Advanced Risk Management – Use of Predictive Modeling in Underwriting and Pricing

By Saikat Maitra
& Debashish Banerjee

Abstract
In this paper, the authors describe data mining and predictive modeling techniques as tools for advanced risk management. An introduction of data mining and predictive modeling is provided and certain terminologies are introduced. We then describe the data mining and predictive modeling process in detail using a case study in which simulated data was used to portray the Indian Personal Auto sector. Authors hope to demonstrate the great value that segmentation and scoring models might add to insurance business in the new de-tariffed Indian non life insurance market. Though the paper mainly target non life insurance sector, we believe the same techniques can be utilized effectively in health and life insurance sectors as well.

I. Background:

I.1 Data Mining and Predictive Modeling

Data Mining is a process that utilizes a number of modern and sophisticated techniques with the present day computation powers to analyze large quantities of related internal and external data to unlock previously unknown and meaningful business relationships. In short, data mining is about "knowledge discovery' and "finding new intelligence" to assist organizations in winning in the market place. For example, the exploratory analysis during data mining in discovering the strength of a wide range of particular variables on one-by-one basis is an example of data mining in practice.

Predictive Modeling techniques, on the other hand, can then be utilized to develop mathematical models that will bring a series of predictive variables together to effectively predict future events or behaviors, e.g. segment future insurance policies based on expected profitability.

Data mining and predictive modeling are part of an integrated process of **learning** from past data and attempt to **predict** the future. These terms are somewhat loosely used in the industry to indicate diverse activities.

I. 2 Supervised Vs Unsupervised Learning

Unsupervised Learning: In unsupervised learning, we are interested in fitting a model to a set of observations without knowing the true values of target variables. The purpose is to find any observable pattern in the data among the variables/characteristics. An example of unsupervised is using cluster analysis to group data with similar characteristics together, such as "young, professional accounts" vs. "mature, family accounts".

Supervised Learning: In supervised learning, we are interested to arrive at models which can reasonably explain a set of observations/target variables (called OUTPUTS) from another set of observations (called INPUTS). An example of supervised learning is to build regression models using loss ratio as the target against a series of policy, driver, vehicle, etc variables.

In insurance, we generally work within the framework of supervised learning as the variables which determine the strategic objective (e.g. underwriting profitability is determined by Loss Ratio) is well known. Few applications in insurance can apply unsupervised learning, such as insurance fraud detection.

I. 3 Insurance Risk Manage Challenges for De-tarrification in India Insurance Market

Risk Management is a procedure to minimize the adverse effect of a possible financial loss by (1) identifying potential sources of loss; (2) measuring the financial consequences of a loss occurring; and (3) using controls to minimize actual losses or their financial consequences. In the past, the risk management techniques never attempted to use the advanced statistical methods like predictive modeling, but relied more on rudimentary process or knowledge based approach.

For example, for insurance, it is critical to establish a more accurate way to separate out "good risks" from "bad risks". In India, some insurance companies maybe attempting the separation at a portfolio level. But, can this be done at a policy level? If so, how can we separate the good or the bad policies based on some measures (Y) and their characteristics (X)?

Most of the matured insurance markets around the globe are attempting to do more precise segmentation of good vs. bad risks, so that they can price and underwrite their books more accurately in the highly competitive markets. The authors feel the same need in India with the current de-tariffed environment: competing on price to increase market share and future sustainability is one the main focus for the management teams for most of the private non life insurance companies here. Providing effective tools to segregate the good policies from bad and understanding pricing gaps should be invaluable at this juncture for the Indian insurance market.

In US, predictive modeling started in the personal lines first because it involves homogenous exposure base and easy to define the coverage. There were some doubts whether the predictive modeling could be applied to more complex commercial business. However, today we are seeing that the results are widely applied for both the personal and commercial lines and have brought significant values to insurance companies in U.S. who embraced the technique early, that is, the first-mover advantage. There is no doubt that a predictive model can immensely boost this effort by identifying the pricing gaps and segregating policies based on risk regardless of the type of business.

For us, an important question is, "would it be of any use in India?" It is natural that there will be doubts among India insurers about the use of predictive modeling techniques in any of the lines. With limited experience among private carriers and only small amount of policy information being captured, it is indeed necessary to evaluate the efficacy in investing on building predictive models to enhance pricing, marketing or underwriting process of the company.

In this paper, we have made an attempt to answer this question using a case study analysis on a realistic Indian scenario. We will try to demonstrate that even with limited data and a small set of variables; it is possible to significantly improve pricing and underwriting process with data mining and predictive modeling.

II. Case Study

II.1 Strategic Objective for a Predictive Modeling Project

Some serious thought has to be given on what is the problem at hand before conducting predictive modeling. We need to tackle one problem /issue at a time:

- Is it pricing or underwriting that we want to improve through modeling?
- Should it be done at policy level or other level, such as vehicle level or portfolio level?
- Should the study be focused primarily on loss or both loss and expense?

A complete understanding of the problem should give the business case for the analysis and the associated model design that we intend to solve. Answering the question will assist in determining the choice of the dependant variable (Y).

For the predictive variables of X's, we need to analyze them from a business perspective as well. One of the drawbacks of statistical methods is that sometimes the relationship between the Y and X's are difficult to explain and there might be spurious correlation. While working on the predictive models, it is very important to analyze each and every predictive variable and review the finding that we observe not only from statistical criteria, but also from business perspective. That is, can we can explain the relationship between the target and the predictive variables or not.

For example, for the selection of the dependent variable for modeling, it could be severity, frequency, or loss ratio, or maybe any flavor of the same, such as severity or loss ratio capped at 95th percentile for elimination of large loss impact. Also, for premium used to calculate loss ratio, we have a choice of historical actual premium or premium adjusted to the current level. For predictive variables, we can consider whether the variables should be completely from policy information, or we can enrich the list with other sources, whether from company internal sources or sources external to the company. These measures have to be really thought through, and the answers should be based on what the objective is for the modeling.

II.2 General Data Mining and Predictive Modeling Process

We could divide the entire data mining and predictive modeling process into the following phases:

1. Data Load: load the raw data, such as premium, loss, and policy data into the system.
2. Variable Creation: from the raw data, create both target variables and predictive variables.
3. Data Profile Analysis: create all the necessary statistics, such as mean, min, max, standard deviation, miss values for all the variables created above, and then analyze the correlation between the predictive variables and the target variables, one at a time.
4. Model Building: build multivariate models
5. Validation: validate whether the multivariate models' performance using independent data set.
6. Implementation: perform business and system implementation of models.

The above steps provide a generic framework in which any data mining and predictive modeling exercise may be carried out for an insurance company. Some of the details involved in carrying out each of these steps will be discussed through our case-study.

II.3 Case Study

We will now describe in details the above phases of data mining and modeling via a case study.

Data Used for the Study:

The data for the case study was simulated to closely resemble Indian personal auto insurance sector. The simulation involved generating information relevant for Indian market, i.e. the ones which are typically used for the current tariff structure.

The following variables were simulated at policy level. We assumed single vehicle policies only, and we will study the physical damage coverage:

- Incurred Loss (Physical Damage): This was aggregate loss over a policy year. We did not split loss into individual claims and loss per claim.
- Sum Insured Or Insured declared value of the vehicle
- Vehicle Cubic Capacity
- Vehicle Seating Capacity
- Financial Year
- Branch Code
- Country Zone of Policy
- Maker of Vehicle: Maker variable in our data was a categorical variable indicating whether the vehicle was manufactured by Indian, US, Asian or European Manufacturer.
- Segment of Vehicle: Segment was a categorical variable indicating whether the vehicle was of Small, Mid-Size, Premium, Luxury OR Utility Segment.

We chose to do a case study with an Indian insurance sector in mind to demonstrate the efficacy of the data mining and predictive modeling methodology in Indian market where relatively few policy information are captured.

Data collection in our market is driven typically by rating, underwriting and marketing requirements. The tariff plan for personal auto has very few factors involved and we believe that insurance companies have typically not collected any data beyond what was required. For example the tariff plan does not involve any driver information (age, sex) to calculate physical damage rates. So, we have not simulated any data beyond the minimal that we expect to be available in the Indian Market.

Business Value Review: Underwriting Scoring/Segmentation Vs Pricing

Prior to starting the actual data mining process, a business case analysis is done to understand the company's processing and data environments to develop specific data mining recommendations. Both opportunities and project risks should be identified at this stage.

Analysis of losses against the policy characteristics that has been captured can yield vital clues as to what type of policies should the company focus its underwriting and marketing resources so as to grow and be profitable. This is achieved by building a scoring or segmentation model which captures a policy's true level of risk. Such models typically are not limited to actuarial applications, but should be, for maximum benefit, integrated into underwriting and marketing applications as well.

The same analysis as above with minor modifications can be used to build a pricing model to calculate the rating factors. Such a rating plan will reflect the companies own experience and would be much more accurate than a tariff plan. It can automatically lead to identify pricing and profitability gaps if properly implemented.

For more developed markets around the world, generally underwriting scoring models involve a lot of additional variables (like Billing, Agent, Demographic data, etc) which typically are not part of the rating plan. These models utilize the underwriting flexibility to apply "credit" or "debit" over the price indicated by the rating plan.

The authors would also like to note that underwriting models are built and applied at policy level, where pricing models are built at risk, exposure, and coverage level. For personal auto, policy level and risk level analysis are the same. However, for commercial lines, difference will exist, so pricing and underwriting models should be built separately.

In the remainder of the paper we will describe the data mining and predictive modeling process and the generic setting assuming that we are building either a scoring or a pricing model. At the end, in the concluding sections we would suggest what improvements and changes in the approach may be used for building a pure pricing model or a pure scoring model.

Phase 1: Data Load

This phase involves actually 3 steps:-

- Data Specification – Specifying in detail what fields and what level of data is required. This allows the MIS department to program the data extraction. A data dictionary, data record layout, and detailed programming specs are sought from MIS at this point.
- Data Extraction – Actual extraction of data by MIS
- Data Load – Loading the extracted data in the statistical analysis platform.

The statistical analysis platform could be chosen based on: (i). flexibility in loading data of different formats; (ii). Robust for handling huge volume of data; (iii). Availability of the necessary tools / methodology for statistical modeling. We believe that using off-the shelf insurance specific modeling software's will make both modeling and application restrictive. Most of these off-the shelf software admit data in a specific format and have limited capability when compared to the statistical software available in the market.

Phase 2: Variable Creation

Raw data obtained after data load mostly likely cannot be directly usable for data analysis. At this stage necessary transformations to the data and variables are made to arrive at the predictive variables and the target variables.

Data Transformations

Appropriate data transformations are done at this step and various predictive variables are created. For example transactional level data is rolled up to policy level to create premium and loss.

Data may also come from multiple systems and tables. At this stage various different data sets have to be merged to arrive at a common modeling data set. This dataset has to be at the level at which the model will be built. For example, for underwriting scoring all data must be brought to policy level and merged by a well-defined policy key.

The simulated data used in our study was already at policy level and no further transformation was needed.

Predictive Variable Transformations

The raw data fields in systems often need great amount of modifications to create the predictive variables.

For example "vehicle age" variable can be created from "year of build" (which most company would be capturing during their underwriting process) variable at this stage using the formula VEH_AGE=POLICY_YEAR-YEAR_OF_BUILT. Another example could be that we may use historical variables for renewal policies like "last year's loss ratio / LR_PREV_YEAR". This can be calculated by looking the previous year's records for the same policy. We did not compute any historical variable for this study. The authors, however, believe that these are very predictive in nature.

Target Variable Transformations

We chose "INCURRED LOSS (Physical Damage) / SUM_INSURED" as our main target variable. We call this RATE in the remainder of the paper. The goal of the remainder of this case study was to arrive at a predictive segmentation based on the rate.

The advantages of using RATE as defined above for the target variable instead of raw loss are:

- We are normalizing the loss by the exposure base, hence removing the bias due to Sum Insured, if any.
- The modeled RATE can be directly used in the pricing plan
- No additional work is required to adjust for inflationary as RATE is already adjusted by a exposure unit (SUM INSURED)

However this is just one choice, and either LOSS or LOSS RATIO can also be used for the same purpose. Please refer to the earlier section of "Strategic Objective" in this paper for more details.

Actuarial Adjustments

While preparing the target variable, actuarial adjustments play a key role and cannot be ignored. Target variables across the dataset should be at same level (ultimate figures). This means that losses should be trended and developed, and premium (in case of LR as a target variable) must be adjusted for all the prior rate changes.

In our case, we chose RATE as the target variable, so this requires little actuarial adjustments as it is safe to assume that both LOSS and SUM INSURED are affected by similar inflation effect. However loss still needs to be developed and it may be prudent to use actuarial estimated loss development factors rather than use the claims department's estimate of outstanding loss.

Treatment of outliers and missing values

After creating the target and predictive variables, one should look into the distribution of each of the variables. Two main issues to resolve at this point are:

1) Treatment of outliers
2) Treatment of missing values

Various methods like exponential smoothing or capping the value of variable at the 99th or 95th percentile can be used to treat outliers.

If number of data points with missing value is very low or non existent (as with our simulated data), we may ignore them. Typically, statistical packages ignore the entire observation with missing values (in any variable) while fitting models.

In case missing values are significant for a variable, a suitable method can be chosen to impute the missing values. Typical methods of involves imputation include:

a) Using mean/media/mode value – this is a very crude approach and depends on the distribution of the non-missing values and business reasons
b) Estimating the value using regression with other correlated variables
c) Binning with missing as a separate category – i.e. creating indicator variables

We used variable binning method to bin all our predictive variables into disjoint groups of uniform size. This was done by analyzing the distributional characteristics of each of the variables. For example the variable CUBIC CAPACITY (a continuous variable) was binned as follows:
Values 0 To 799 in Bin 1
Values 800 To 999 in Bin 2
Values 1000 To 1299 in Bin 3
Values 1300 To 1599 in Bin 4
Values 1600 To 1799 in Bin 5
Values 1800 To 1999 in Bin 6
Values 2000 To 2199 in Bin 7
Values 2200 To 2399 in Bin 8
Values 2400 To 2599 in Bin 9
Values 2600 To MAXIMUM in Bin 10
Post which the distribution of non-missing values where studied and the missing observation was imputed by the median of that distribution.

Phase 3: Data Profiling

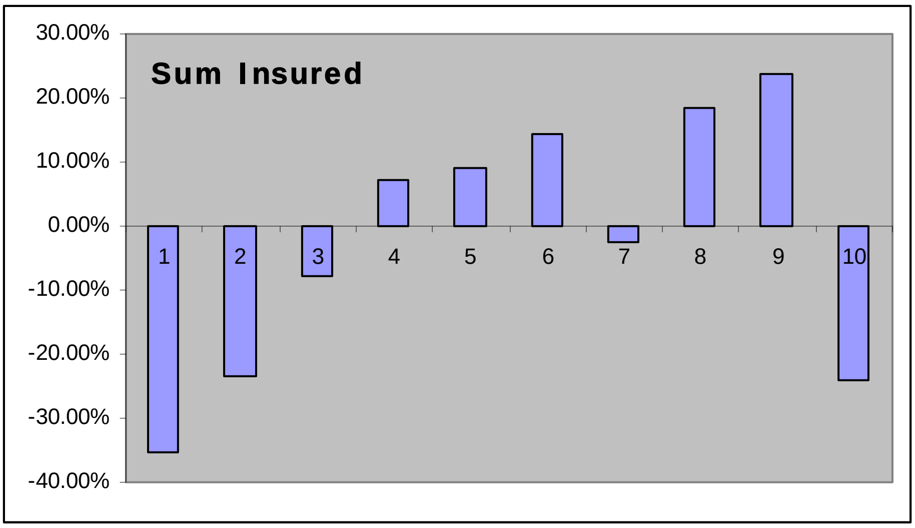
Before moving forward with multivariate modeling, it is advisable to proceed by analyzing each predictive variable vis-à-vis the target variable (RATE in our study). When number of predictive variables are large, data profiling helps us to discard variables which have little predictive power and keep the "strong" variables during actual modeling.

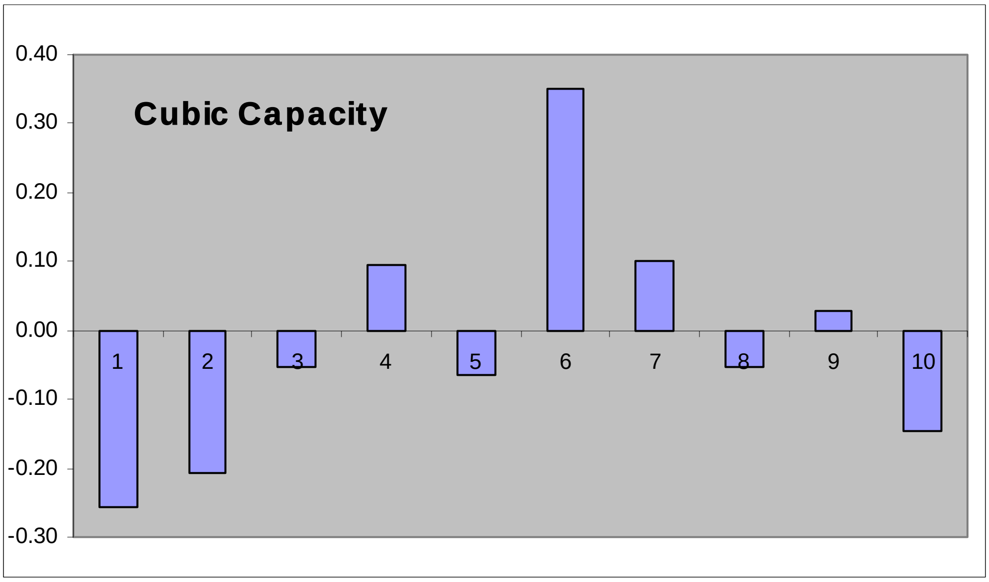
In our study we started with 8 predictive variables and after variable transformation & binning, all of them were analyzed using the following lift charts.

Lift Charts

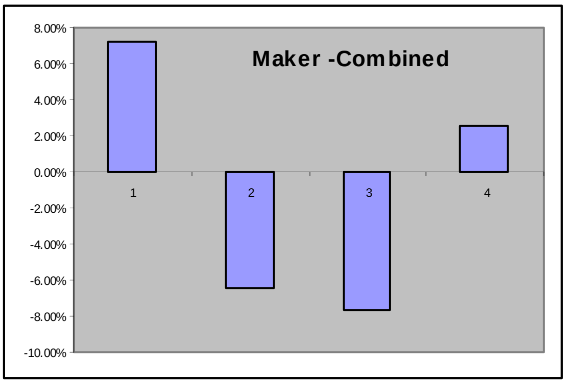
We plot the difference of the value of RATE with the average value of RATE against each value of the predictive variable. This chart shows the relative improvement in the value of RATE with changing values of the predictive variable. We not only plotted lift charts for the entire data but also for each financial year separately to study any trend existing in the variable.

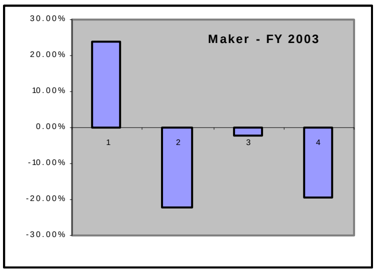
The following lift charts show that the variable, SUM_INSURED, showed a good lift as did CUBIC_CAPACITY.

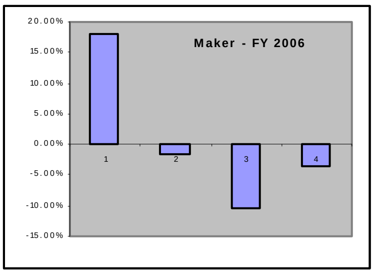




The variable MAKER showed minimum lift, however was not entirely non-predictive. We can see the BIN 1 of this variable was significantly worse than the rest. BIN 1 was all vehicles which were manufactured by Indian Makers.







In summary, we found that all variables in the study had significant effect on the target variable, so none of them should be dropped from the multivariate modeling. This is expected since all variables are the core rating variables.

Phase 4: Modeling and Validation

All 8 variables were considered for multivariate modeling as none of the variables were very weak.

Training Vs Validation Datasets

Before commencing modeling, data must be divided into training data and validation data. Typically, for insurance application validation data should be latest couple of years of data to provide an independent validation of the model results. The true indication of the power of a model to predict the future can be done by building models on an older training data set and selecting the based model based on performance in the latest validation dataset.

In our study we used data for 2007 as the validation dataset, and the rest of the previous years' data was used for training /model building.

Correlated Predictive variables

Presence of highly correlated predictive variables increases the variance of the parameter estimates for regression (either OLS or GLM). As a result the individual factor estimates from the model becomes unreliable for drawing any inference.

Various methods may be used to solve this correlation problem including:-

a) Select a smaller subset of less correlated variables based on business understanding
b) Use the data profiling results to drive the selection of more important variables from the set of correlated variables
c) Use stepwise regression to select the more important variable
d) Multivariate techniques like principal components analysis and partial least squares.

The above methods can be combined with actuarial judgment to eliminate the correlation effect.

Again, with limited number of variables in our case study the correlation problem was proven to be somewhat insignificant. Pair-wise correlation analysis was done and the result showed that CUBIC_CAPACITY, SEGMENT and SUM_INSURED were somewhat highly correlated.

CUBIC_CAPACITY	SEGMENT	0.842484
CUBIC_CAPACITY	SUM_INSURED	0.704662

Please note that both of CUBIC_CAPACITY and SUM_INSURED showed very strong lifts in data profile analysis. SEGMENT also had a reasonable lift.

Similarly, BranchCode and Zone were found to be reasonably correlated, 0.72, as could be expected. From business perspective, branch code can be dropped since pricing zone is used in the tariff plan.

Model Options: OLS Vs GLM

Due to presence of discrete predictive and non-normality of the target variable, GLM is the first choice amongst modelers for such case.

However GLM framework allows OLS as a special case (Normal Distribution, Identity Link) we can easily try out OLS while doing GLM, rather than rejecting OLS outright. The method of binning allows us to use the same variable as continuous and discrete.

Classically choice of the distribution in GLM may be indicated by looking at the distribution of the target variable (RATE). Choice of the link in GLM can be rigorously done by fitting box-cox transformations.

However, in reality, the easiest and most reliable method is to fit many models using a variety of combination of distribution, link functions and variables. This is an iterative process which is very much dependent on the experience and business knowledge of the modeler.

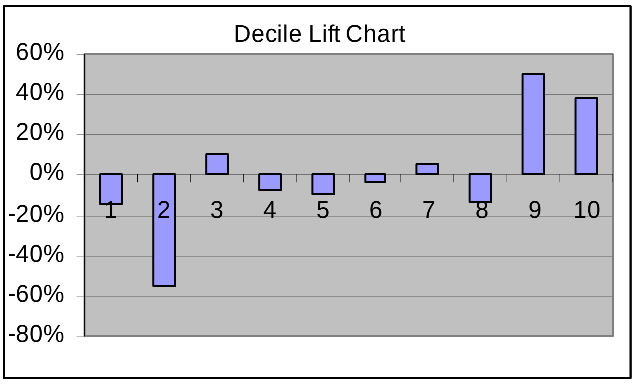
Significance and strength of the predictive power of a variable are provided by the statistical software using P-Values and Deviance based Type I statistics. These statistics should be used to discard weak predictors from the model.

The performance of the model as a whole can be tested using Type III tests which show the improvement in deviance when fitting a set of nested models. However, we want to stress the point that the only reliable test of a model as whole is its performance against the validation dataset. Statistical measures based on deviance which are calculated by the software in the training dataset should not be relied upon since the result overstates the performance of the model.

After fitting a model in the training dataset we plot the lift chart for the validation dataset. A model is deemed better than another can be relied on the basis of lift chart on the independent validation dataset.

Modeling Results

We started of with all variables, normal distribution, identity link and taking all variables as continuous. The lift chart (in the validation dataset) for this first model is shown below.



Iteratively many models were then tried till we settled on a Gamma Distribution, Log Link for the final model. We used the first principal component of SUM_INSURED and CUBIC_CAPACITY instead of using the raw variables separately to avoid multi-collinearity. This principal component and SEATING_CAPACITY were used as continuous predictors.

BRANCH was dropped as a predictor from the final model mainly because it doesn't make sense for a pricing model. It may still be used for a Scoring model. ZONE, MAKER and SEGMENT were treated as Categorical predictors.

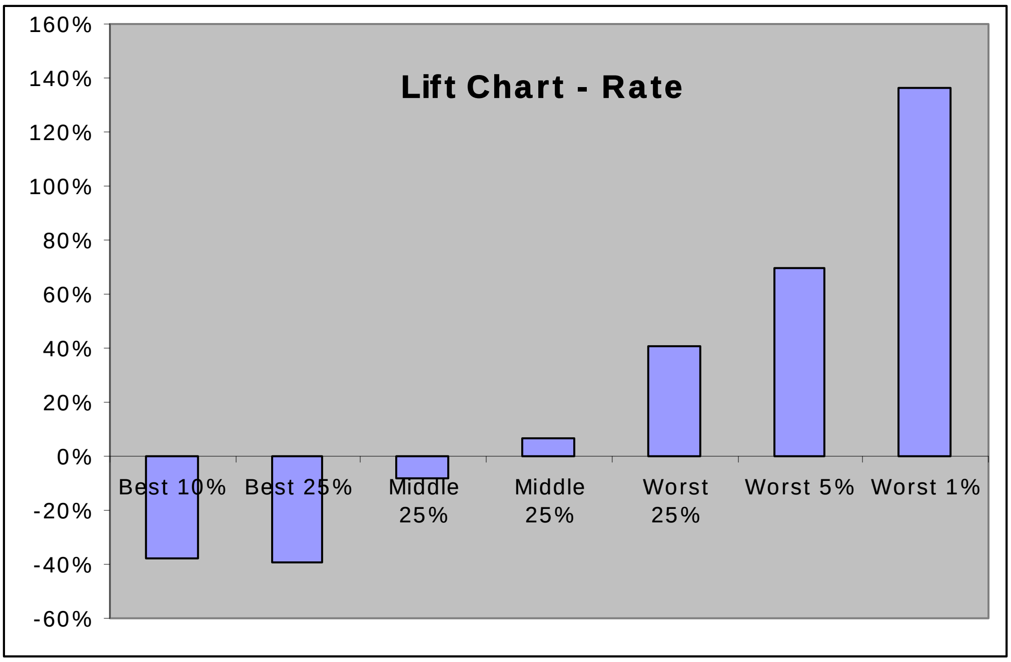
Financial Year (FINYEAR) was used as an offset variable. This is to remove any bias which may be creeping in from year to year.

The significance of the predictors is shown below which indicates all variables included were strong.

Likelihood Ratio Statistics for Type 1 Analysis

			Chi-	2*Log
Source	Likelihood	DF	Square	Pr > ChiSq
Intercept	1698669.61			
MAKER	1698685.69	1	16.08	<.0001
SEGMENT	1699971.34	4	1285.65	<.0001
ZONE	1705978.67	3	6007.33	<.0001
FINYEAR	1705978.67	1	0.01	0.9401
SEAT_CAP	1706085.25	1	106.58	<.0001
pc_size1	1706155.47	1	70.22	<.0001

The lift chart of the final model on validation dataset is shown below.



Using the Model

This model can be used for both pricing and underwriting. While for underwriting we can use the linear predictor (maybe LR) as the raw scoring formula (for segmentation), for pricing we need to take exponential of the linear predictor and arrive at a multiplicative formula to set the base price (pure premium). Further loading of base price needs to be done for expenses, profit margin and contingency provisions.

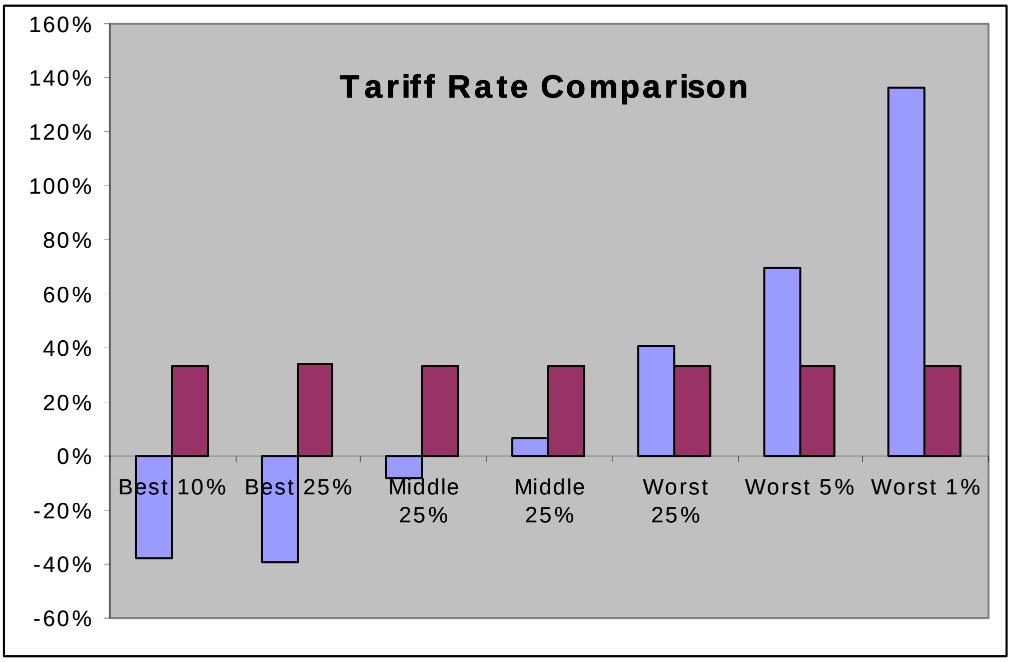
For scoring, we will use the following formula to get the raw scores (this is the final equation of our model):

```
SCORE = 0
 + (-1.63760717 )
 + CUBIC_CAPACITY  * (-0.002514933 )
 + MAKER * (-0.011415772 )
 + SEATING_CAPACITY  * (0.0627351222)
 + SEGMENT_1  * (-0.048671275 )
 + SEGMENT_2  * (-0.022205737 )
 + SEGMENT_3  * (-0.068712889 )
 + SEGMENT_4  * (-0.094753683 )
 + SUM_INSURED  * (-0.002573695 )
 + ZONE_1  * (0.1677551306 )
 + ZONE_2  * (0.0843169933 )
 + ZONE_3  * (0.011902719 )
```

Further in scoring application, additional tasks of creating decile or centile cuts and derive appropriate business rules need to be worked out. However we cease from going into details of a rules engine in this paper.

Conclusion of the Model

To evaluate the impact of the model – we plotted the Rate obtained using the Tariff chart (for each segment) & the lift curve provided by the model. The result is shown below.



The tariff rate hovers around 3% of insurance value as a look into the tariff chart will suggest. Even though this may be adequate over the portfolio, it does a poor job of segmentation the policies based on risk. The tariff plan does use vehicle age, cubic capacity etc. but in a simplistic way (e.g. all vehicles less than 5 years age and are less than 1000cc are charged a flat price).

The above chart indicates that proper modeling even with the basic rating factors available can indicate areas where there is a significant gap between actual cost and tariff price. These gaps can be utilized by companies and the power of the data collected through predictive modeling is clearly indicated.

Phase 5: Implementation

This phase entails on how the developed model is used. A model built just to arrive at a pricing formula which may be then used to finalize the actuarial rating plan needs no special implementation. Same may be true for a scoring model to understand market dynamics to be used for marketing strategies.

A scoring model with associated rules engine may be, on the other hand, implemented and integrated within the underwriting system. In Indian market, we are not in a position to comment if it would be possible to integrate a predictive model in underwriting or policy systems of the company. This would vary from company to company and the system / platform they have.

III. Conclusion

We have now seen that using a simulated (albeit realistic) personal auto data that even using the minimum core variables suggested by the tariff structure has significant price gaps over the tariff/current rate which may be discovered by using multivariate predictive modeling.

With real industry data, we may have significant opportunities to improve further the model that is being suggested in this paper. Insurers may be capturing additional data (e.g. vehicle age) which can be included and tested for modeling. Subject to regulatory requirement, past performance of a policy may be used as a predictive variable as well, and such information is likely to lead to improvement in model performance for policies with prior experience.

Finally, to capitalize the possible maximum benefit associated with predictive modeling for underwriting risk management and pricing, insurers are encouraged to collect additional policy information's. For personal auto considered in this paper, there was no policyholder (driver, vehicle owner) information that was used. International experience suggests that such policyholder information are generally have very strong predictive power.

To end this paper we would like to reiterate that methodology presented here a generic framework of building predictive models without any special attention as to the application. The methodology can be further improved based on whether we are doing a pricing exercise or building a scoring application, Following are a few suggestions which should be considered during actual application.

Improvements for Pricing

In pricing we are interested in actual point estimate of the predicted value and how close it is to the real value (not just segmentation). Following methods are likely to produce better point estimates of loss cost (or RATE in our case).

Tweedie Distribution: While we used a gamma distribution as our final model, tweedie distribution provides the theoretically correct model for taking into account the large percent of exact zeroes in the loss variable, Tweedie family of distributions is a sub-class of the exponential family with variance function given by *(mu)^p* where *mu* denotes the mean and *p* belongs to interval (0,1). If p=1 tweedie distribution reduces to Poisson distribution (frequency modeling). For p=2 tweedie reduces to a gamma distribution (severity modeling). For intermediate values a compound Poisson distribution can be easily modeled. Choice of p can be done based on analysis of residuals.

Frequency-Severity Approach: The classical actuarial approach in pricing is fitting separate distributions to model frequency of claims per policy and severity of a claim. Same approach may be used for predictive modeling where separate predictive models are built to model the frequency and severity components. GLM approach easily lends itself to the task of modeling count data (frequency) using Poisson models.

Improvements for Scoring

Inclusion of historical variables (past performance) in modeling as previously indicated and building separate models for New vs. Renew business can provide major improvements.
Instead of building independent and separate models for New business and Renew business respectively another approach is to build a common base model and test additional variables for Renew business over the base rating factors.

Let RATE denotes the variable containing observed values of RATE. Let RATE_NEW denote the fitted value of RATE using the base model. Define E=RATE-RATE_BASE. We may now test the additional variables taking E as our dependent or target variable, which will then provide refinements over the base model.

IV. References

1. Wu, C. P., Guszcza, J., “Does Credit Score Really Explain Insurance Losses? - Multivariate Analysis from a Data Mining Point of View,” *2003 CAS Winter Forum*, Casualty Actuarial Society (2003).

2. Mildenhall, S. J., “A Systematic Relationship Between Minimum Bias and Generalized Linear Models,” *Proceedings of Casualty Actuarial Society,* Vol. LXXXVI, Casualty Actuarial Society, (1999).

3. Feldblum, S. and Brosius, J. E., “The Minimum Bias Procedure--A Practitioner's Guide”, *Proceedings of Casualty Actuarial Society,* Vol. XC, Casualty Actuarial Society, (2003).

4. Neter, J. , Wasserman, W., Kutner, M. H., “Applied Linear Regression Models“, (2nd Edition), Richard D. Irwin, Inc., (1989)

5. Kass, Rob, “Compound Poisson distribution and GLM's - tweedie's distribution“

6. P. McCullagh and J.A Nelder, "Generalized Linear Models" - Pub: Chapman and Hall

About the Authors:

Saikat Maitra

Saikat Maitra works as an Assistant Manger in Deloitte Consulting's Advanced Quantitative Services team, at the firm's India office in Hyderabad. He has over 5 years of experience working in the actuarial profession in non life reinsurance pricing and predictive modeling field. With Deloitte he has primarily worked in building and implementing underwriting scoring engines for several of Deloitte's US clients. Prior to Deloitte he was with Swiss Re (Genpact) and has worked on development of re-insurance pricing models, pricing of reinsurance deals, catastrophic modeling and actuarial rate adequacy studies. Mr. Saikat is a student member of Casualty Actuarial Society, US and has cleared the preliminary exams of the society. His professional interests are in the fields of predictive modeling and model evaluation. He has bachelors and masters degrees in statistics from Indian Statistical Institute, Kolkata.

Debashish Banerjee

Debashish Banerjee has over 7 years of experience in non-life insurance and re-insurance analytics. Most of his work involves in the data mining and predictive modeling space. He started his career with GE Insurance and was instrumental in establishing and leading the non-life reinsurance pricing team for GE Insurance in India. He was awarded the most prestigious "Summit Award" by GE Insurance. He has the expertise in reinsurance pricing, statistical & parameter studies, building pricing models, exposure curves & market studies, and predictive modeling. He moved to Deloitte in 2005 with the primary goal to set up the Advanced Quantitative Solutions practice in India. His consulting experience is mainly focused on the commercial lines. His clients include both insurance companies and self-insured entities. He is instrumental in creating some of the best practices within the actuarial group in Deloitte. He is currently the Service Line Lead for the Actuarial group of Deloitte Consulting LLP, Hyderabad, India. Mr. Debashish is a student member of Casualty Actuarial Society, USA. He has bachelors and masters degrees in statistics from Indian Statistical Institute, Kolkata.