Data Mining and Predictive Modeling with Excel 2007

Spyridon Ganas

Abstract

With the release of Excel 2007 and SQL Server 2008, Microsoft has provided actuaries with a powerful and easy to use predictive modeling platform. This paper provides a brief overview of the SQL Server system. It then discusses the "Data Mining Client for Excel 2007" and explains how actuaries can use Excel to build predictive models, with little or no knowledge of the underlying SQL Server system.

Keywords: predictive modeling, data mining, exploratory data analysis, neural networks, regression modeling

1. INTRODUCTION

Microsoft Excel is the data analysis tool most frequently used by members of the actuarial community. Spreadsheets are ideal for basic actuarial tasks, such as analyzing loss triangles or building actuarial indications. More advanced actuarial tasks, such as building predictive models and conducting advanced statistical analysis, have historically required specialized software, such as SAS, S-Plus or SPSS.

With the release of Excel 2007 and SQL Server 2008, it is now possible to build complex statistical models directly in Excel. This paper begins by describing the SQL Server Business Intelligence platform. It then discusses the "Data Mining Client for Excel 2007" and shows how actuaries can take advantage of SQL Server's advanced capabilities without leaving the familiar Excel interface.

2. SQL SERVER

SQL Server is Microsoft's enterprise-class database solution. It consists of four components: the Database Engine, Integration Services, Analysis Services, and Reporting Services. These four components work together to create business intelligence. Creating business intelligence is the goal of all actuarial work; business intelligence is knowledge that is extracted from data and delivered to the individuals who can use it to improve the company's bottom-line.

The Database Engine is essentially a grown-up version of Microsoft Access. Like Access, it stores data in tables and allows users to analyze the data using queries written in the SQL language. Unlike Access, it is a true enterprise-class database system; its capabilities are limited only by the available hardware. From a business intelligence perspective, the primary function of the Database

Engine is to store the raw data. Terms such as "data mart" and "data warehouse" refer to large collections of raw data that are stored and managed by the Database Engine. It is not uncommon for SQL Server databases to store billions of rows or hundreds of gigabytes of data.

Although data warehouses and data marts have become increasingly common, most companies continue to store data in a variety of formats. Integration Services is an extract, transform and load (ETL) tool. Its primary purpose is to transfer data between different storage formats. For example, Integration Services can pull data out of an Excel file and load it into a SQL Server table. Integration Services is also the primary data-cleansing tool. Dirty data makes it difficult to develop valid statistical models. Integration Services includes a number of data cleaning techniques, such as fuzzy logic, that can be used to clean data by identifying or removing suspicious values.

Analysis Services allows users to analyze the raw data using Online Analytical Processing (OLAP) cubes and data mining algorithms. An OLAP cube is essentially a pre-calculated pivot table. It resides on the server and stores the raw data, along with pre-calculated summarized data, in a multi-dimensional format. The data in an OLAP cube is usually viewed using an Excel pivot table. OLAP cubes are valuable because they allow users to "slice and dice" data sets that would otherwise be too large for Excel.

Analysis Services also contains the tools used to build data mining models. SQL Server 2008 includes seven data mining algorithms, such as neural networks and time series models. These algorithms allow actuaries to extract the valuable knowledge hidden within the massive amounts of raw data.

The final component of the SQL Server business intelligence platform is Reporting Services. Reporting Services is a web-based reporting tool. Essentially, it creates a web page where users can view reports that were built using the data in the SQL Server tables. It also includes a web application, known as Report Builder, which allows users to create ad hoc reports without knowing the SQL language. Reporting Services is the easiest way to distribute business intelligence to a large number of users.

The four components of the SQL Server system are tightly integrated and designed to produce enterprise-class business intelligence. While SQL Server is an extraordinarily powerful system, it is also quite complex. In 2007, the average salary for a Microsoft-certified business intelligence developer was $132,000[1]; this is the best evidence that developing enterprise-class business intelligence requires a great deal of education and experience.

3. THE DATA MINING CLIENT FOR EXCEL 2007

The complexity of the SQL Server system has discouraged actuaries from using it as a business intelligence platform. Most actuarial business intelligence is built and distributed using Microsoft Excel. Excel can store moderately large amounts of data (over one million rows per worksheet in Excel 2007). Tools such as VBA macros and the "Remove Duplicates" button provide a means of cleaning the data, while pivot tables and a respectable set of built-in functions allow actuaries to perform a wide range of analysis. Perhaps the most common method of distributing actuarial business intelligence is the simple act of e-mailing an Excel spreadsheet to the end-users.

Although Excel is the business intelligence platform of choice for most actuaries, its statistical modeling capabilities are limited. Neural networks, classification and regression trees, and other data mining algorithms simply are not available in standard Excel installations. There is a good reason for this; most data mining algorithms require fast processors and large amounts of memory, which are typically available only on servers.

The "Data Mining Client for Excel 2007" allows users to build complex statistical models in Excel, while processing those models on a high-performance server running Analysis Services. This greatly reduces the time and effort required to extract information from the raw data. It also allows actuaries with limited statistical programming experience to analyze their data using the most powerful algorithms currently available.

3.1 Installing the Data Mining Client for Excel 2007

The Data Mining Client acts as a link between Excel (which is typically installed on a laptop or desktop computer) and a server running Analysis Services. This client-server architecture allows multiple users, each working on their own desktop or laptop computer, to share the computational power of a single Analysis Services Server.

As a result, both the server and the client computers must be set up. Installing the Data Mining Client on each of the client machines and telling the client the name of the server running Analysis Services is relatively straightforward; a wizard guides users through the installation process. It is important to note that the Data Mining Client requires Excel 2007; a trial version of Excel 2007 is currently available [2]. Also noteworthy is the fact that the data mining feature is not installed by default; it must be selected from the list of installable features. The Data Mining Client is available as a free download on the Microsoft website [3].

Setting up the server can be more difficult. Analysis Services must be installed and running on

the server [4], and a user with administrator privileges must set up an Analysis Services database. When the Data Mining Client is installed, a tool called the "Server Configuration Utility" is also installed [5]. This is a wizard that allows a user with administrator privileges to set up an Analysis Services database for use with the Data Mining Client. This wizard is also used to grant users of the Data Mining Client access to the Analysis Services Database. Since many users can use a single Analysis Services database at the same time, the server only needs to be set up once.

When the server has been set up and the Data Mining Client has been installed, users can begin building data mining models. Models can be selected from either the "Data Mining" or "Table Tools" menus. These menus allow users to access the advanced data mining capability of the Analysis Services server.

3.2 Analyzing Tables using the Table Tools

Excel 2007 introduces "Tables", a new feature than allows users to quickly sort, filter and format data. Clicking the "Table" button on the insert menu creates a table. The process of creating, formatting and using tables is described in the article "Working with Tables in Excel 2007" [6].

When a cell in a table is selected, the "Table Tools" menu appears at the top of the screen. This menu contains two submenus, "Design" and "Analyze". The design menu is primarily used to format the table. One menu item of interest to actuaries is the "Remove Duplicate Records" wizard. This tool can be used to clean data by identifying and removing identical rows.

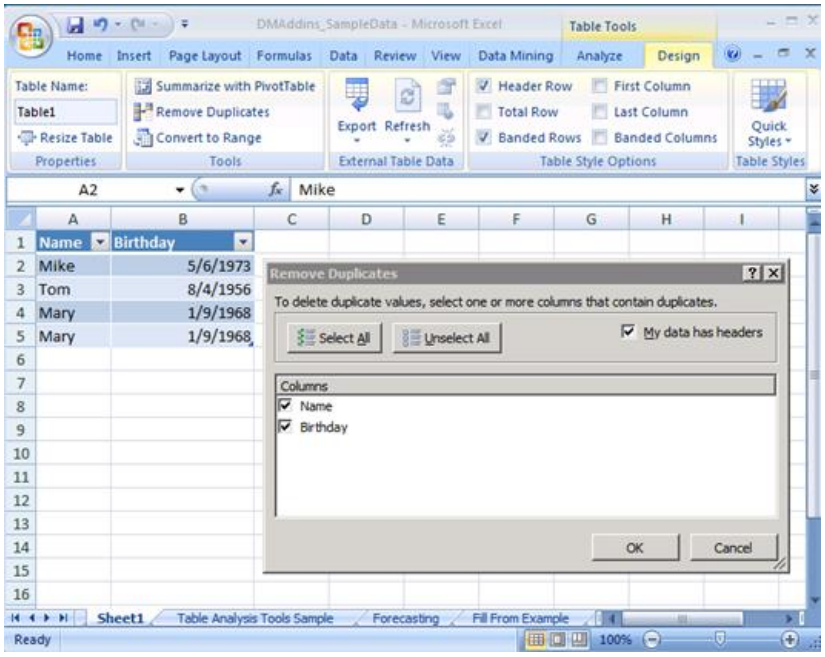

Figure 1. Removing duplicate records from a table.

The "Analyze" menu allows the user to perform eight task-oriented data mining functions. These are essentially "Black Box" functions, which allow a user to perform a task without forcing them to know anything about the underlying data mining algorithms. As seen in figure 2, these tasks are:

- Analyze Key Influences
- Detect Categories
- Fill From Example
- Forecast
- Highlight Exceptions
- Scenario Analysis
- Prediction Calculator
- Shopping Basket Analysis

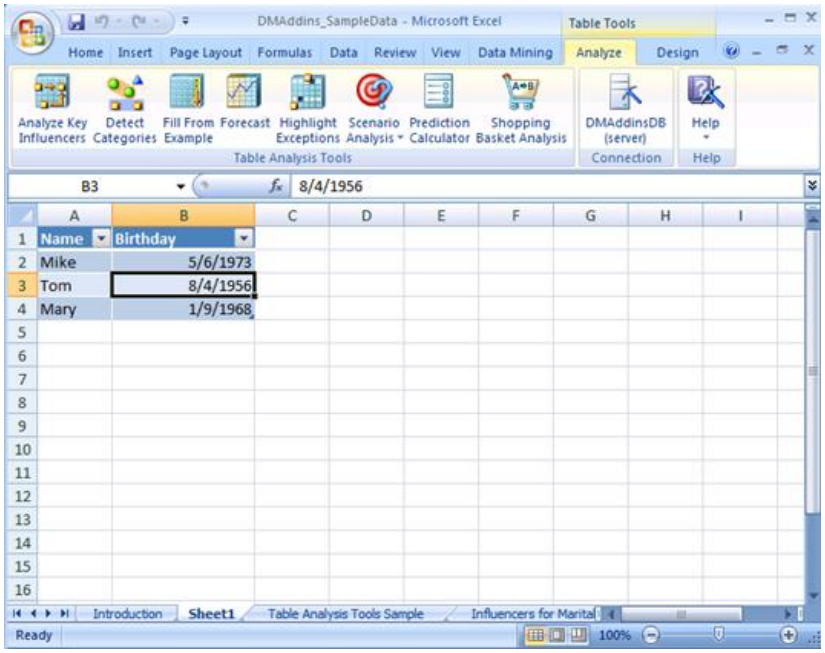

Figure 2. The Analyze menu.

3.2.1 Analyze Key Influencers

The Analyze Key Influencers button allows users to select one column in the table and then shows how the other columns are related to that column.

This can be used to predict zero-claim status for personal automobile insurance customer. For example, a table can be created that shows age, gender, marital status and if the customer had zero claims in a given time period [7]. The Analyze Key Influences tool will create a report that shows how strongly age, gender and marital status affect the likelihood of the customer having zero claims.

The report shown in Figure 3 shows the output of the Analyze Key Influences tool. In the data that was fed into this tool, married customers who are at least 51 years old are likely to have had zero claims. Single customers, especially those younger than 35, are more likely to have had a claim. This report shows that gender has no impact on zero-claim status. It also shows that age only predicts zero-claim status if the customer is younger than 36 or older than 50.

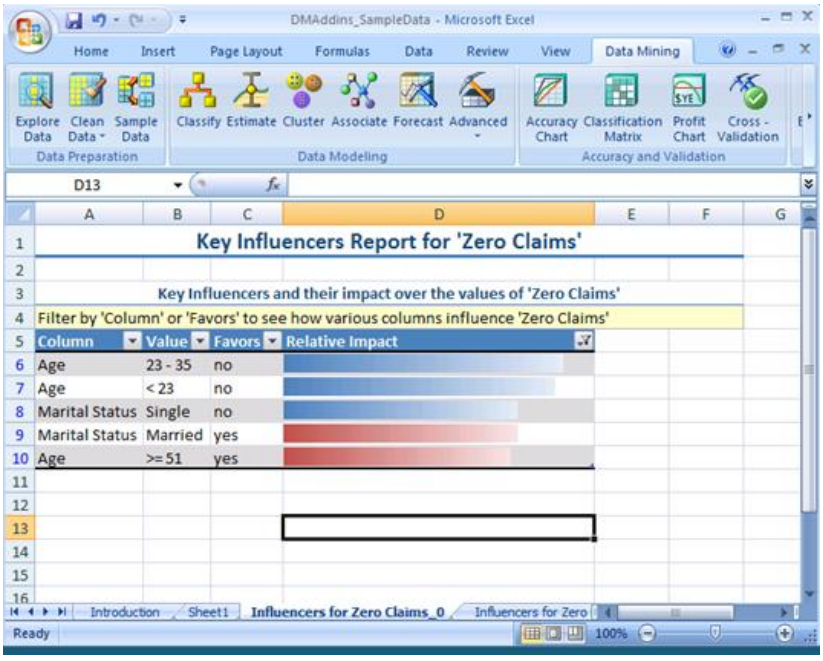

Figure 3. The output of the Analyze Key Influencers tool.

3.2.2 Detect Categories

The Detect Categories tool uses a clustering algorithm. Clustering algorithms are typically used to segment customers for marketing purposes. It identifies rows in the table that are similar and assigns the similar rows to a category. The final number of categories will depend on the number of identifiable groups of similar rows. Once the categories have been detected, a column will be appended to the table of data. The value in this column will show which category the row was assigned to.

A report containing three sections will also be created. The first section shows the number of categories and the number of rows assigned to each category. The second section of the report is similar to the Key Influencers Report. It allows a user to understand why a row was assigned to a given category. The third section of the report allows users to compare categories.

From the sample data, the Detect Categories tool found four categories in the zero-claim dataset. Looking at the key influencers and comparing the categories using the chart in the third section of the report can help explain these categories. For example, the first category contains mostly men whose age is "high" or "very high". Rows assigned to this category are clearly different from those in the second category (which primarily contains younger women).

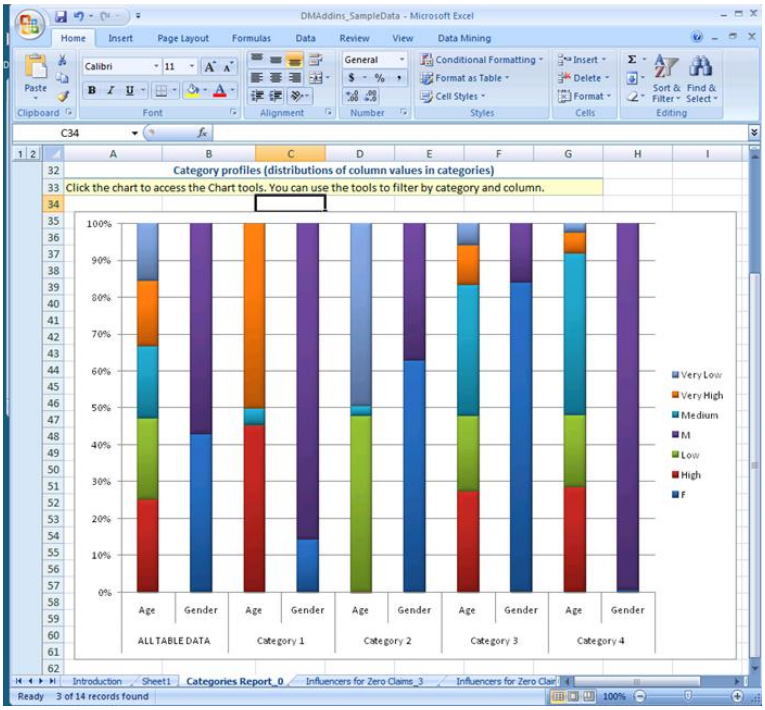

Figure 4. Comparing categories based on age and gender.

The Detect Categories tool may be a useful risk classification tool. Customers can be categorized using current or potential rating variables. Once customers are assigned to a category, the historical loss ratio of that category can be determined. If the loss ratios vary a great deal between categories, there is a strong chance that the current risk classification methodology needs to be reviewed.

3.2.3 Fill From Example

The Fill From Example tool provides a means of imputing missing data. If several rows of the

zero-claim sample dataset had no value in the married column, this tool could be used to predict the marital status of those records. A column containing the predicted values is appended to the table and a report similar to the Key Influencers report is created.

3.2.4 Forecast

The Forecast button uses a time series model to predict future values based on historical data. The algorithm automatically detects the periodicity of the data. This tool can be used by actuaries to predict the new business premium for the next quarter or the medical trend for the next year. The tool adds the predicted values to the bottom of the table and produces a line chart that shows the historical data and the predicted future values.

3.2.5 Highlight Exceptions

The Highlight Exceptions tool identifies anomalies in the data. Rows in the table that do not match the data's general patterns are highlighted. Often these anomalies are caused by data entry errors, making the Highlight Exceptions tool a powerful data-cleaning tool. In other cases, the tool identifies valid values that are simply outliers. These outliers often require further analysis before they feed into other data mining algorithms.

Outlier detection algorithms are often used to identify fraudulent claims. The technique can also be used to identify policies that are over/under priced, or policies that are rated using an inaccurate rating variables.

3.2.6 Scenario Analysis

The Scenario Analysis tools are primarily used for sensitivity analysis; the "Goal Seek" and "What-if" tools show how a model changes when the raw data is altered.

Goal Seek is used to determine which input variables need to be changed in order to reach a desired output. For example, a young customer who is judged to be a bad risk may become a good risk in a few years. The Goal Seek tool will determine how old the customer needs to be before the model judges them to be a good risk.

The What-If tool allows the user to see how the model changes if one of the variables changes. If all of the customers suddenly married, the model would be unable to use marital status to predict zero-claim status. The What-If tool allows the user to see the impact of this change on the model and its predictions.

3.2.7 Prediction Calculator

Since the cost of having an underwriter review a policy can be high, automated underwriting systems have become common. Policies that meet predetermined criteria are automatically issued, while other policies are forwarded to underwriters for review.

The Prediction Calculator tool can be used to create an automated underwriting system. If the average cost of a bad customer and the average profit from a good customer are known, the predictive model can be used to maximize estimated profits.

This is presented to the users as a series of drop-down boxes, which allow the users to select different values for each of the rating variables. Once the rating variable values are entered into the spreadsheet, the Prediction Calculator determines the odds of a true positive, a false positive, a true negative and a false negative. These odds are used to weight the costs and profits associated with each of the four possible outcomes. If the weighted average of the costs and profits is positive, the policy should be issued.

The Prediction Calculator is one example of an end-user tool that integrates data mining technology. Perhaps its greatest value to an actuary is that it clearly demonstrates how data mining can be used to improve a company's profitability.

3.2.8 Shopping Basket Analysis

The "Association Rules" data mining algorithm is used to conduct Shopping Basket Analysis. The algorithm identifies goods or services that are frequently purchased together. The output of the algorithm is a set of if-then statements along with a measure of accuracy; e.g. if a customer purchases Product A and Product B, then X% of the time they will also purchase Product C.

Shopping Basket Analysis is most frequently used in the retail industry, but it does have other applications. Healthcare organizations have used it to identify patients who are likely to have undiagnosed illnesses. Property and casualty insurers often use Shopping Basket Analysis to identify cross-selling opportunities (for example, customers who have a homeowners policy and an automobile policy are much more likely to purchase a policy for a recreational vehicle). In general, Shopping Basket Analysis provides an efficient method of analysis related transactional data.

3.3 The Data Mining Menu

The Data Mining menu is designed for users who are familiar with data mining concepts. The tools on the data mining menu are split into five sections: Data Preparation, Data Modeling, Accuracy and Verification, Model Usage, and Management. These tools allow users to complete every step in the data mining lifecycle without leaving the familiar Excel interface.

The tools available on the Data Mining menu provide full access to all the capabilities available in SQL Server Analysis Services. For example, the classify button allows users to categorize data using a CART algorithm, logistic regression, a naïve Bayes model, or a neural network. Users can also adjust algorithm parameters, such as the minimum support for a leaf in a decision tree or the number of nodes in the hidden layer of a neural network. The tools and options available in the data mining menu are much more flexible, and thus more powerful, than those available on the Table Tools menu discussed in the previous section.

3.3.1 Data Preparation

Data preparation is often the most time consuming part of the data mining process. A common rule of thumb is that 70% of a data mining project is spent cleaning data. The Data Mining Client provides Excel with some limited data cleaning capabilities. While theses tools are not as powerful as those including in a true ETL tool (such as SQL Server Integration Services), they may allow users to avoid the more complex data cleaning tools.

Traditionally, the first step in the modeling process is to look at the distribution of the data and its basic characteristics. The Explore Data button simply creates a histogram for a single column of data. This provides an easy means of identify data from a non-normal distribution.

The Clean Data button provides limited data cleansing capabilities. It allows users to identify outliers within a single column. Non-robust algorithms, such as linear regression, can be greatly influenced by outliers. The Clean Data button also allows users to re-label discrete values. For example, teenagers and twenty-year-olds could be re-labeled to have a single value for an age variable.

Sampling is the final step in the data preparation step of the data mining lifecycle. The Sample Data button can be used to select a random subset of the raw data. It can also be used to "oversample" the data. Oversampling is used when the raw data contains important, but rare, rows. Oversampling ensures that these important rows are not drowned out by less important data. This is especially important when developing models to predict rare events, such as large losses or fraudulent claims.

3.3.2 Classification Algorithms

Classification algorithms are used to separate data into distinct groups. Personal automobile clients can be classified as good drivers or bad drivers, or current customers could be classified as likely to renew their policy or likely to switch to a different company (this is known as churn modeling). The Classification button allows users to classify data using one of four algorithms: the naïve Bayes algorithm, logistic regression, decision trees, or neural networks.

The naïve Bayes algorithm is the fastest classification algorithm, but it often misses complex patterns that can be found by the other algorithms. The algorithm calculates conditional probabilities and then uses Bayes Theorem to predict the data's class. Since this involves relatively simple calculations, the naïve Bayes algorithm is often used on very large data sets or when a model needs to be trained very quickly.

Logistic regression is used when then classification is binary (i.e. true vs. false, yes vs. no, good vs. bad). Logistic regression is a type of generalized linear model. The output of a logistic regression is a number between zero and one, which represents the probability that the data falls into the "True" category.

Decision trees are often the preferred classification algorithm. Decision trees can capture more complex relationships than the naïve Bayes algorithm, and are not limited to binary classes. They are also not "black box" algorithms; the output of a decision tree algorithm is a set of rules that any individual can easily understand and use to classify data. These rules are often presented in the form of a tree-like graph, which give the algorithm its name.

Neural networks are the most advanced classification algorithm available. Neural networks can find complex nonlinear patterns within the data. Neural networks will often provide the best predictions. However, they can take a great deal of time to train and are "black box" algorithms. Explaining the reason behind a neural networks prediction can be difficult or impossible. This has prevented neural networks from being used in highly regulated insurance environments.

There are countless insurance applications of classification models. They can be used to identify fraudulent claims, or used to identify insureds who are likely to leave for another company. One popular use of classification models is to automatically sort potential customers into high risk or low risk groups. A new policy that is classified as high risk can be referred to an underwriter for review, while the low risk policies can be issued automatically.

3.3.3 Estimation Algorithms

Estimation algorithms are used to predict continuous values. They can be used to predict the lost

cost of a specific workers compensation policy or the number of accidents a driver will get into during the next year. SQL Server uses four estimation algorithms: decision trees, linear regression, logistic regression and neural networks.

Decision trees, linear regression and logistic regression are all special cases of the Classification and Regression Tree (CART) algorithm. The decision tree algorithm is the most general form of the CART algorithm and is also used as a classification algorithm. When a decision tree does not contain any "branches" (i.e. binary splits, such as separating the data into male and female subsets), the decision tree algorithm produces a liner regression model[8].

When estimating a binary value (true/false, yes/no, etc.), the desired output is a probability that the value is true/yes, etc. This probability is measured as a percent or as a value between 0 and 1. Logistic regression is often used to model binomially distributed data[9]. A logistic regression model is essentially a linear regression model, where the dependent variable is transformed using the logit transformation, logit=Ln(p/1-p). This transformation causes the model to output a value between 0 and 1, making it possible to model the probability of a true/yes outcome.

The same features that make neural networks powerful classification algorithms allow them to accurately estimate continuous data. However, the previously discussed limitations of neural networks limit their use as an estimation algorithm. In spite of their "black box" nature, neural networks can be used to confirm the results of other algorithms or to solve problems where the results, rather than the logic behind the results, are all that matters.

3.3.4 Clustering Algorithms

Classification and estimation algorithms are two classes of supervised algorithms. The term "supervised" is used to describe data mining algorithms that model a pre-selected dependant variable. "Unsupervised" algorithms, such as clustering algorithms, look at all of the available data in order to identify patterns. All patterns, rather than just those affecting the dependant variable, are reviewed and analyzed.

Clustering algorithms try to split records into similar groups. For example, clustering a video store's rental database might show three major groups of customers: those who enjoy action films, those who like romance movies, and customers who rent cartoons. This type of analysis can allow businesses to understand who their clients are and how to better serve them.

The Data Mining Client for Excel make it easy to interpret the groups identified by the clustering algorithm. The Browse window offers four view of the clusters. The first view is a cluster diagram,

shown in Figure 5.

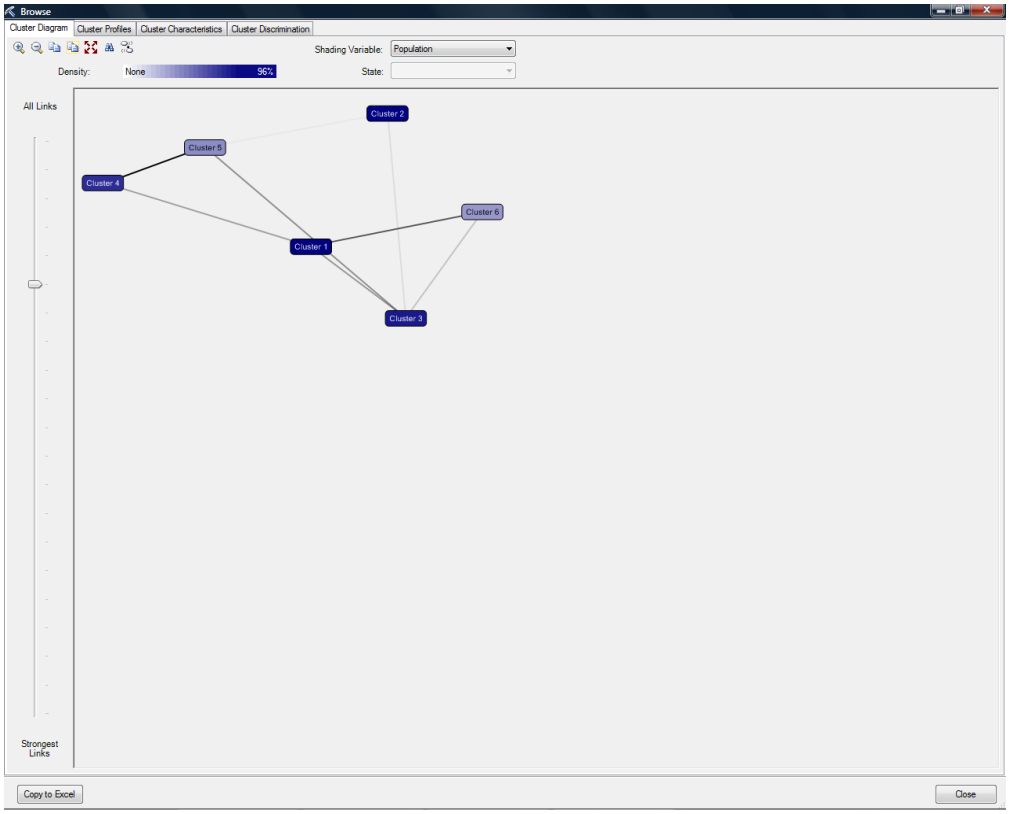

Figure 5. A Cluster Diagram.

A cluster diagram allows the analyst to see the relationship between various clusters, based on different attributes. The darkest line in Figure 5 shows that clusters 4 and 5 are most similar, while the various shades of color show how the cluster are related with respect to the "population" variable.

The second tab in the cluster viewer is shows "cluster profiles", which are essentially a univariate statistical analysis of each of the variables, for both the overall population and each cluster individually. This view makes it possible to identify the differences between the clusters.

The third tab, called "cluster characteristics", makes it possible for an analyst to understand the nature of a given cluster. Characteristics are values of a given variable that help distinguish one cluster from another. For example, if one cluster contains only young men while another contains women of any age, an analyst may consider gender to be the "most interesting" characteristic. The final tab, "cluster discriminants", uses these cluster characteristics to show the differences between two chosen clusters.

Clustering is a valuable data mining technique. While it is most commonly used by marketers to

segment a business's customers, it is important to remember that clustering can be used to pre-process data. By reducing a large number of variables into a single variable (the cluster names), clustering can make computationally intractable data mining problems solvable.

3.3.5 Time Series

Time series are statistical models that represent data that is collected over time. These models usually decompose data into a series of periodic components. For example, the daily sales at a pizzeria will depend on the day of the week (Fridays are busy, Mondays are slow), the day of the month (customers tend to eat out on the 1st and the 15th of the month, when they get paid), the season (summer is busy, winter is slow) and the year (inflation increases the dollar amount of the sales). A time series analysis will identify patterns like these and allow the analyst to forecast future data points.

3.3.6 Association Algorithms

The Associate button is used to create a shopping basket analysis. It performs the same analysis as the "Shopping Basket Analysis" button described in section 3.2.8, but allows the user to change more of the algorithms parameters.

3.3.7 Accuracy and Validation

The four buttons in the Accuracy and Validation section of the data mining toolbar allow analysts to evaluate the quality of a data mining model. Models can be evaluated using accuracy charts, a classification matrix, profit charts or the cross-validation method.

Figure 6 shows an accuracy chart. Also known as a lift chart, this graph shows how a chosen model compares to a perfect model and a model based on random guessing. Figure 6 was produced by applying the decision tree algorithm to the "Bike Buyer" data in the sample file that is installed with the Excel Data Mining Add-In. This lift chart shows how accurate the data mining model is; in this example, the 5% of individuals whom the model selects as most likely to buy a bicycle represent 22.91% of the total population of individuals who actually buy a bicycle.

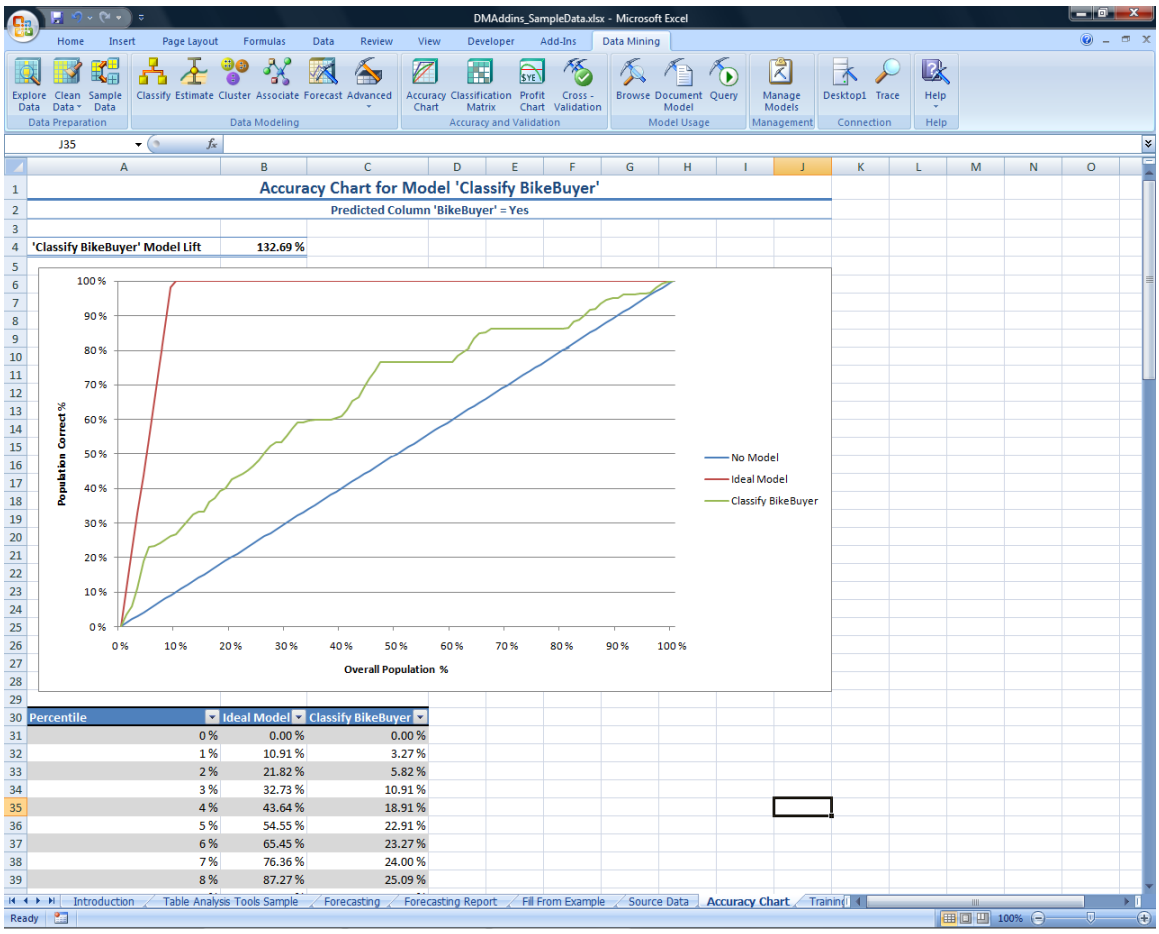

Figure 6. Using an Accuracy Chart to Evaluate a Data Mining Model

The Classification Matrix button produces a pivot table which shows the number of records that are correctly or incorrectly classified. For example, the profit chart button can be used to assist marketers planning a mailing campaign. The profit chart would ask for the cost of sending out a mailing and the profit from a successful mailing. It then uses the data mining model to determine the optimal number of mailing to send out.

The Cross-Validation button is the final means of evaluating a data mining model. Cross-validation randomly splits the data into disparate training and testing subsets. The training data is used to parameterize the mining model, while the testing data is used to measure the accuracy of the model. This process is repeated a number of times and the results of all of the models can then be reviewed. Variance among the outputs of these models is a sign that the models are over-fitting the data, while similar outputs are a sign that the model is an accurate depiction of the patterns found within the data.

These Accuracy and Validation tools often show that a data mining model does not meet the user's needs. An analyst may need to try other modeling algorithms, or conduct more data cleaning, or even add additional data in order to reach the desired level of accuracy and applicability.

3.3.8 Managing the Data Mining Models

Models that are built using Excel can be temporary or permanent. Temporary models disappear when Excel is closed. Permanent models are saved in the Analysis Services database.

The Manage Models button is used to delete, export or reprocess permanent models. A data mining model can only be accurate if the data used to train it is similar to the data it is trying to predict. If new or better data becomes available, the model should be reprocessed.

The model viewer that opens after a model is created can be reopened using the Browse button. This tool allows users to understand how the model is making it's predictions. The Document Model button copies some of the information presented in the model viewer into an easy-to-read Excel spreadsheet. It also lists the algorithm's parameters and basic information about the columns of data used in the model.

3.3.9 Using a Model to Make Predictions

The final step in the data mining process is applying the model to new data in order to make predictions. The Query button opens a wizard that allows users to select the model, new data and output for a prediction query. This wizard creates a prediction query, which runs new data through the trained model. The results of a prediction query include the prediction, the probability that the prediction is correct, and the prediction's support (the number of rows in the training data set that suggest this prediction is correct).

The query button is important because it allows Excel users to easily run new data through a data mining model on the SQL Server Analysis Services system,. Allowing end users to "score" their own data is a valid and effective means of deploying a data mining model. The Data Mining Extensions (DMX) code that is produced by the Query button can also be used for other purposes, such as building a Visual Basic application that returns data from the mining model.

4. CONCLUSION

Predictive modeling has become an essential actuarial skill. Since the value of a model is determined by both its quality and the amount of time required to build it, the ability to build models quickly may become a sustainable competitive advantage for some insurance companies.

The Data Mining Client for Excel 2007 allows actuaries to build a variety of powerful models very quickly. While specialized statistical packages will continue to serve as the primary data mining platforms at most companies, Excel 2007 and SQL Server Analysis Services certainly deserve a place in an actuary's toolbox.

5. REFERENCES

[1] Michael Domingo, RedmondMag.com 2007 Salary Survey, http://redmondmag.com/salarysurveys/
[2] An Excel 2007 trial edition is available at: http://us1.trymicrosoftoffice.com/product.aspx?sku=3203819&culture=en-US
[3] The SQL Server 2008 version of the data mining client for Excel 2007 is available at: http://www.microsoft.com/downloads/details.aspx?familyid=896A493A-2502-4795-94AE-E00632BA6DE7&displaylang=en
[4] A trial edition of SQL Server 2008 is available at: http://www.microsoft.com/sqlserver/2008/en/us/trial-software.aspx
[5] This video explains the process of setting up a SQL Server Analysis Services database using the Server Configuration Utility: http://www.microsoft.com/sql/technologies/dm/DMAddinConfigure/DMAddinConfigure.html
[6] JKP Application Development Services, "Working with Tables in Excel 2007", visited on 9/8/2008 http://www.jkp-ads.com/articles/excel2007tables.asp
[7] The author created the zero-claim dataset. It is not based on any real-world data and therefore the sample "results" are completely fictional.
[8] ZhaoHui Tang & Jamie MacLennan, *Data Mining with SQL Server 2005*, page 153
[9] Fred L. Ramsey & Daniel W. Schafer, *The Statistical Sleuth - A Course in Methods of Data Analysis* , page 614

Abbreviations and notations

CART, classification and regression tree
ETL, extract, transform and load
OLAP, online analytical processing
GLM, generalized linear models

Biography of the Author

Spyridon Ganas is an actuarial analyst at a non-profit healthcare organization. His work focuses on building actuarial business intelligence tools using the SQL Server platform. He has a bachelor's degree from Babson College and read applied statistics at St. Cross College, Oxford. He can be contacted at spyridon.ganas@stx.oxon.org