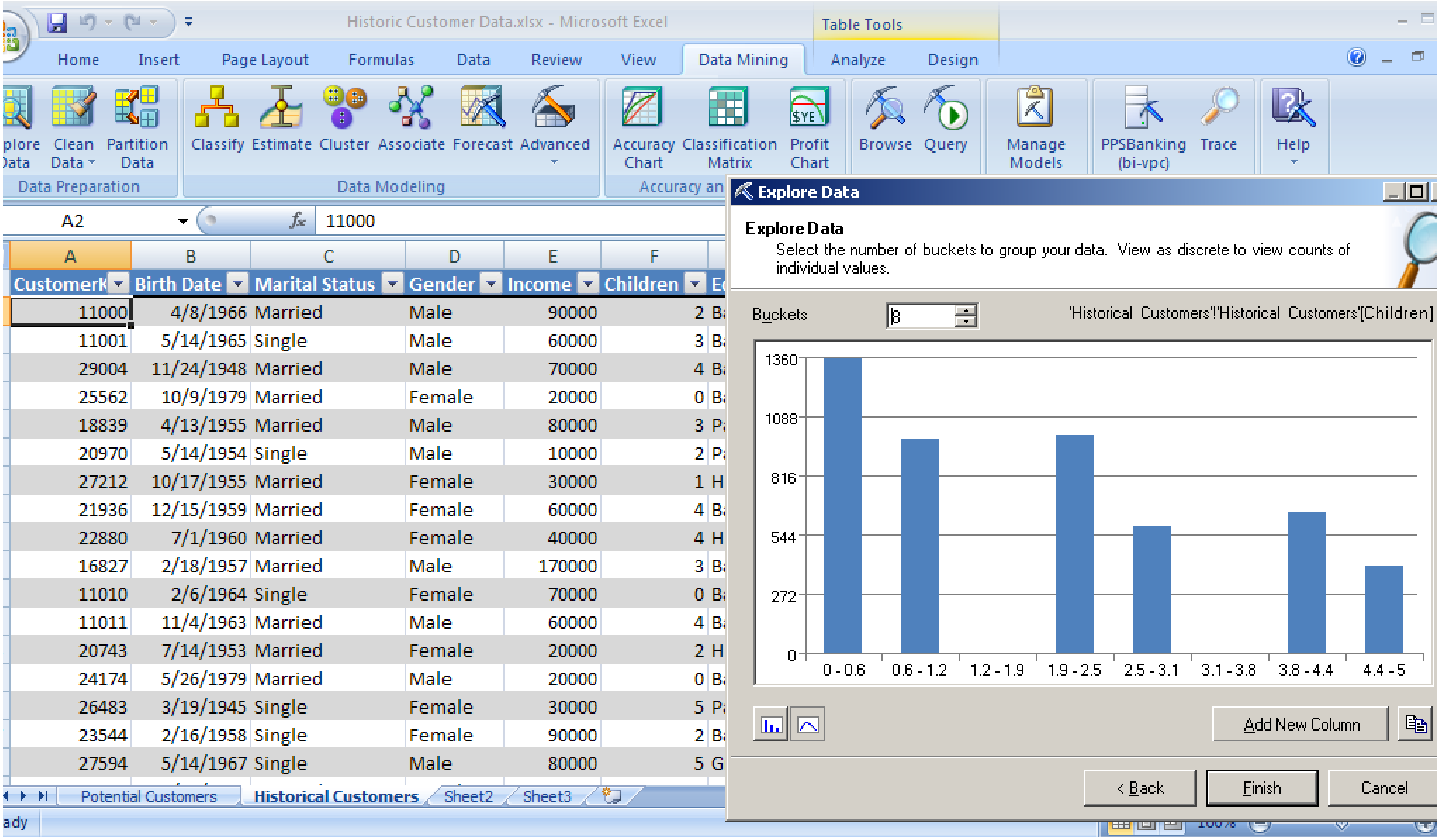
Data Mining – Demographic Information

Now, let's go back and look at a column of data that we know contains useful demographic information. We'll choose the number of children that a customer has. Looking at the results, we can see that the data follows the typical long tail distribution, where most of the values fall within a certain range, with a small number of outliers. In this case, most customers have six or less children, with only a few customers having more than that.



Data Mining – Data Cleanup, Remove Outliers

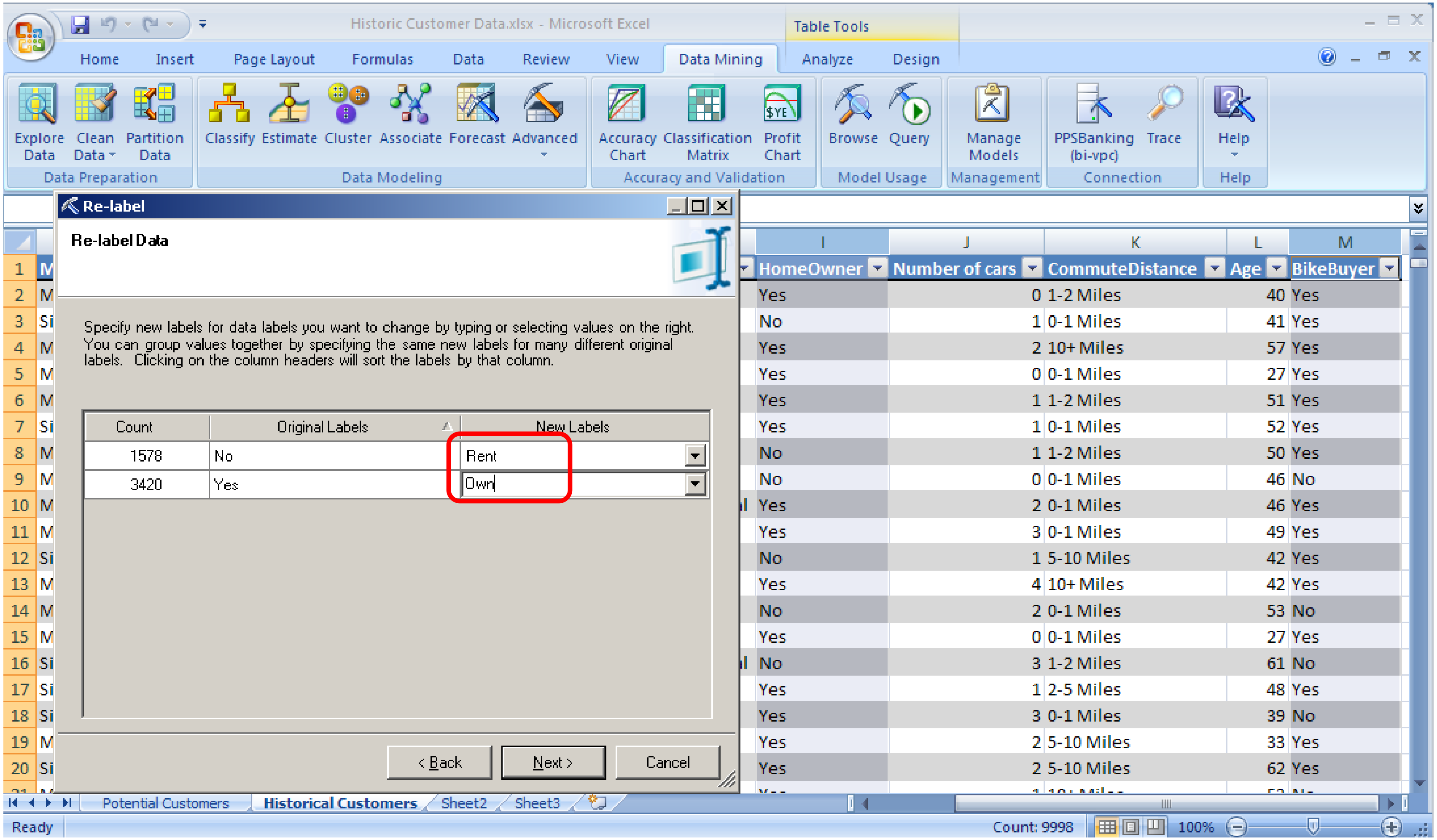
What we want to do here is remove those rows containing the outliers, so they do not skew the data model when it is built. Using this Clean Data wizard, we'll quickly remove all rows for customers that have more than six children.



Data Mining – Data cleanup, Review Distribution of Values

When we now go back and review the distribution of values for the children column, we can see that the outliers have been eliminated from the data set.

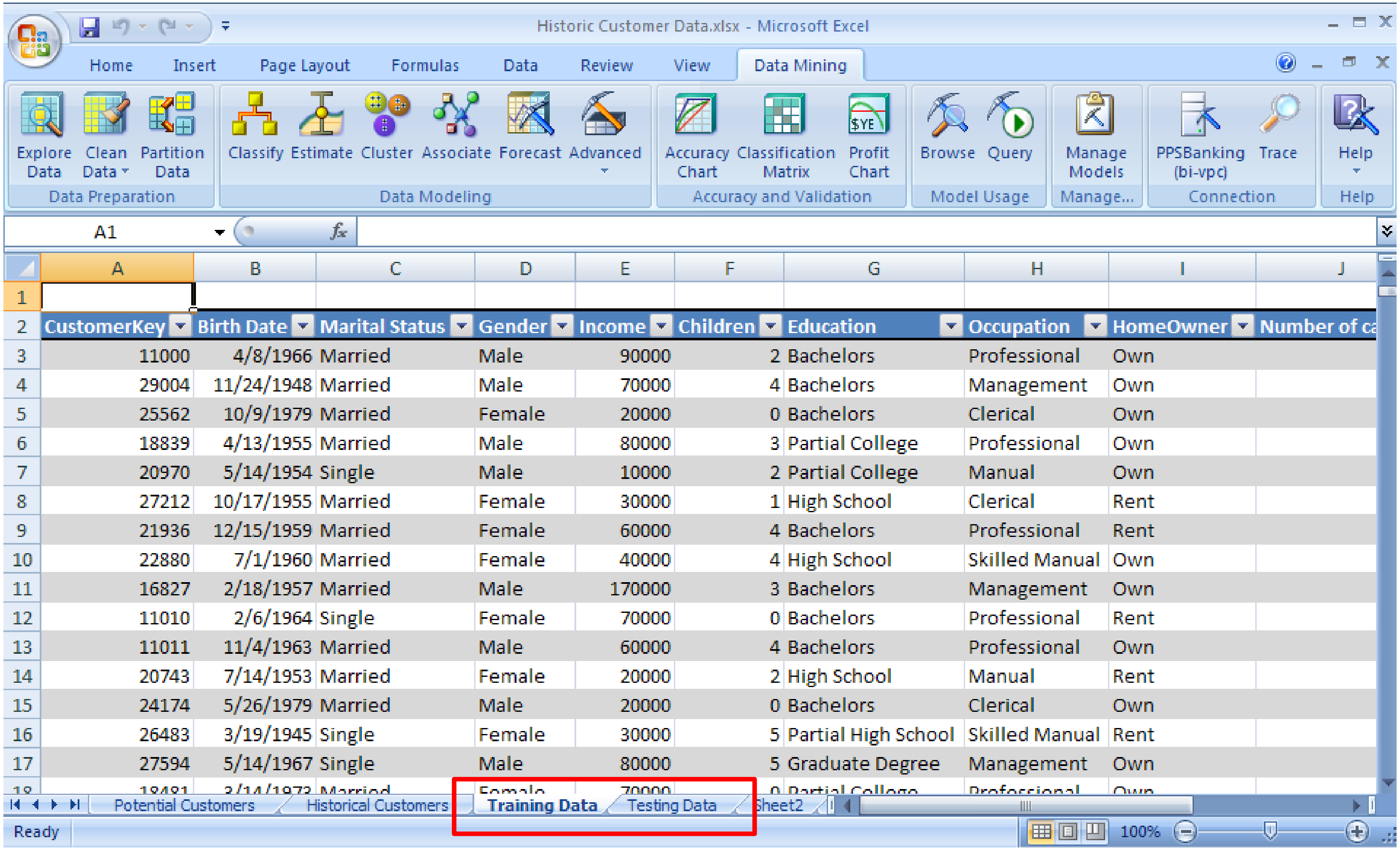
This same process can be used to eliminate the outliers from the other columns as well.



Data Mining – Data cleanup, Re-label

Returning to the Historical Customers worksheet, there's one other issue we want to correct. Notice that the values for the HomeOwner column are the same as the values for the BikeBuyer column. Although this won't impact the data modeling, for clarity we might as well take the opportunity to clean this up as well.

Using the Re-Label Wizard, we can easily change the values for the HomeOwner column from "Yes" and "No" to "Own" and "Rent".

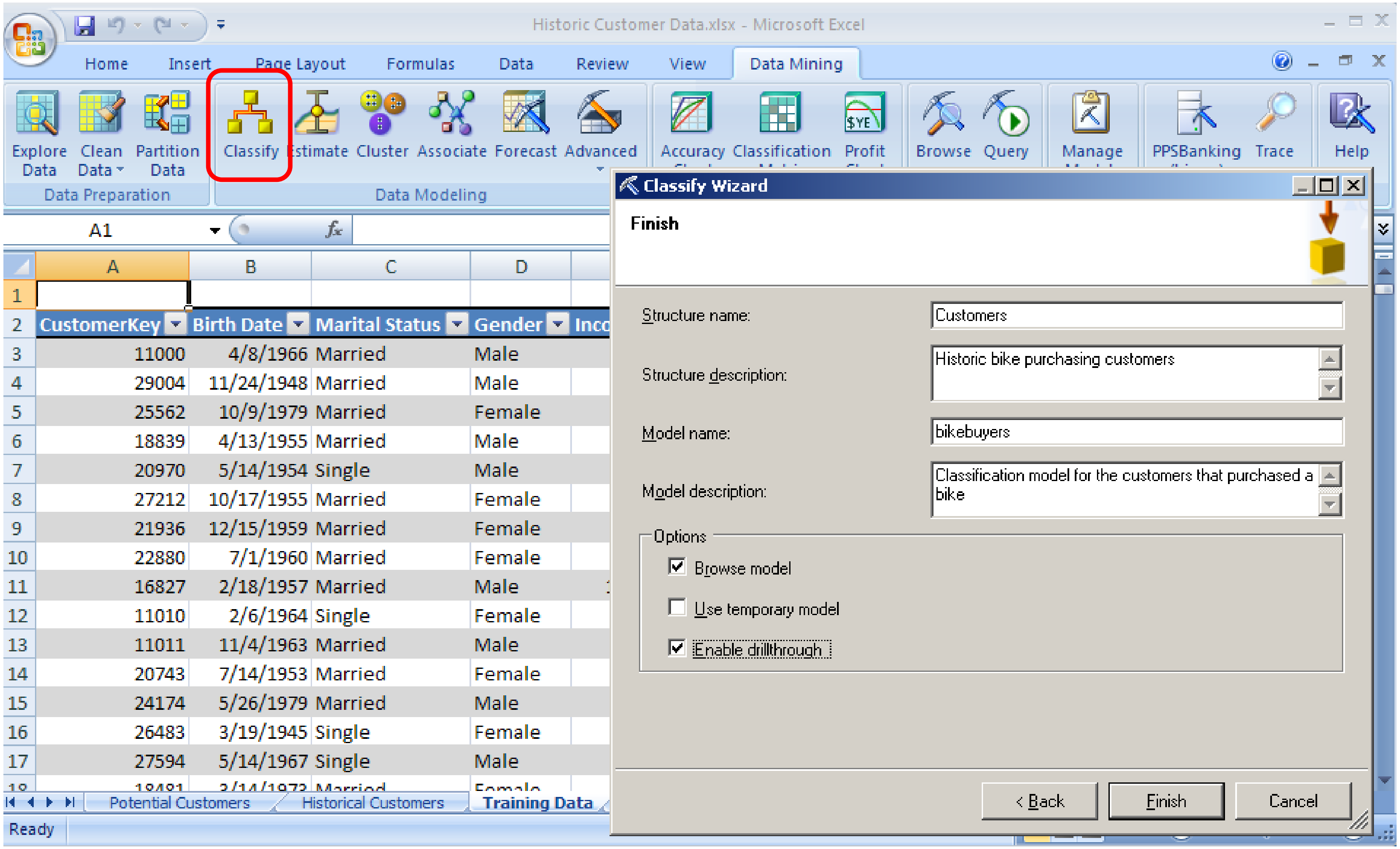


CustomerKey	Birth Date	Marital Status	Gender	Income	Children	Education	Occupation	HomeOwner	Number of ca
11000	4/8/1966	Married	Male	90000	2	Bachelors	Professional	Own	
29004	11/24/1948	Married	Male	70000	4	Bachelors	Management	Own	
25562	10/9/1979	Married	Female	20000	0	Bachelors	Clerical	Own	
18839	4/13/1955	Married	Male	80000	3	Partial College	Professional	Own	
20970	5/14/1954	Single	Male	10000	2	Partial College	Manual	Own	
27212	10/17/1955	Married	Female	30000	1	High School	Clerical	Rent	
21936	12/15/1959	Married	Female	60000	4	Bachelors	Professional	Rent	
22880	7/1/1960	Married	Female	40000	4	High School	Skilled Manual	Own	
16827	2/18/1957	Married	Male	170000	3	Bachelors	Management	Own	
11010	2/6/1964	Single	Female	70000	0	Bachelors	Professional	Rent	
11011	11/4/1963	Married	Male	60000	4	Bachelors	Professional	Own	
20743	7/14/1953	Married	Female	20000	2	High School	Manual	Rent	
24174	5/26/1979	Married	Male	20000	0	Bachelors	Clerical	Own	
26483	3/19/1945	Single	Female	30000	5	Partial High School	Skilled Manual	Rent	
27594	5/14/1967	Single	Male	80000	5	Graduate Degree	Management	Own	

Data Mining – Historical Data Partition

Now that we've cleaned up the data set, we're ready to start building the data model that will be used to identify likely customers. The first step in this process is to partition the historical data into two separate spreadsheets. This is necessary because we need both a training data set and a testing data set. The training data set will be used in detecting patterns in the data and building the actual model, while the testing data set will be used in validating the patterns in the model.

We'll accomplish this task by using the Partition Data Wizard. After the wizard has run, note that two new spreadsheets have been created in the workbook. 70% of the data has been randomly assigned as training data, while the rest is now our testing data.



Data Mining – Data Classification

We're now ready to build the actual model, using the Classify task on the ribbon.
Stepping through this wizard will allow us to generate a decision tree model for our predictive analysis.

Data Mining Modeling

Note that the Data Mining Client for Excel supports other kinds of models in addition to decision trees, as described by the other options on the ribbon.

These other models include:

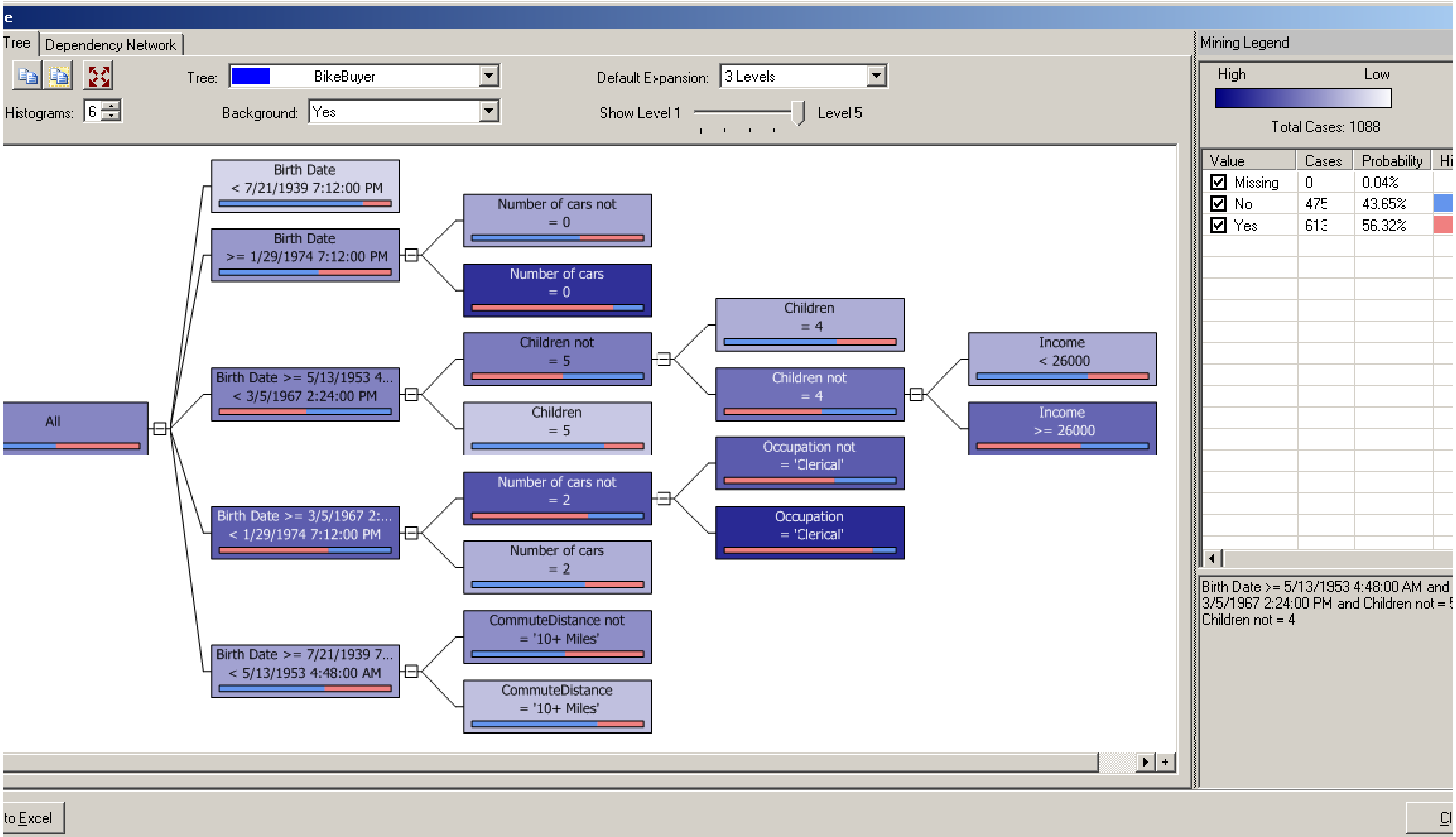
- Estimate – Builds a regression model for predicting numeric values.
- Cluster – Builds a clustering model for detecting groups of rows with similar characteristics.
- Associate – Detects associations between items that appear together in multiple transactions.
- Forecast – Detects patterns in a series of cells and uses these patterns to forecast the evolution of the series.

In the Classify Wizard, we're going to choose to analyze the BikeBuyer column. This will allow us to understand how the values in the other columns impact the decision of a customer to purchase a bike. We also want to remove the CustomerKey value from the model, since we determined previously it isn't relevant to the model.

After we have finished creating the data model, we'll have an opportunity to preview the decision tree that was generated. Let's adjust the Background filter, so that it will shade cells on the decision tree darker that are more strongly associated with the decision to purchase a bike.

We can also drill through on one of these nodes. Here we can see a new Excel spreadsheet containing all the rows that were associated with that node.

14 Wall Street, 20th Floor New York, NY 10005 Phone 1 646 827-2291 http://www.cmlgroup.com



Data Mining – Data cleanup, Review Distribution of Values

These other models include:
Estimate – Builds a regression model for predicting numeric values.
Cluster – Builds a clustering model for detecting groups of rows with similar characteristics.
Associate – Detects associations between items that appear together in multiple transactions.
Forecast – Detects patterns in a series of cells and uses these patterns to forecast the evolution of the series.

	A	B	C
1	Counts of correct/incorrect classification for model 'BikeBuyers' on column 'BikeBuyer'		
2	Rows correspond to predicted values		
3	**Total correct:**	**64.24 %**	**963**
4	**Total misclassified:**	**35.76 %**	**536**
5			
6	Results as Percentages		
7		No(Actual)	Yes(Actual)
8	No	**66.75 %**	38.49 %
9	Yes	33.25 %	**61.51 %**
10			
11	**Correct**	**66.75 %**	**61.51 %**
12	**Misclassified**	**33.25 %**	**38.49 %**
13			
14	Results as Counts		
15		No(Actual)	Yes(Actual)
16	No	**522**	276
17	Yes	260	**441**
18			
19	**Correct**	**522**	**441**
20	**Misclassified**	**260**	**276**

Data Mining – Testing Data, Classification Matrix

In order to use the mining model in our targeted mailing campaign, we need to make sure that the patterns indicated by the model are correct.

This Classification Matrix Wizard allows us to evaluate the performance of our newly-created model against the test data we partitioned.

After running this wizard, we're presented with a report summarizing the accuracy of the mining model. In an ideal model, all of the customers would be classified correctly and we could predict with 100% accuracy which customer would purchase a bike. However, we can see from this report that roughly 65% of the customers were classified correctly in our model. The rest of the results were either false positives or false negatives, and ended up being misclassified by the model. This is still considerably better than just guessing which customers to market to, making this a good candidate for our mining model.

Data Validation – Accuracy Chart

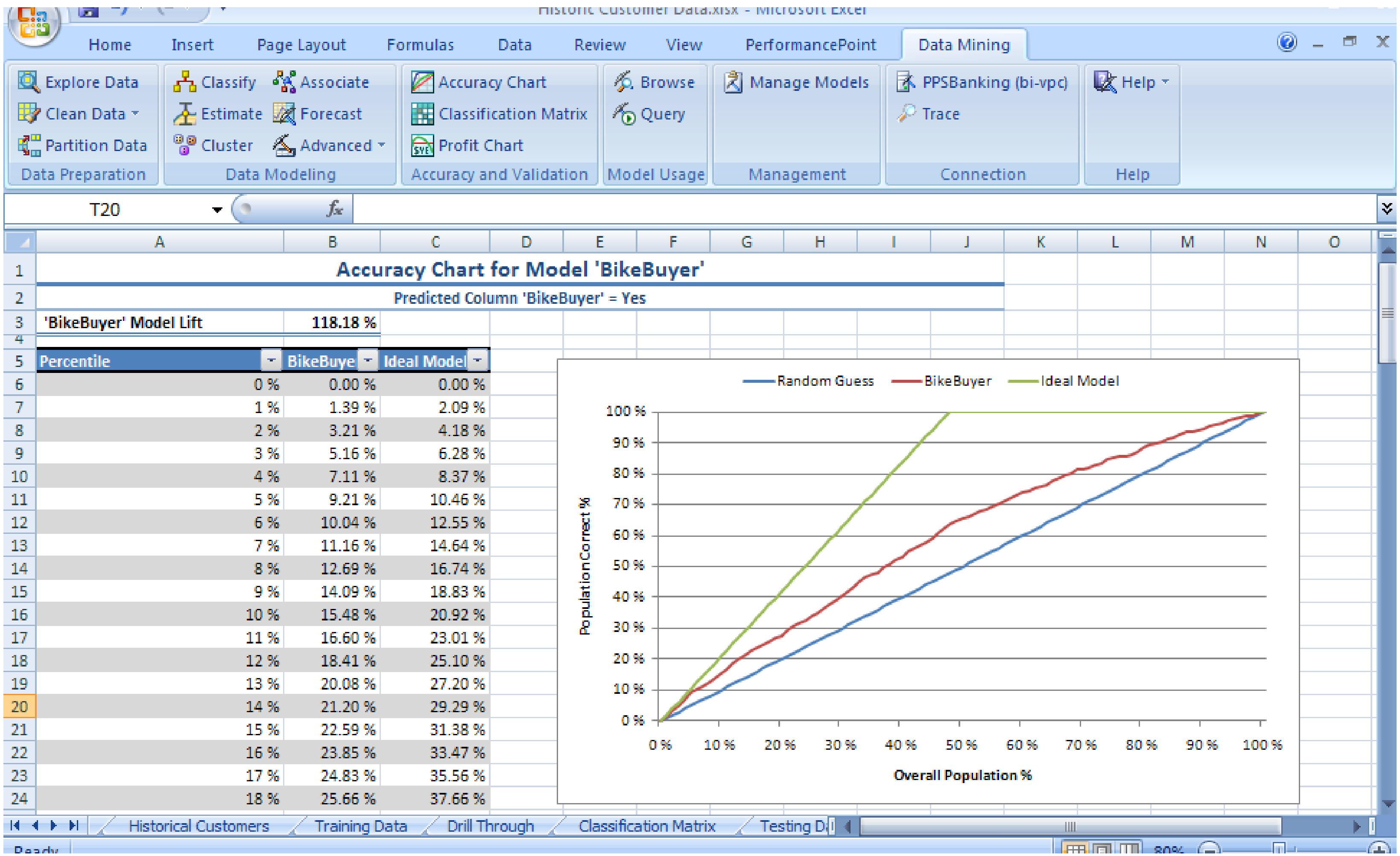
The second step in validation is verifying how good a job the model does with predicting the result that is interesting to us. In our case, we want to know how good this model is for finding customers who will purchase a bike. This is different from our previous analysis, which just looked at how good the model was for classifying all outcomes correctly.

For this, we will use the Accuracy Chart Wizard. This wizard will build a lift chart that presents the model performance compared against a hypothetical ideal model (which performs classification without errors) and a random guess (equivalent to flipping a coin in determining whether a given customer is likely to purchase a bike).

Here we can see the lift chart that the wizard has generated. The vertical axis represents the target population (customers that actually buy a bike) while the horizontal axis represents the total population. Let's assume that 50% of our customers are likely to purchase a bike, and 50% are not. What we want to know is how accurate our model is at finding the correct 50% of our customers.

In this chart, the green line represents the ideal model. This line flattens out at 50% of the total population. This is because the ideal model correctly predicts every customer that will purchase a bike, and targeting any single customer after that 50% would not generate any additional bike purchases.

The blue line represents a random guess. As you would expect, the accuracy of this guess is like flipping a coin. To reach the same number of bike purchases as the ideal model, we would have to market to twice as many people by randomly guessing which ones are likely to purchase a bike.



Data Mining – Testing Data, Accuracy Chart

The reality of our mining model is the red line, which is in between the ideal model and the random guess. This means that our model is indeed better at predicting outcomes than just guessing. The gain between our model and a random guess is what is known as the lift. We can see by the measure at the top of the spreadsheet that the lift for our model is about 17%.

Marketing Campaign Model

Now that our analysis has shown that the predictive model is reasonably accurate, we're ready to model the marketing campaign. This modeling involves knowing more than just which customers to market to. We also have to know how many of these customers we should market to.

The number of customers we should market to is affected by the costs of the marketing campaign itself. For example, there will be fixed costs associated with creating the marketing materials, and variable costs for postage and shipping. If we market to too few people, our profit will be reduced by the fixed costs. Similarly, if we market to too many people, the postage and shipping costs will eat into the profit. What we want to determine is the sweet spot where we determine the exact number of people we should market to in order to maximize our profit.

To determine this optimal value, we'll use the Profit Chart Wizard, which allows us to build a profit chart for an existing classification model. A profit chart displays the estimated profit increase that would be associated with using a mining model to determine which customers should be contacted in a business scenario. We'll base this profit chart on our existing test data.

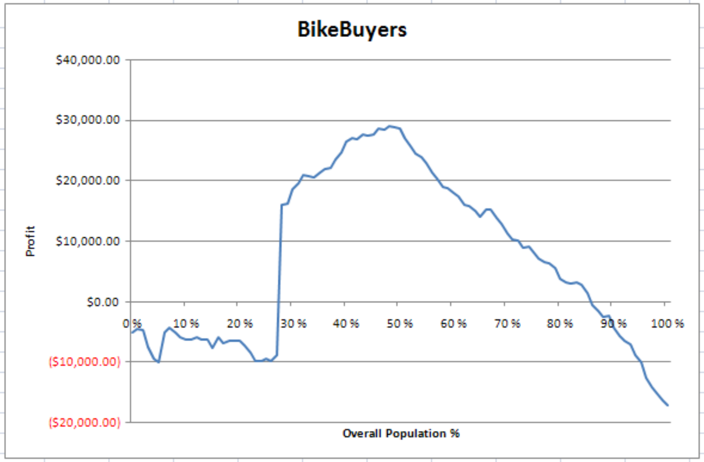
As we build this profit chart, there are a few values we'll be adjusting to make sure the profit is modeled correctly. We'll change the target population to match the number of potential customers we're considering marketing to. In this window, we're adding the individual costs associated with marketing to each customer. Notice that as we market to more customers, efficiencies in the direct mailing campaign process allow us to achieve some cost savings. Finally, we'll update the revenue per individual.

Profit Chart for Model 'BikeBuyers'

Target Population: 'BikeBuyer' = Yes

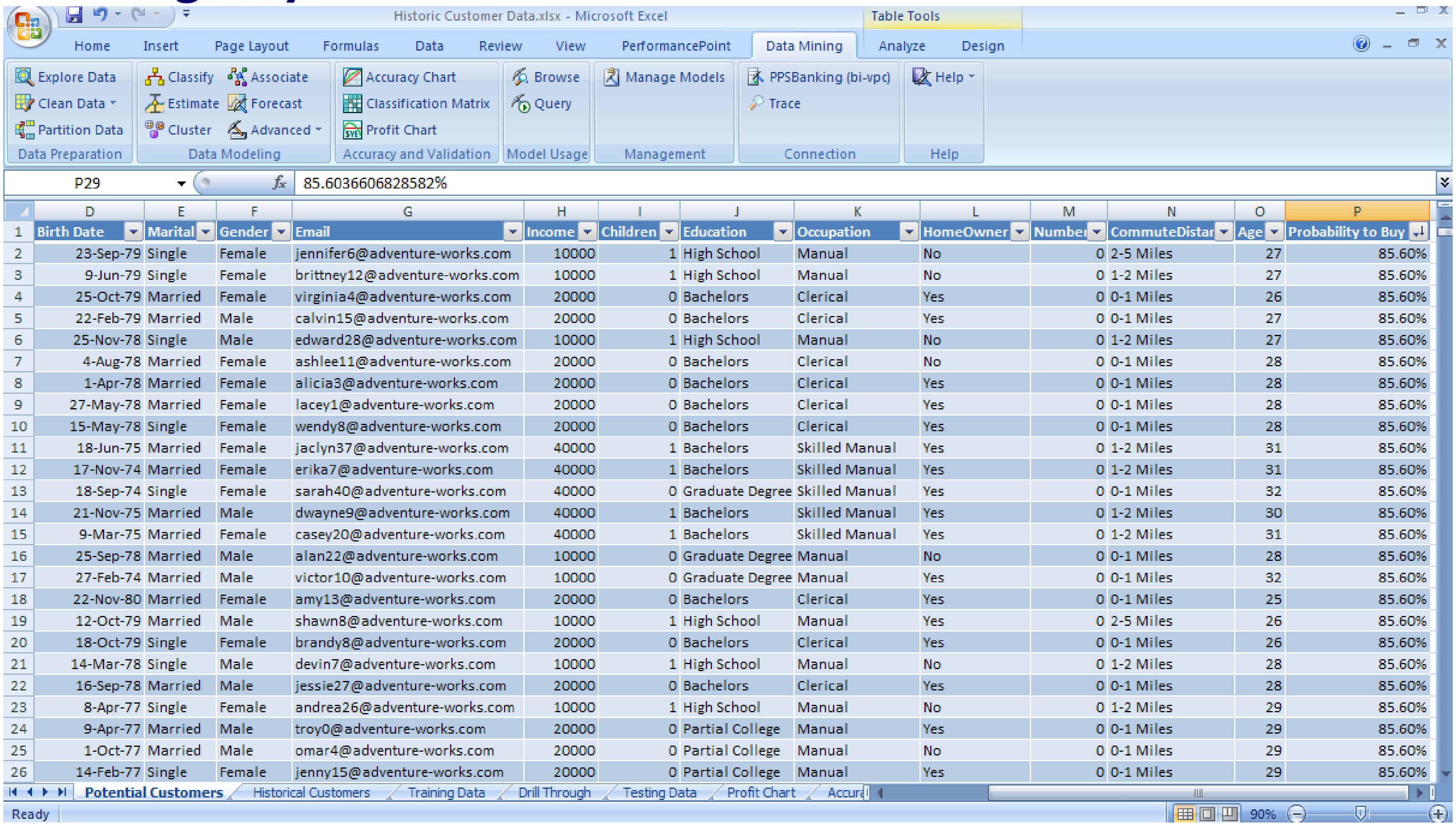
Fixed cost	5000
Population	18500
Individual cost	Up to 1000: $25.00 Up to 5000: $20.00 More than 5000: $15.00
Revenue per individual	30
Maximum profit	**$29,062.71**
Probability threshold	**42.20 %**

Percentile	BikeBuyers	Profit	Probability
0 %	0.00 %	($5,000.00)	100.00 %
1 %	1.95 %	($4,444.63)	85.60 %
2 %	3.63 %	($4,629.75)	69.66 %
3 %	4.32 %	($7,406.60)	69.66 %
4 %	5.30 %	($9,442.96)	69.66 %
5 %	6.83 %	($9,998.33)	69.66 %
6 %	8.37 %	($5,000.00)	69.66 %
7 %	10.04 %	($4,259.51)	69.66 %
8 %	11.16 %	($5,000.00)	69.66 %
9 %	12.27 %	($5,740.50)	69.66 %
10 %	13.53 %	($6,110.74)	69.66 %
11 %	14.92 %	($6,110.74)	69.66 %
12 %	16.46 %	($5,740.49)	69.66 %
13 %	17.71 %	($6,110.74)	69.66 %
14 %	19.11 %	($6,110.74)	69.66 %
15 %	19.94 %	($7,591.73)	69.66 %
16 %	22.04 %	($5,740.49)	69.66 %
17 %	23.01 %	($6,851.24)	69.66 %
18 %	24.55 %	($6,480.99)	69.66 %
19 %	25.94 %	($6,480.98)	69.66 %
20 %	27.34 %	($6,480.99)	59.85 %
21 %	28.45 %	($7,221.48)	59.85 %
22 %	29.43 %	($8,332.22)	59.85 %
23 %	30.26 %	($9,813.20)	59.85 %
24 %	31.66 %	($9,813.21)	59.85 %
25 %	33.19 %	($9,442.96)	59.85 %
26 %	34.45 %	($9,813.21)	59.85 %
27 %	36.26 %	($8,702.47)	59.85 %
28 %	37.24 %	$16,104.07	59.85 %
29 %	38.35 %	$16,289.20	59.85 %
30 %	40.31 %	$18,695.80	59.85 %
31 %	41.70 %	$19,621.41	59.85 %



Data Mining – Profits Chart Analysis

Here's the final profit chart that has been created. From these results, we can see that in order to maximize profits, we should market only to customers with a probability threshold of roughly 60% or higher.



	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Birth Date	Marital	Gender	Email	Income	Children	Education	Occupation	HomeOwner	Number	CommuteDistar	Age	Probability to Buy
2	23-Sep-79	Single	Female	jennifer6@adventure-works.com	10000	1	High School	Manual	No	0	2-5 Miles	27	85.60%
3	9-Jun-79	Single	Female	brittney12@adventure-works.com	10000	1	High School	Manual	No	0	1-2 Miles	27	85.60%
4	25-Oct-79	Married	Female	virginia4@adventure-works.com	20000	0	Bachelors	Clerical	Yes	0	0-1 Miles	26	85.60%
5	22-Feb-79	Married	Male	calvin15@adventure-works.com	20000	0	Bachelors	Clerical	Yes	0	0-1 Miles	27	85.60%
6	25-Nov-78	Single	Male	edward28@adventure-works.com	10000	1	High School	Manual	No	0	1-2 Miles	27	85.60%
7	4-Aug-78	Married	Female	ashlee11@adventure-works.com	20000	0	Bachelors	Clerical	No	0	0-1 Miles	28	85.60%
8	1-Apr-78	Married	Female	alicia3@adventure-works.com	20000	0	Bachelors	Clerical	Yes	0	0-1 Miles	28	85.60%
9	27-May-78	Married	Female	lacey1@adventure-works.com	20000	0	Bachelors	Clerical	Yes	0	0-1 Miles	28	85.60%
10	15-May-78	Single	Female	wendy8@adventure-works.com	20000	0	Bachelors	Clerical	Yes	0	0-1 Miles	28	85.60%
11	18-Jun-75	Married	Female	jaclyn37@adventure-works.com	40000	1	Bachelors	Skilled Manual	Yes	0	1-2 Miles	31	85.60%
12	17-Nov-74	Married	Female	erika7@adventure-works.com	40000	1	Bachelors	Skilled Manual	Yes	0	1-2 Miles	31	85.60%
13	18-Sep-74	Single	Female	sarah40@adventure-works.com	40000	0	Graduate Degree	Skilled Manual	Yes	0	0-1 Miles	32	85.60%
14	21-Nov-75	Married	Male	dwayne9@adventure-works.com	40000	1	Bachelors	Skilled Manual	Yes	0	1-2 Miles	30	85.60%
15	9-Mar-75	Married	Female	casey20@adventure-works.com	40000	1	Bachelors	Skilled Manual	Yes	0	1-2 Miles	31	85.60%
16	25-Sep-78	Married	Male	alan22@adventure-works.com	10000	0	Graduate Degree	Manual	No	0	0-1 Miles	28	85.60%
17	27-Feb-74	Married	Male	victor10@adventure-works.com	10000	0	Graduate Degree	Manual	Yes	0	0-1 Miles	32	85.60%
18	22-Nov-80	Married	Female	amy13@adventure-works.com	20000	0	Bachelors	Clerical	Yes	0	0-1 Miles	25	85.60%
19	12-Oct-79	Married	Male	shawn8@adventure-works.com	10000	1	High School	Manual	Yes	0	2-5 Miles	26	85.60%
20	18-Oct-79	Single	Female	brandy8@adventure-works.com	20000	0	Bachelors	Clerical	Yes	0	0-1 Miles	26	85.60%
21	14-Mar-78	Single	Male	devin7@adventure-works.com	10000	1	High School	Manual	No	0	1-2 Miles	28	85.60%
22	16-Sep-78	Married	Male	jessie27@adventure-works.com	20000	0	Bachelors	Clerical	Yes	0	0-1 Miles	28	85.60%
23	8-Apr-77	Single	Female	andrea26@adventure-works.com	10000	1	High School	Manual	No	0	1-2 Miles	29	85.60%
24	9-Apr-77	Married	Male	troy0@adventure-works.com	20000	0	Partial College	Manual	Yes	0	0-1 Miles	29	85.60%
25	1-Oct-77	Married	Male	omar4@adventure-works.com	20000	0	Partial College	Manual	No	0	0-1 Miles	29	85.60%
26	14-Feb-77	Single	Female	jenny15@adventure-works.com	20000	0	Partial College	Manual	Yes	0	0-1 Miles	29	85.60%

We've determined which customers to market to, and how many of them we should market to in order to maximize our profits. Let's now apply these results to our list of potential customers, and identify which ones we should send the direct marketing campaign to.

Here is our list of potential customers. We're going to run a data mining prediction query against this list using our existing data mining model. In this window, we're specifying that we want to create a new output column in the list showing the probability that the customer will purchase a bike.

Here is the results of our query. Scrolling over, we can see the new column that has been added, showing the probability that each individual customer will purchase a bike. We can now sort this data so that the highest probabilities are on top.

Data Mining – How to Maximize your profits

In our previous modeling of the marketing campaign, we determined that to maximize our profits we should only market to customers that were more than 60% likely to purchase a bike. We can take that information to extract just the customers we want from this potential customer list, and only market to those individuals.

As you've seen, the new Data Mining Client for Excel reduces much of the complexity involved in performing common data-mining tasks. In just a short period of time, we've been able to perform a complete data-mining scenario, including the construction of a mining model, the modeling of a profitable marketing campaign, and the selection of which potential customers we should market to.

Contact us today to schedule a presentation and discuss your particular requirements

CMLgroup

Info@cmlgroup.com