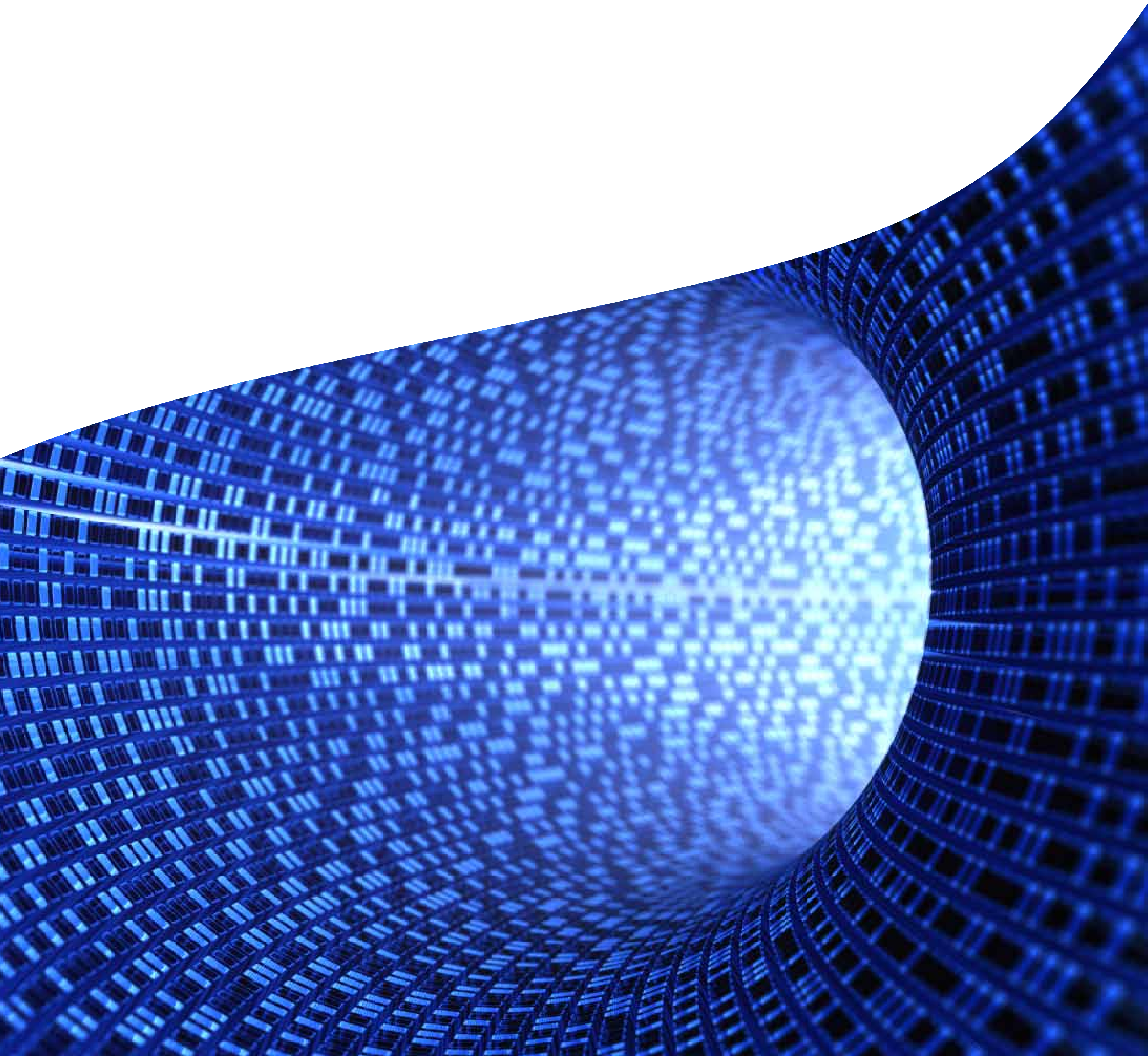
Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

Financial Services | the way we see it

Predictive Modeling Using Transactional Data



Contents

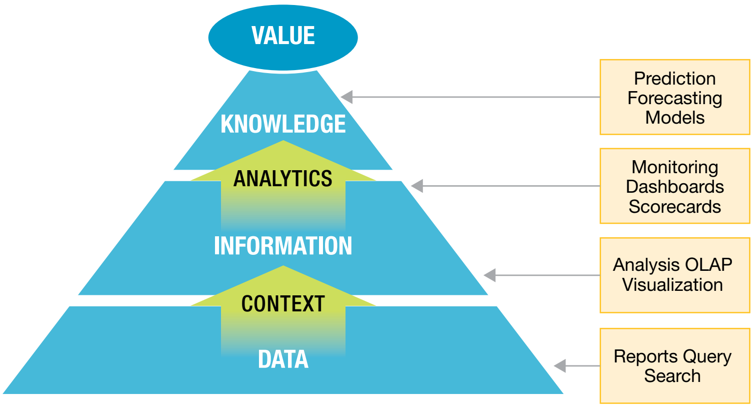
1 Introduction

The real benefit of analytics is in using past data to forecast or predict future events, providing firms with a strategic capability to be proactive.

In a world where traditional bases of competitive advantages have dissipated, analytics driven processes may be one of the few remaining points of differentiation for firms in any industry[1]. This is particularly true in financial services, which has progressed rather fast along the analytical path in the last couple of decades.

Analytics can be used to slice and dice historical data to analyze past performance and to produce reports. Here analytics helps firms react to past events. The real benefit of analytics is in using past data to forecast or predict future events, providing firms with a strategic capability to be proactive.

Figure 1: Reactive vs. Proactive Decision Making

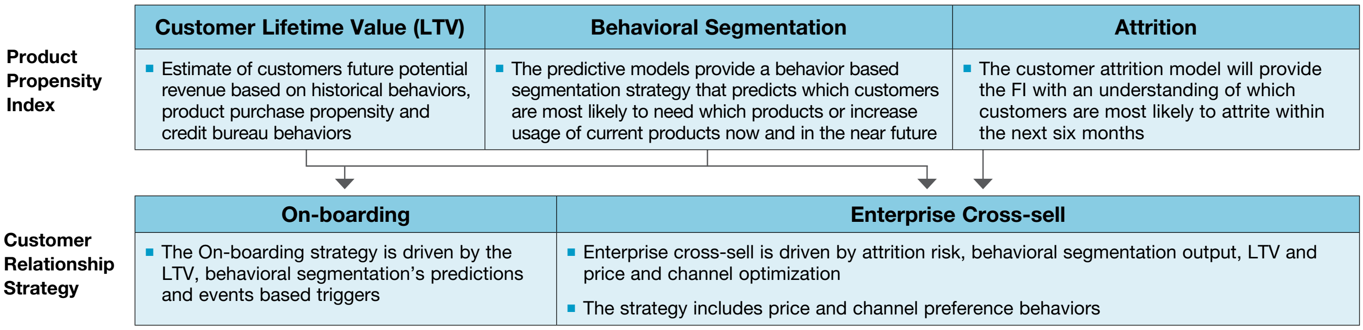

VALUE
KNOWLEDGE
ANALYTICS
INFORMATION
CONTEXT
DATA
Prediction Forecasting Models
Monitoring Dashboards Scorecards
Analysis OLAP Visualization
Reports Query Search

Source: Capgemini

Predictive modeling involves creating a model that outputs the probability of an outcome given current state values of input parameters. In banking and insurance industries, it is typically used in the context of predicting customer behavior. Historical data related to past customer activity is used to create a predictive model that captures attributes which seem to have greatest influence on future customer activity.

This provides marketing departments with a great tool to optimize their marketing campaigns, channel performance, customer on-boarding and cross-sell. These are typically driven by predictive models for customer life-time value, behavioral segmentation and attrition.

Figure 2: Customer Strategy driven by Predictive Analytics



	Customer Lifetime Value (LTV)	Behavioral Segmentation	Attrition
Product Propensity Index	▪ Estimate of customers future potential revenue based on historical behaviors, product purchase propensity and credit bureau behaviors	▪ The predictive models provide a behavior based segmentation strategy that predicts which customers are most likely to need which products or increase usage of current products now and in the near future	▪ The customer attrition model will provide the FI with an understanding of which customers are most likely to attrite within the next six months

	On-boarding	Enterprise Cross-sell
Customer Relationship Strategy	▪ The On-boarding strategy is driven by the LTV, behavioral segmentation's predictions and events based triggers	▪ Enterprise cross-sell is driven by attrition risk, behavioral segmentation output, LTV and price and channel optimization ▪ The strategy includes price and channel preference behaviors

Source: Capgemini

[1] Competing on Analytics: The New Science of Winning by Thomas H. Davenport, Jeanne G. Harris. Harvard Business School Press

2 Using Transactional Data

Transactional data potentially offers additional levels of insight into customer's activity, but poses some challenges that need to be addressed before analytics can derive valuable insights from it.

A customer's historical activity typically comprises of a few accounts and transactions around those accounts. For example, a customer may have a checking and savings account, a mortgage loan and a credit card from a bank. Banks also offer services like Electronic Bill Pay (EBP) and ATM/debit cards which generate Electronic Funds Transfer (EFT) transactions.

Data associated with accounts are typically stored in an Accounts Processing (AP) system. They may contain transactions, but AP systems usually carry only the last month's history. Prior months' transactions are reflected in monthly balance snapshots.

Unlike AP data, transaction data is typically maintained as is in corresponding transaction processing systems, whether it is EBP or EFT. Banks may have many months or years worth of daily transactional data archived and stored. Therefore, transactional data potentially offers additional levels of insight into customer's activity.

The richness of transactional data poses some challenges that need to be addressed before analytics can derive valuable insights from it. The rest of this paper details these challenges and possible solutions by referring to a case study as an illustrative example.

3 Data Quality

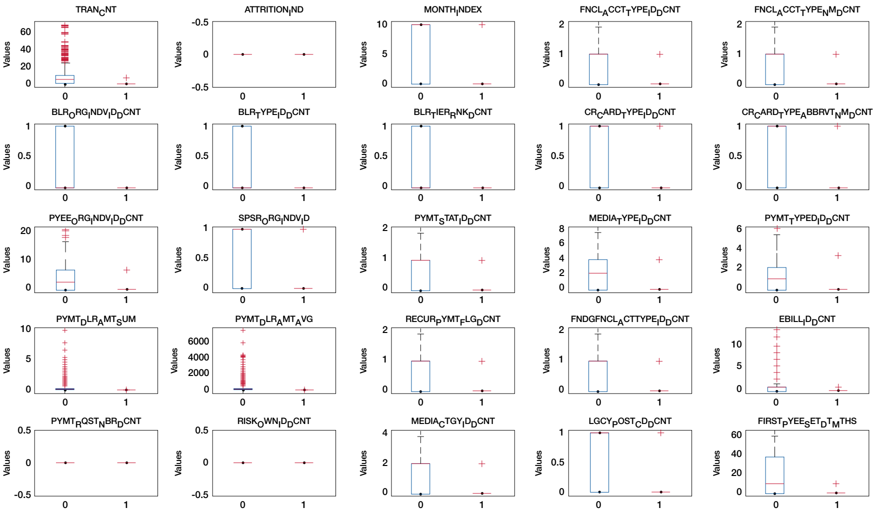
As with any kind of data for any kind of analytics, data quality is the first issue to be tackled. In order to understand the structure of data and identify issues, the key steps are to perform data profiling and exploratory data analysis.

3.1. Data Profiling

Data profiling involves creating summary statistics for each and every column and looking at simple plots of the data to identify trends, clusters or outliers. Summary statistics can include count, number of missing records, mean / mode / median values, ranges and quartiles. Box plots are useful tools to visualize some of this information graphically.

Data profiling helps understand which columns warrant additional attention from data quality perspective. The appropriate course of action for each column has to be carefully determined. For some columns, missing values may be replaced by mean or mode or a constant. Some columns may need to be simply dropped from analysis.

Figure 3: Box Plots to identify clusters and outliers



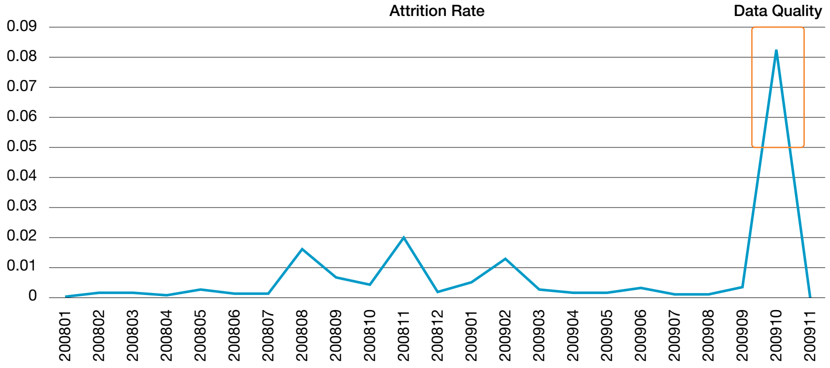
Source: Capgemini

The next step is to look further into the columns at the values represented by the data and identify any inconsistency. For example, in a transaction file, the transaction date cannot be earlier than the customer's account start date. There may also be subtle issues that cannot be caught by such logic, but can be observed simply by plotting the corresponding attribute. As an example, the plot below shows the number of customers who attrited each month from a bank.

In this case, the spike was caused by default values entered for some customers whose data was migrated from one source system to another. The resolution in this case was to not rely on the end date provided in the data column, but to define attrition as a period of inactivity as depicted by the transaction data.

This definition also opens up the possibility of defining and detecting lower levels of customer engagement that typically precedes attrition. Instead of defining attrition as period of no activity, it could be defined as a period of declining activity.

Figure 4: Data Quality issue identified using a trend plot



Source: Capgemini

3.2. Exploratory Data Analysis

In exploratory data analysis, data is examined further to identify attributes that seem significant or anomalous. This step also involves creating derived attributes by applying transformations to original data columns. The simplest of such transformations would be computing an Age attribute from a Birth-Date column by differencing against current date.

For transactional data, this step often implies rolling up daily transactions into a weekly or monthly aggregate for analysis purposes. For example, EBP data which contains daily bill-pay transactions for all customers can produce an aggregation of monthly transactions for each customer per month. These can include count of transactions, total dollar amount of transactions, average dollar amount of transactions. If individual transactions had flag values associated with them, then an aggregate count of flag value occurrences might make sense.

While modeling customer attrition, one of the first steps is to look at periods of inactivity to determine the appropriate definition of attrition. This is sometimes referred to as activity analysis. The example analysis below can be extended to determine that 3 or more consecutive months of inactivity can be considered as attrition, and customers with more than 25 transactions per month can be classified as small businesses.

Figure 5: Activity analysis to determine attrition definition

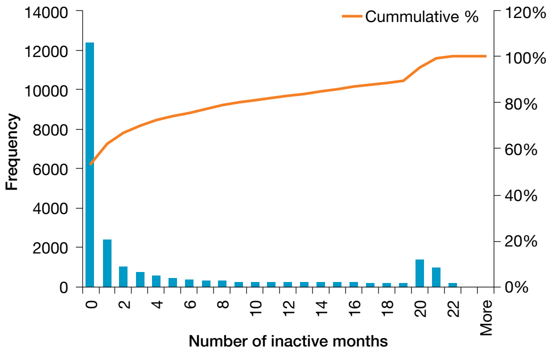
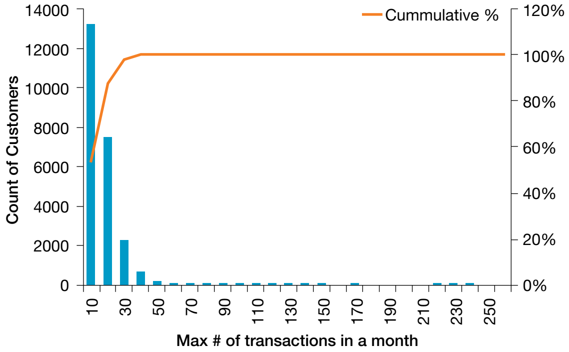


Figure 6: Activity analysis to identify small business customers



Source: Capgemini

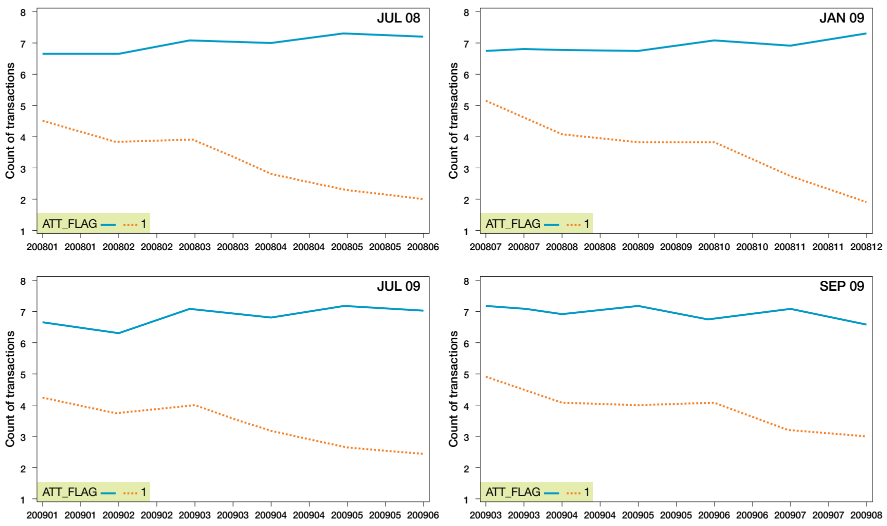
4 Cohort and Trend Analysis

Once a prediction segment has been defined (e.g. attriter or high transactor), the next step is to look at groups of customers that belong to that segment. In the case of an attrition model, we can identify customers who attrited in each month and bucket them into a cohort. For example, JAN09 cohort would be customers whose three consecutive months of inactivity started in January 2009. This approach leads to a cohort for nearly every month of data in consideration.

It is possible that each cohort is different – i.e. customers who attrited in one month exhibit different behavior than customers who attrited in another month. Unless there are seasonal effects, it is usually unlikely that cohorts are significantly different from each other. To confirm this, one can compare some attributes of attriters and non-attriters from different cohorts.

In the example below, average monthly transaction counts of attriters and non-attriters are plotted for 12 months prior to month of attrition for the cohort. The four months chosen are Jul 2008, Jan 2009, Jul 09 and Sep 2009.

Figure 7: Cohort analysis to compare behavior across cohorts



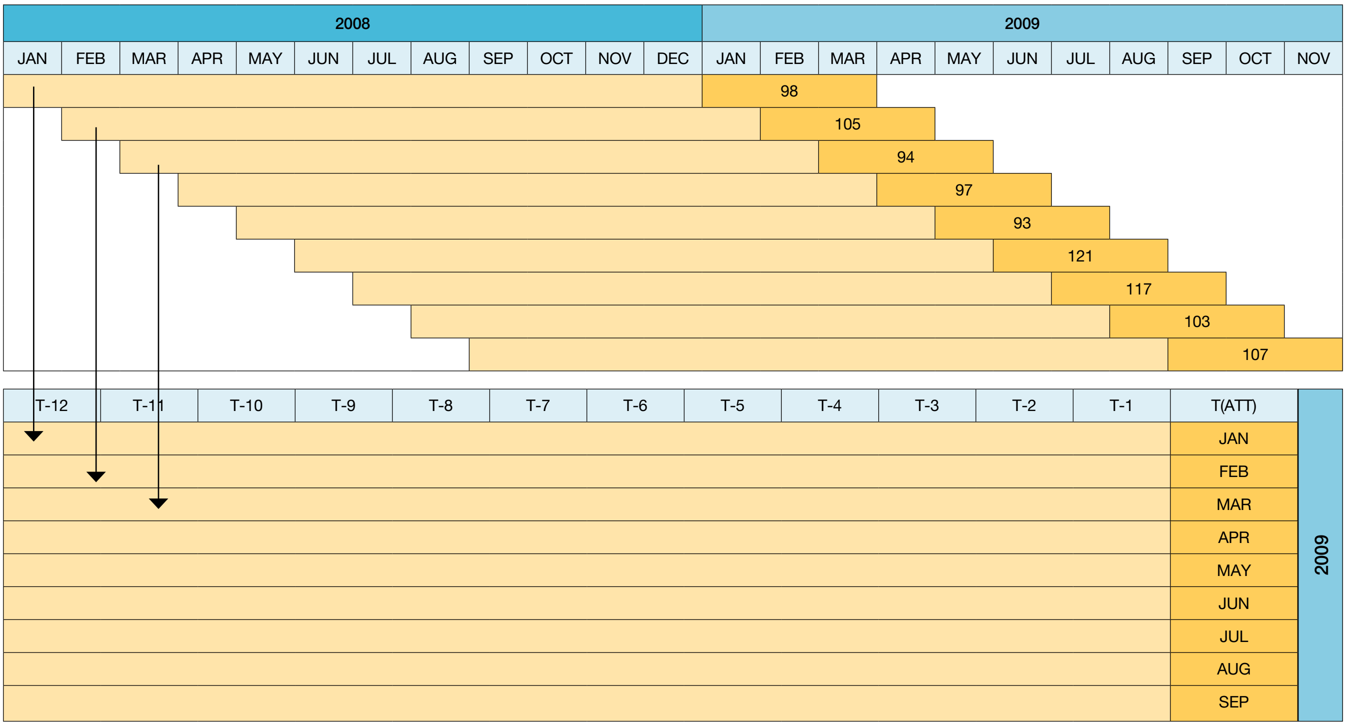
Source: Capgemini

The plots indicate that there is no significant difference between cohorts – whether it is across years or across months. In each case, there is a difference in level of activity between attriters and non-attriters. Also, attriters tend to show declining activity in months close to attrition. These patterns are consistent across all cohorts.

These observations allow one to combine all cohorts into one single large segment of attriters. While combining cohorts, care has to be taken so that monthly activities are tagged correctly with respect to the month of attrition. If a customer attrited in Jul 2009, his activity in Jun 2009 will be tagged T-1 and activity in May 2009 will be tagged T-2. Similarly, for someone who attrited in Jan 2009, activity in Dec 2008 will be tagged T-1 and activity in Nov 2008 will be tagged T-2. Once these tags are in place, all activity in T-1, T-2 and so on can be aggregated across cohorts.

For example, in the first diagram below, JAN09 cohort had 98 attriters, FEB09 cohort had 105 attriters and so on. Each cohort has 12 months of history that is considered for analysis. When aggregated, the cohorts stack up as shown in the bottom diagram.

Figure 8: Aggregating across cohorts



5 Model Variable Definition

Once cohorts are analyzed and combined (if appropriate), the next important step is to define the set of variables that will be used for modeling.

One obvious set of variables are those associated with the customer and not with the transactions. These are demographic type of information like Gender, Age, Location, Marital Status etc. They fluctuate very little over time (except age, which steadily increases) and are sometimes referred to as stock variables.

While dealing with transactional data, it is useful to look at trends to identify patterns of customer behavior across time, as shown in the cohort analysis section. Such attributes are often referred to as time-varying attributes or flow variables. Since flow variables exhibit high volatility, they are typically aggregated rather than used as is in models.

Linear trends in flow variables can be captured using two types of variables – one to capture the level of activity (sometimes referred to as intercept) and one to capture the trend itself (sometimes referred to as slope). Below is a summary of the types of variable and the analysis performed on each one.

Figure 9: Types of variables and analysis

Variable Type	Description	Type of Analysis	Example	Used for Modeling
Stock Variable	Static value for customer during the analysis period	Distribution	Age	YES
Flow Variable	Value changes month-to-month	Time Series (Trend)	Monthly Transaction Count	NO
Aggregated Flow Variable	Capture intercept and slope of flow variable trend ▪ Average ▪ Average (M-M Difference) ▪ Average of Directional Flag ▪ M12 – M1 ▪ M9 – M1	Distribution	Average Monthly Transaction Count	YES

6 Model Selection

Once model variables are defined, various kinds of models can be created. The most common ones for predicting customer behavior are logistic regression and decision trees. These models can be easily created using a tool like SAS or SPSS.

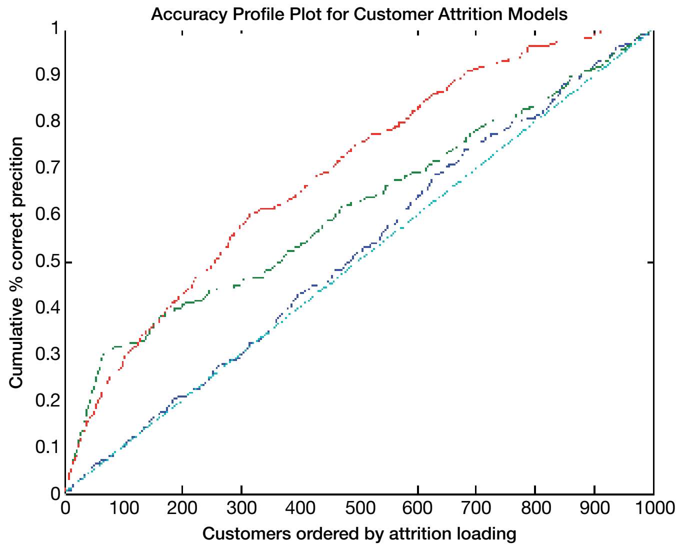
Logistic regression model is created to predict the probability of occurrence of an event (like attrition) by fitting data into a logistic curve. It uses many predictor variables that may be numerical or categorical. In case of customer attrition, variables may reflect the amount of fees paid by the customer in recent months or change in status of the customer.

Decision trees use tree-like graph or model of decisions to determine the conditional probability of an outcome (like attrition). It also uses numerical and categorical variables similar to logistic regression.

Since there are many possible predictive models to choose from, it is useful to have metrics to compare models and select the best one. Some commonly used metrics are Receiver Operating Characteristics (ROC) curve, Cumulative Gains Chart and Lift Chart. All of these provide metrics by trading off desirable outcomes (i.e. correct predictions) against undesirable outcomes (false positives or false negatives). These metrics are obtained by running the model on the training data set (used to create the model) or on an out-of-sample validation set.

ROC Curve plots True Positives along the y-axis and False Positives along the x-axis. Visually, the higher the curve is above the 45 degree line, and the closer it is to the top left corner, the better the model.

Figure 10: Sample ROC Curves

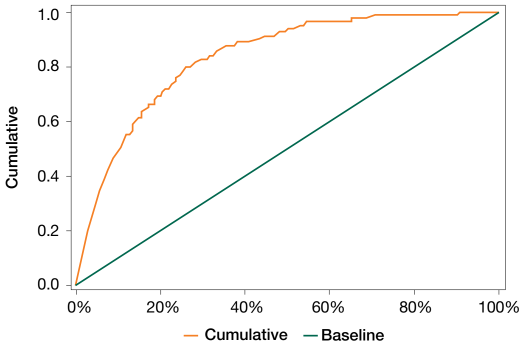


Source: Capgemini

Cumulative Gains Chart and Lift Charts are more commonly used by marketing departments as they allow for direct visual comparison and interpretation of results with respect to marketing campaigns.

Cumulative Gains Chart shows the cumulative percentage of target segment captured (on y-axis) by increasing the number of campaign audience (on x-axis). This curve typically shows that beyond a certain percentage, additional expansion of marketing produces little additional benefit in terms of target capture.

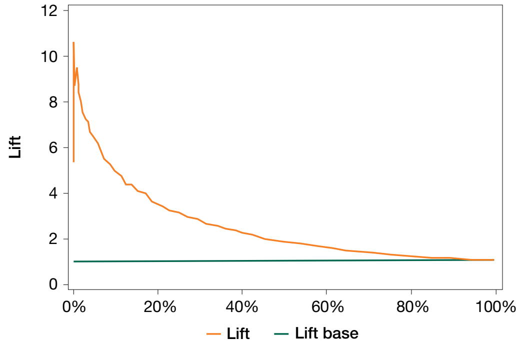
Figure 11: Sample Cumulative Gains Chart



Source: Capgemini

Lift chart directly shows the gain of using the model versus no-model approach. For example, in the figure above the model works 10 times better when a small percentage of audience is selected. The effectiveness decreases as the audience widens.

Figure 12: Sample Lift Chart



Source: Capgemini

7 Conclusion

Predictive modeling offers the potential for firms to be proactive rather than reactive. Predictive modeling using transactional data poses particular challenges which need to be carefully addressed to create useful models. Some of the key issues identified in this paper are data quality, cohort and trend analysis, model variable definition and model selection.

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

www.capgemini.com/financialservices

About Capgemini and the Collaborative Business Experience

Capgemini, one of the world's foremost providers of Consulting, Technology and Outsourcing services, has a unique way of working with its clients, called the Collaborative Business Experience.

Backed by over three decades of industry and service experience, the Collaborative Business Experience™ is designed to help our clients achieve better, faster, more sustainable results through seamless access to our network of world-leading technology partners and collaboration-focused methods and tools. Capgemini utilizes a global delivery model called Rightshore® which aims to offer the right resources in the right location at competitive cost, helping businesses thrive through the power of collaboration.

Capgemini reported 2009 global revenues of EUR 8.4 billion and employs over 90,000 people worldwide.

More information about our services, offices and research is available at **www.capgemini.com.**