Churn analysis using data mining

Churn is the process of customer turnover. Churn is not restricted to the telecommunications market—it occurs wherever stiff competition provides incentives for customers to switch providers. Industries such as credit card issuers, insurance companies, and ISPs are all subject to increasing levels of churn. ISPs have hinted at churn rates as great as 50 percent, while the mobile telecommunication rate appears to vary around the 30 percent mark. As the market plateaus in these industry sectors, the problem of churn is becoming ever more critical.

Churn analysis using data mining

In order to combat the high cost of churn, increasingly sophisticated techniques can be employed to analyze why customers churn and which customers are most likely to churn in the future. Such information can be utilized by marketing departments to better target recruitment campaigns and by active monitoring of the customer call base to highlight customers who may, by the signature in their usage pattern, be thinking of migrating to another provider.

Data mining techniques bring powerful modeling analytics to the identification of the “Why” and “Most Likely” groups. This paper describes a typical application of such techniques to a mobile telecommunication provider’s customer base.
Customer churn analysis with IBM® SPSS® Modeler

IBM SPSS Modeler is a knowledge discovery workbench recognized as an industry leader in the integration of data mining technologies with ease of use. Based on an intuitive visual data flow process of analysis, models and reports can be created to tackle otherwise difficult and time consuming problems. Consider the following dataset:

The data represents a sample of customers with descriptors such as number of minutes on long distance and local calls, financial information such as billing types and payments methods, some demographics such as age, sex, income. The final field “Churned,” indicates whether the customer has ceased usage on their own accord (Vol), or is still a current paying customer (Current). The InVol customers are those who have been removed by the telecommunications provider for reasons not due to the churn we are interested in, i.e. payment defaults, etc. The task is to profile those customers most likely to be voluntary churners.

This data shows us a variety of customer information. The last field, Churned, indicates what the current status of each customer is. We will use this information to find which customers are most likely to be voluntary churners.
Types of analysis

Decision trees and neural networks are the staples of advanced knowledge discovery workbenches. These techniques can be effectively applied to the churn problem. Neural networks are able to uncover complex patterns in the types of customers and rank the customer base based on a score, or likelihood, to churn. Although powerful, neural networks are difficult to interpret and do not readily give up the process by which a customer has been scored.

Decision trees, on the other hand, build very open and interpretable models that show the analyst the patterns discovered. For example, a decision tree algorithm may discover the following rule.

*If international call time is 10 minutes and the long distance bill type is standard, then churn likelihood is high.*

This rule can lead to actionable knowledge: it may be worth calling this customer and suggesting a change of billing type to one more suited to this level of international calling.

Neural networks, decision trees and even clustering algorithms can be helpful in the analysis of customer churn.

Feature selection

In any data analysis there is a process. Specifically:

- **Business question definition** – In this case churn analysis: Who is churning? And can we develop an action plan to combat that churn?
- **Data gathering** – Build the data sample. Summarize customer calling behavior and join with relevant demographics and billing information
- **Feature selection** – A combination of identifying poorly populated (noisy?) attributes and attributes not deemed relevant to the task at hand. This is done by visually inspecting the data and, together with a domain expert, identifying those attributes most likely to impact the churn decision
- **Modeling** – Build the models
- **Analysis and report generation**

IBM SPSS Modeler supports all aspects of the churn analysis process apart from the business question definition. This has to be defined by domain experts, because the actual business problem drives the data mining task. Once data has been gathered from various database sources, IBM SPSS Modeler can be used to visualize and explore. Histograms, scatterplots and association graphs can be created to help uncover patterns. The following picture shows an association graph (sometimes known as a web diagram).
The darker the line connecting any two attribute values, the stronger that association. It can be seen that females seem likely to churn (Vol), while standard long distance billing types seem more related to not churning (Current) as does free local bill types. This type of visual exploration of the data is often invaluable in allowing the business user to uncover patterns that the analyst “knew” was in the data – “I always thought our highest churn rate was on that type of billing plan.”

This is an association graph, otherwise known as a web diagram. This lets you do visual exploration to understand the patterns in your data.

After exploring the data, the next stage is to begin modeling. The process of feature selection is important. The problem is that some modeling algorithms cannot cope with large numbers of attributes. In many real datasets, a user may have hundreds of attributes, and this can lead to poor performance of the modeling algorithm. One technique is to reduce the number of attributes to the core set of highly indicative attributes by using decision trees. Decision trees (such as C5.0 within IBM SPSS Modeler) do not have a problem with large numbers of attributes.
The screenshot below shows an IBM SPSS Modeler visual data flow used to create the graphics and report included within this paper.

Once the core set of attributes has been selected (a three-step process in IBM SPSS Modeler) a neural network (or other scoring algorithm) may be trained using the remaining attributes. Once trained (again an easy process in IBM SPSS Modeler – as the setting up and training mechanism are all selected automatically by the system) the customers can be sorted into a ranked list of high probability to churn down to the least likely to churn (the loyal customers).

**Value and profitability analysis**

Even when scoring models have been built, the analysis is by far from over. In some ways it has only just begun. The reason is that there are many ways to use a churn-scoring model – all governed by the business task and problem being tackled. If the goal is to “touch” all customers no matter the cost, then we don’t even need a model. If the goal is to maximize the return on a marketing budget of a fixed amount, then modeling is critical. Other issues such as customer value begin to play a part. What to do with customers who are likely to churn but do not in fact use their mobile phone? If the telecommunications provider is not making money from this customer, should they not just let this customer go? This problem is perhaps even more apparent with credit card issuers.
In IBM SPSS Modeler, reports can be created that indicate where to draw the line in customer targeting campaigns, taking into account propensity to churn and customer value (profitability). The following chart indicates a maximum net return in a targeting campaign.

Here, we drew a line to indicate a maximum net return in a targeting campaign.
The middle line shows the net campaign gain based on parameters such as the cost of contacting each customer and the discount rate offered to entice that customer to stay. Other factors relevant to the telecommunications company’s particular business problem can easily be added into the IBM SPSS Modeler analysis.

The final report may be a list of customer IDs and a predicted maximum campaign gain, such as the following example. As can be seen, a gain of US $128,226 is made by targeting (touching) 50.1 percent of the customer base. (Note only a partial list of “hot” ID’s is actually shown)

Finally, IBM SPSS Modeler can generate a list of customer ID’s who are the best prospects for a future promotional campaign.
About IBM Business Analytics
IBM Business Analytics software delivers complete, consistent and accurate information that decision-makers trust to improve business performance. A comprehensive portfolio of business intelligence, predictive analytics, financial performance and strategy management, and analytic applications provides clear, immediate and actionable insights into current performance and the ability to predict future outcomes. Combined with rich industry solutions, proven practices and professional services, organizations of every size can drive the highest productivity, confidently automate decisions and deliver better results.

As part of this portfolio, IBM SPSS Predictive Analytics software helps organizations predict future events and proactively act upon that insight to drive better business outcomes. Commercial, government and academic customers worldwide rely on IBM SPSS technology as a competitive advantage in attracting, retaining and growing customers, while reducing fraud and mitigating risk. By incorporating IBM SPSS software into their daily operations, organizations become predictive enterprises – able to direct and automate decisions to meet business goals and achieve measurable competitive advantage. For further information or to reach a representative visit www.ibm.com/spss.

© Copyright IBM Corporation 2010
IBM Corporation
Route 100
Somers, NY 10589
US Government Users Restricted Rights - Use, duplication of disclosure restricted by GSA ADP Schedule Contract with IBM Corp.
Produced in the United States of America
May 2010
All Rights Reserved

IBM, the IBM logo, ibm.com, WebSphere, InfoSphere and Cognos are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. If these and other IBM trademarked terms are marked on their first occurrence in this information with a trademark symbol (® or TM), these symbols indicate U.S. registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the web at “Copyright and trademark information” at www.ibm.com/legal/copytrade.shtml.

SPSS is a trademark of SPSS, Inc., an IBM Company, registered in many jurisdictions worldwide.

Other company, product or service names may be trademarks or service marks of others.

Please Recycle