B-64

Kahane, Yehuda, Nissan Levin, Ronen Meiri, and Jacob Zahavi (2007). "Applying Data Mining Technology for Insurance Rate Making: An Example of Automobile Insurance," *Asia-Pacific Journal of Risk and Insurance*, Volume 2, Issue 1: 33-50.

Applying Data Mining Technology for Insurance Rate Making: An Example of Automobile Insurance

Yehuda Kahane[+]
Nissan Levin[+]
Ronen Meiri[+]
Jacon Zahavi[+]

Abstract

In this paper we discuss the use of modern data mining (DM) methods to design risk-based insurance premiums for motor vehicles. Our objective is to predict the likelihood and expected value of future claims for each insured based on a myriad of attributes available in the database on the customers and their "peers." The model results may then be used for underwriting and for rate making. We employ a two-stage approach, involving a survival analysis model and a linear regression model, to estimate the risk level of each customer and the proneness to file a claim. The study was performed on actual data set obtained from a small insurance company. We demonstrate our ability to discover new underwriting parameters, build accurate predictive models and to distinguish between distinct groups of policies. The new method creates a new ordering of the policies where the most risky people were, on the average, 12 times more expensive than the least risky people. The importance of the study is not in the particular results, which are specific for the particular company and its environment, but rather in the demonstration of the general ability to use data mining for insurance rate making purposes, and in the original use of the concept of survival analysis and the concept of mean time between claims for this purpose.

Key Words: Automobile insurance, Data mining, Insurance pricing, Insurance rate making, Insurance underwriting decisions, Motor car insurance, Predictive models, Survival analysis.

I. Introduction

Insurance companies create and maintain large databases, which are often utilized only partially for deriving useful underwriting and actuarial knowledge. Modern Data Mining (DM) technologies may now be used with these vast data sources, in order to automate and improve data analysis and create useful predictive models. The insurance industry is a ripe area for DM. Not only because it is a data-rich industry, but also because of the recognized huge potential of DM to support a variety of decisions problems, ranging from targeting decisions for cross and up selling, fraud detection, customer acquisition, pricing and actuarial decisions, and others. The purpose of this paper is to demonstrate the use of DM in estimating the desirable net premium for motorcar insurance. We demonstrate the ability

---

[+] Yehuda Kahane (kahane@post.tau.ac.il) is Professor, Faculty of Management, Tel Aviv University, Israel. Nissan Levin is with Q-Ware Consulting, Israel. Ronen Meiri teaches at the Faculty of Management, Tel Aviv University, Israel. Jacon Zahavi is Professor Emeritus, Faculty of Management, Tel Aviv University, Israel.

to discover new underwriting parameters, to build accurate predictive models, and to distinguish between distinct groups of policies. The particular example discussed in the paper yields a very different rate structure than the one currently used by the investigated insurance company. The new method creates a new ordering of the policies where the most risky decile is on the average 12 times more expensive than the lower decile. The importance of the study is not in the particular results, which are specific for the particular company and its environment, but rather in the demonstration of the general ability to use data mining for insurance rate making purposes, and the novel concepts used in the study.

II. Data Mining and Insurance

Statistical methods and models have been, for years, the primary tools for data analysis. Typically, statistical methods begin by making assumptions (hypotheses) about the population and then testing their validity using data analysis (for example that the probability of an accident depends on the age of the driver). But in large databases the number of possible hypotheses is huge, rendering this process very long and tedious, practically prohibitive. Data mining (DM) is a new generation of computerized methods for extracting useful information from massive amounts of data. Often, DM does not require any prior hypotheses. Instead, the DM engine interrogates the database for feasible hypotheses and then test their validity and significance.

Take predictive models, for example, say, predicting the probability of next purchase for each customer in the database. Predictive models often involve hundreds of potential predictors, only a handful of which usually suffice to explain the phenomenon at hand. Choosing the most influential predictors is a very tough combinatorial problem known as the feature selection, or the specification, problem.

The beauty of DM is that this process may be automated to a large degree. The DM process investigates a large variety of candidate variables to introduce to the model, checks the recommended grouping of variables, the type of linear and non-linear relationships that exist in the data and the variables to include/exclude from the model. Only the significant relationships make it to the final model. A validation test is essential, to guarantee against over-fitting and ensure the model results are stable and reliable. Of course, best results are obtained by introducing domain knowledge into the process, e.g., specific data transformations that can "explain" the phenomenon under study.

The process of building and implementing a DM solution is referred to as Knowledge Discovery in Databases (KDD) (Fayyad et al., 1996). The process contains three main components: pre-processing, data mining (modeling and analysis), and post-processing.

Pre-processing. This is often the most time consuming of the KDD process. It comprises of several tasks to prepare the data for modeling, including defining the application problem, creating a target data set for the analysis, checking data integrity, cleansing the data, transforming and collapsing the data, and others. Each of the possible predictors (e.g., the driver's age) is being tested for its quality and the quantity of "knowledge" it contains. New predicting variables are generated from the original set using several transformations that may improve the amount of "knowledge."

Data mining (DM). It is about interrogating the data for patterns and trends for decision-making. DM is an interdisciplinary field involving query tools, regression models, association rules, decision trees, rule-based expert systems, neural nets, genetic algorithms, visualization, and others. Perhaps the most intriguing component of DM is the feature selection problem mentioned above. In some data mining models, the feature selection

process is inherent in the algorithm, such as in decision trees. But in most other models, feature selection is part of the model building process.

Post-processing. Evaluating and interpreting the modeling results to make sure the models are adequate. Most important is validating the model results to guard against over-fitting. Often, several candidate models are used to study the data, requiring that one analyze the performance of the models and choose the final model for implementation.

The KDD process is basically an interactive and iterative process for extracting knowledge from data, requiring the active involvement of the decision-maker and the domain experts. The users can intervene at any stage, by offering certain variables, certain relationships between variables, suggesting a mathematical formulation, etc.

DM tools and methods have already been applied in a variety of industries, including the automotive, banking, insurance, communication, book club, retail (catalog), travel, health care, and others. Typical applications include targeting of audiences for cross selling (e.g., offering a life insurance policy to a customer holding a car insurance policy) and up selling (e.g., increasing the life insurance coverage in return for a small increase in premium), acquiring new customers, loyalty and retention programs, churn management, fraud detection, credit scoring, and others.

In the insurance industry, DM can be used for prediction and diagnostics. In prediction, DM analyses past claims records, lifestyle characteristics, demographic data and risk information in order to predict the likelihood and expected value of a future event, say the future claim level, the propensity to purchase another insurance policy, and others. In diagnostics, DM is used to develop an understanding of customer characteristics and profile in order to improve marketing decisions. Several studies have already been done to harness the data mining technology to address some of these issues, to mention a few: Williams and Huang (1996), Viveros et al. (1997), Staudt et al. (1997), Kietz et al. (1997), Cheo Yeo et al. (1997) and several IBM Research Reports (Pednault et al.,(2000); Apte et al., (1999)).

In this paper we focus on the application of DM to design risk-based motor insurance premiums. Our objective is to predict the likelihood and expected value of future claims for each insured in terms of a myriad of attributes available in the database on the customer and his/her "peers." The modeling results may then be used for underwriting and for rate making. We employ a new approach, based on survival analysis, to estimate the customer risk level and the proneness to file claims. The study was performed on actual data of a small insurance company, which is unidentified here due to commercial reasons.

III. The Rate Making Problem

For years insurance companies have been trying to improve their rate-making scheme, to develop new methods, which will enable them to determine a reasonable risk-based rate structure, and to appropriately differentiate between the policyholders according to the likelihood of claims and the expected value of the claims.

The core of such a customer-specific differential rating scheme is predicting the expected claim level for each customer as a function of certain rating factors (e.g., in automobile insurance: the car characteristics, the claim history and the demographic attributes of the car owner and the other drivers). This is a complex prediction problem where the dependent variable is the (continuous) claim amount, and the explanatory variables (or predictors) are the car characteristics, the customer attributes and the claim history.

Traditional methods to set up insurance rates are based on segmentation. The objective has been to divide the population into "homogenous" segments (e.g., policyholders having homogeneous claim pattern). Then, the claim level for each customer is given by the average claim values of all policyholders belonging to the same segment. This method assumes that all individuals in the segment are "alike." The segments are created to yield groups that are not-too-small and not-too-big. Small segments suffer from lack of statistical significance and poor prediction. Large segments may not be homogenous enough for decision-making. Sometimes, segmentation methods were replaced by the analysis of rate relativities, obtained from the analysis of the margins of the tables (see critical analysis of these methods by Hart et al., 1996, and Hossak et al., 1983). Both segmentation and rate relativities methods may yield erroneous results because they usually do not consider interaction terms, which are very prevalent in the insurance industry.

In the 1970's, following the development of sufficiently strong computational power, multi-regression analysis was suggested as a possible remedy for these problems. One of the earlier studies in this direction is the study by Kahane and Levy, who suggested and used this approach in 1969-70 for the determination of motorcar insurance rates for Israeli insurers (a summary of part of this study has been published in Kahane and Levy, 1975). Such methods are currently used quite commonly by the insurance industry (and rate-making bureaus), alongside with traditional segmentation methods.

Regression models are individual-based models regarding each customer as a separate entity. Unlike segmentation methods, regression models use all the information available on customers to predict the desirable effect (claims, in our case). This is not just a technical difference, as it enables one to probe into smaller effects that would be lost in segmentation. For instance, assume that a given attribute (say gender) has a significant impact on the claim level. When considered in a context of a regression model, which involves the entire population, the gender effect may turn out significant. However, this same effect may be washed out in the segmentation analysis due to the small number of observations in each segment.

Regression models can also handle non-linear effects of attributes on the claim level, which segmentation methods cannot (unless we use a very refined segmentation which may result in segments which are too small to yield significant results). For example, it is possible that the probability of an accident may depend on the driver's age. This relationship may be non-linear, with higher probabilities for young and old drivers, and lower probabilities for middle age groups. Ignoring such non-linear relationship may result in a wrong model and, in turn, in poor prediction. In addition, regression may help in the analysis of co-relationships (or interactions) among variables. And with data mining technology, many such relations can be handled simultaneously. However, when we have a very large number of potential variables, and their effects are non-monotonic, regression analysis could become a very difficult task.

Yet, while relatively very common, linear regression suffers from several deficiencies making it somewhat inappropriate for modeling claims. First and foremost is the inherent assumption that the dependent variable is observed in all cases and may take any value – positive or negative. Clearly, this assumption does not hold in our case since the claim level can assume only non-negative values, either 0 if there was no claim, or a positive value if a claim was filed, in violation of the basic assumption above. Since the linear regression model is not bounded from below, some predicted claim levels may turn out negative, in contrast to the fact that the actual claim levels are nonnegative (either the customer files a claim or does not). To overcome these difficulties, we pursue a two-stage approach, after Heckman (1979), for modeling claim levels, using a combination of survival analysis and linear regression, as discussed below.

IV. The Mean Time Between Claims (MTBC) Approach

The class of problems where the dependent variables is restricted to be non-negative is very common in direct marketing applications. For example, in the catalog industry one observes the order size only for people who responded to the offer and bought at least one item from the catalog. No observation of the order size is available for people who did not respond to the offering. In this case, the response (dependent) variable is said to be *censored*. As mentioned above, this type of response variable is not suitable for modeling by linear regression as it violates the basic assumption of the linear regression model.

Several approaches have been devised in the literature to handle censored non-negative response, e.g., the Tobit model (Tobin, 1958). An alternative approach is the two-stage model of Heckman (1979) according to which one estimates the expected claim level for each customer in two stages. In the first, estimating the probability of filing a claim based on all customers in the data set, in the second stage, estimating the average claim (given that a claim had been filed). The (unconditional) expected claim per customer is obtained by multiplying the resulting two "scores": the probability of a claim, and the conditional average claim. The predicted claim value per customer serves as the basis for setting up the differential rate.

The most common approach to estimate the probability of claims in the first stage, also suggested by Heckman, is by means of a discrete choice model, such as logistic regression, where the dependent variable is a yes/no variable. (Yes - for filing a claim; No - for not filing a claim). The explanatory variables are the customer's demographics, claim history and vehicle characteristics. All customers, including those who did not file a claim, participate in this stage. The second stage is usually addressed by means of a linear regression model and involves only customers who actually filed claims in the period involved. The dependent variable is the actual claim value and the explanatory variables the predictors that pertain to claimants.

But in trying to implement the two-stage approach to predicting claim levels in our study, we ran into another complication caused by the data structure that we worked with. As discussed in the next section, each amendment to a policy during the period, either because of a change in the attributes of the policy or due to filing a claim, results in the previous policy being shut down and a new policy opened up. Thus policy records in our data set are not of the same duration. Consequently and in order to bring all records to an equal footing for analysis, one needs to weigh each record by the proportion of time that the policy/amendment was in effect during the period involved. Clearly, the claim events and the duration of the records are highly correlated. Suppose, for example, that a customer has filed two claims in a given year, six months apart. This customer will have two amendment policies, each with an average duration of six months, ending with a claim. By using the proportional duration as weights, each of these claims will carry a weight of 1/2. In an additive model, such as logistic regression, these two six-month policies will receive the same weight in calculating the probability of claim as a single 12-month policy ending with a claim. Thus, unless the study is restricted to policies with uniform duration, it is incorrect to weigh each record in proportion to its duration.

In order to overcome this difficulty we have adopted a different modeling approach to for the first stage, which is based on survival analysis (Cox, 1984). Survival analysis is an advanced statistical method, originated in the field of life sciences and medicine, which is concerned with estimating the time until an event occurs. In medicine, the event could be death due to a sickness or recuperation from an operation. In our case, the event is a claim made by a customer, either as a result of an accident or a theft. Our objective is to estimate the mean time between claims (MTBC) in days. Then, the average number of claims per day is given by the reciprocal of the MTBC. Multiplying by 365 yields the average number of claims per year. For example, MTBC of, say, 122 days translates into an average of three

claims per year; and MTBC of, say, 1095 days, translates into an average of one claim every three years or .33 claims per year. In survival analysis the dependent variable is the time until the next claim. Thus, a customer that has made two claims in the year, as in the above example, will "contribute" three records to the study, two amendment records and a third right censored record with no event, with the dependent variable in each record expressing the time elapsing since the previous accident. If a customer has not filed a claim during the study period, the dependent variable is set equal to the total number of days in the period, but is regarded as censored, implying that we have not waited long enough for the event to occur. Thus it is not required to weigh each record by its duration, as the weight is basically imbedded in the value of the dependent variable.

Furthermore, the MTBC approach enables us to study policies in force for any duration, and there is no need to wait until the end of a month, the end of the quarter or the end of the year before one starts the analysis. Also, each record can be treated on its own, even if a policy is split into various sub-periods. Moreover, under the MTBC approach, all the statistical information may be used, even if no claims were filed in the period. This allows us to use the most recent data, since the method is not affected by the length of the study period. The use of recent data definitely improves the stability of the model and the prediction accuracy. What makes the MTBC approach especially attractive in this study is the assumption that hazard rate in car insurance area is constant, which significantly facilitate the estimation problem. Further discussion of the MTBC approach can be found in Appendix A.

Finally we note that the MTBC approach is not new to the insurance world. Actuaries refer to it as a Poisson process and have been using this process and its extension, the compound Poisson process, for many years (Hossak et al., 1989). The novelty of this paper is that it uses the MTBC approach in the context of a two-stage model to estimate an individual risk level for each customer in the database, which should serve as the basis for setting up car premium levels.

V. The Data

The study was conducted on actual passenger car insurance data (mainly family cars), relating to a particular segment of the motorcar insurance portfolio of an insurance company. In this particular environment policies are issued to each car, although typically more than one driver in the family uses the car. The data set consists of approximately 180,000 insured cars, during 1999-2000. Each customer record contained in excess of 200 attributes divided into four groups:

- **Policyholder attributes** (name, address, young driver flag, new driver flag, etc.);
- **Car characteristics** (make, model year, fleet flag, safety devices, anti-theft devices, and car value);
- **Policy attributes** (policy number, start and end dates of policy, renewal dates, premium, extras, and agent); and
- **Claims details** (date of claims, payment against filed claims, estimated outstanding claims, and type of loss).

A detailed list of the predictors that took part in the model appears in Appendix B.

Except for the actual premiums, all other attributes were used as predictors in our two-stage model. The actual premiums were used in order to compare to the predicted premiums and assess the quality of our model

We note that most policies had a large number of amendments during the policy year including changes of ownership, addition or deletion of drivers, changes in safety devices and others. In addition, whenever a claim was filed, another amendment was made, to reflect the additional premium (no automatic reinstatement of policy value). Technically, each amendment results in the original policy being "closed" and a new policy, with the updated information, opened. Consequently, there could be multiple records for each customer in the file, each with different time duration. As discussed above, this introduced a major complication in the modeling process and, in a way, led us to using the MTBC approach to model the risk level.

As in many real life cases, also in our case the data set was incomplete. The claim history corresponds to a limited period only, due to the very frequent changes of insurers in the particular environment studied. Most insured family cars were driven by more than one driver, but information was available only on the car owner. Young drivers authorized to drive the car, and those holding a license for less than a year were identified in the policy only by using "flags." In addition, the data was far from being "clean" and much work was spent to bring the data to the level it would be useable for the analysis.

VI. The Data Mining Process

The fairly large dimension of our data file (180,000 records, each with several hundred predictors) made it necessary that we automate the model building process using data mining technologies. We built our own computer programs to conduct the analysis, using SAS procedures for the survival analysis and linear regression models.

In building large predictive models, as the one we have, one must guard against over fitting. Over fitting pertains to the phenomenon where one gets a very good fit on the data which is used to build the model, but poor fit when the model results are applied on a new set of observations. To check for over fitting, it is very common to split the examined population into two data sets (files)—a training set, for building the model, and a validation set, containing all the remaining records, for validating the model (Shepard, 1995). The model is first calibrated based on the training set. Then, the output of the model, in our case the two-stage MTBC/regression model, is applied to predict the claim level for each customer in the validation set. A model is deemed "OK", i.e., exhibits no over fitting, if the actual claim levels for the training and validation sets are more-or-less aligned. Furthermore, because the actual claim levels in the validation set are known, one could also compare the actual claim levels to the predicted claims levels to assess how good is fit of the model. The closer the predicted value to the actual results, the better the fit. Note, that time dependent changes may affect the stability of the model over time. One therefore needs to account for these time-related features, perhaps judgmentally, in order to ensure that the models findings based on the current period will also extend to future periods.

In our case, we split the database randomly into two separate and independent populations. One, including about 40 percent of the records, was used for training the model. The second, including the remaining 60 percent of the records, was used for validation. In order to make the two data sets mutually exclusive, all records corresponding to a single original insurance policy were included in the same data set.

It should be noted that we have actually developed two different models, one for claims due to accidents and the other for claims due to theft. The reason is that accidents are affected more by the driver characteristics whereas thefts are affected more by the vehicle characteristics. Combining results of these two sub-models yields the predicted average damage for the insurer.

Table 1: Average Number of Accident per Year

Decile	Predicted number claims per year	Actual number claims per year	Total exposure day.	Number of claims
1	0.4937	0.4307	958,407	1,134
2	0.3168	0.3263	1,033,937	924
3	0.2579	0.2828	1,384,345	1,034
4	0.2225	0.2340	1,549,716	993
5	0.1980	0.2106	1,583,465	914
6	0.1802	0.1923	1,594,194	840
7	0.1643	0.1602	1,609,047	707
8	0.1483	0.1456	1,615,455	644
9	0.1268	0.1274	1,604,721	560
10	0.0933	0.0858	1,639,917	386

VII. The Results

All the results pertain to the validation file, and show the predicted probabilities (based on the MTBC) and average claims, using the models derived from the analysis of the training set. To create the output table, we have ordered the results in descending order of the predicted claim values. For convenience, as well as for the ease of presentation, we have grouped the results by deciles. We use several metrics, which are very common in the direct marketing industry, in order to assess the quality of the predictive modeling results (Shepard, 1995):

- The distribution of the actual claim values across deciles. In a "good" model, one should expect these values to decrease monotonically as one move from the top decile to the bottom decile.

- The ratio of the actual claim value between the top and the bottom deciles. A large difference between these ratios indicates that the model is capable of distinguishing between the more risky people (i.e., people at the top decile) and the less risky people (people at the bottom decile). The larger the ratio, the higher the distinctive power of the model.

- The proximity between the predicted claim values and the actual claim values at the decile level, the closer the values, the more accurate the model.

To recall, two models were built here, one for accidents and one for theft. In Appendix B we describe and analyze the coefficient estimates obtained for the MTBC and the claim value models. In this section, we present and discuss the detailed results of the two models. For confidentiality reasons, we have removed identifying data and disguised the financial results by expressing them as normalized values. Nevertheless this does not affect the significance of the results and no bias was introduced because of this.

The Average Number of Accidents. Table 1 presents the average number of accidents, grouped by deciles, in decreasing order of the predicted number of claims. By definition, each decile contains the same number of records, but each one may have a different number of exposure days (the sum of days for all policies). Two important results are evident from this table. First, the actual average number of claims per year are monotonically declining across deciles, implying that the model is capable of distinguishing between the "risky" and the "less risky" insureds. For example, the most risky insureds are filing, on the average, 0.4307 claims per year, whereas the least risky people have only 0.0858 claims per year. The monotonicity implies that the less risky the insured, the lower is

Table 2: Average Claims Due to Accidents (Predicted vs. Actual)

Decile	Expected claims	Actual claims
1	3.095	2.688
2	2.609	2.589
3	2.412	2.435
4	2.251	2.266
5	2.114	2.224
6	1.977	1.997
7	1.832	1.952
8	1.661	1.821
9	1.429	1.495
10	1.000	1.276

Table 3: Average Number of Thefts per Day

Decile	Predicted number of claims per year	Actual number of claims per year	Total exposure days	Number of claims
1	0.0537	0.0416	786,480	90
2	0.0300	0.0244	911,381	61
3	0.0300	0.0265	1,185,375	86
4	0.0191	0.0205	1,280,784	72
5	0.0164	0.0145	1,334,493	53
6	0.0145	0.0135	1,375,646	51
7	0.0128	0.0114	1,330,310	43
8	0.0111	0.0108	1,384,100	41
9	0.0094	0.0045	1,374,537	17
10	0.0069	0.0037	1,279,257	13

Table 4: Average Total Claims per Day (Actual vs. Predicted)

Decile	Predicted claims per day	Actual claims per day	Total predicted claims	Total actual claims	Total exposure (days)
1	12.39	12.05	17.50	17.02	1,412,498
2	7.43	6.97	10.69	10.03	1,438,887
3	5.93	5.67	8.43	8.07	1,422,153
4	5.04	4.31	7.43	6.36	1,474,154
5	4.36	3.79	6.53	5.67	1,498,612
6	3.79	3.16	5.69	4.75	1,504,133
7	3.21	2.76	4.77	4.11	1,487,454
8	2.57	2.39	3.87	3.60	1,507,054
9	1.79	1.41	2.57	2.02	1,434,783
10	1.00	0.95	1.39	1.32	1,393,476
Total			81.80	74.75	14,573,204

the average number of claims per year. Note, that the actual number of claims per year in the upper decile is more then 5 times larger than the average number of claims per year in the last decile. Second, the differences between actual and predicted values are relatively small, often less than 10 percent.

The Expected Damage Due to Accidents. The claim level may strongly depend on the car value, as more expensive cars may bring about larger damages. This raises the question which dependent variable to use in the second stage of our two-stage model—the actual claim value or the relative claim value (the ratio between the claim and the vehicle value)? While we have experimented with both types of dependent variables, we present below only the results corresponding to the actual claim values (without normalization by car value). We used the vehicle value as an additional predictor in the linear regression model to "explain" the differences in the claim level, which are attributable to the vehicle value.

Table 2 presents the average damage (loss) resulting from an accident, by deciles, in a decreasing order of the predicted (expected) damage. All results in Table 2 were normalized such that the predicted claim level at the lower decile is one.

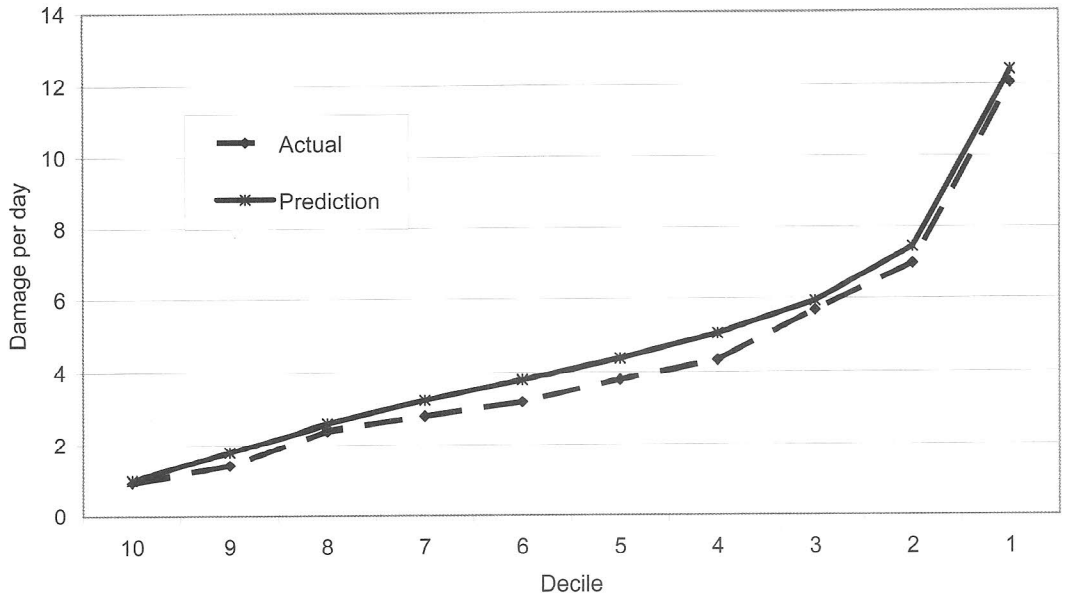
Even though expressed in normalized terms, it is easy to see that both the expected and the actual damage exhibit a monotonic declining pattern as one goes from the top to the bottom decile. The prediction accuracy of the model appears to be very high, with the predicted damage values being quite close, within small percentage points, to the actual damage value at each decile.

The Average Number of Thefts. Table 3 presents the average number of thefts per year, grouped into deciles, in decreasing order of the predicted values. While the number of customers in each decile is the same, the number of exposure days (the sum of days for all policies) varies across deciles. Two important results stand out from this table. First, the average number of actual stolen cars per year is monotonically declining across deciles, except for some minor variations (e.g., in the second decile). Second, the predicted and actual values are quite close to within a few percentage points. The model has an excellent distinctive power with the actual average number of claims per year in the first decile being more then 11 times larger than the average number of claims in the last decile.

We note that in the second stage of the car theft model we used the vehicle value as the dependent variable. Hence, no modeling was involved in this stage and the expected theft value was obtained by multiplying the probability of theft, coming out from the MTBC model, by the car vehicle value. Indeed, this method assumes that the stolen car is completely lost and there is no recovery, which may introduce a small bias in predicting the theft value. However, we believe that in our case this bias was negligible, if any, because the company involved did not require the policyholders to install anti-theft devices as part of the insurance terms. And without these devises the recovery rate are very small.

Combined Results. Finally, we have combined the results from the accident and theft models to yield the total expected claim per customer. The results are presented in Table 4. To create Table 4, we have summed up the expected (predicted) claim level due to accidents and the expected claim level due to theft for each customer in the list, and then summarized the results by deciles, in decreasing order of the total predicted claim per day. All entrees in Table 4, except the total duration days, are normalized such that the predicted claim level at the lower decile is 1.00. The table indicates that our model is indeed capable of distinguishing between the high-risk and the low-risk customers. This is evident by the monotonic declining pattern of the actual claim per day, as one goes from the top to the bottom decile.

Figure 1: Average Damage per Day



Note that the actual average damage per day in the first decile is almost 12 times higher than in the last decile. The first decile itself is responsible for 27 percent of the total claims, almost three times of its relative size. Other than small fluctuations, the results are very accurate across deciles. These results are presented graphically in the chart below.

VIII. Concluding Remarks

The objective of this study is to demonstrate the ability of using sophisticated data mining technology to design risk-based insurance premiums, in a sense that we let the computer interrogate the data and pick the most influential predictors to introduce to the model. We have used a two-stage approach to estimate the risk level of each customer. The first stage estimates the probability to filing a claim using a survival analysis approach, and the second stage estimates the conditional (on filing a claim) claim value using a linear regression model. The findings demonstrate the ability to predict average claim values, to separate between high-risk and low-risk customers, and to give consistent results from a decision-making point of view. The model appears to be useful for the determination of risk-based net premium levels, as well as for underwriting purposes.

Due to obvious reasons, we had not discussed the remarkable differences between the actual rate structure used by the company prior to the study, and the rate scheme suggested by our model. However it is noteworthy that the selected parameters and the weights given to some parameters were substantially different than the actual rate structure. We like to draw the attention of the ability of data mining technology to offer more objective tools rather than basing insurance rates on what seems to be unsubstantiated "beliefs" about the relevant features which distinguish between "good" and "bad" insured.

Clearly, this study could be further pursued in several directions. One is improving the data mining model by adding additional predictors to improve prediction accuracy, thus allowing the model to better identify the high-risk from the low-risk people. To name a few variables: urban/rural indicators, estimated annual mileage, and more details about the drivers. Second is developing an improved algorithm to allocate expenses (loading) between policies according to a more sophisticated method. Other issues are related to the implementation of a risk-based insurance scheme, which we did not touch upon.

Finally, although this study is focused on a particular data base relating to automobile insurance, it is clear that the technology described can also be used to estimate the mean time between claims in almost any type of insurance problem. The automation of the process enables us to develop an accurate model for each claim category and significantly improve the overall prediction accuracy of risk-based claims, which is mandatory in order to serve as a basis for determining risk- prone insurance rates.

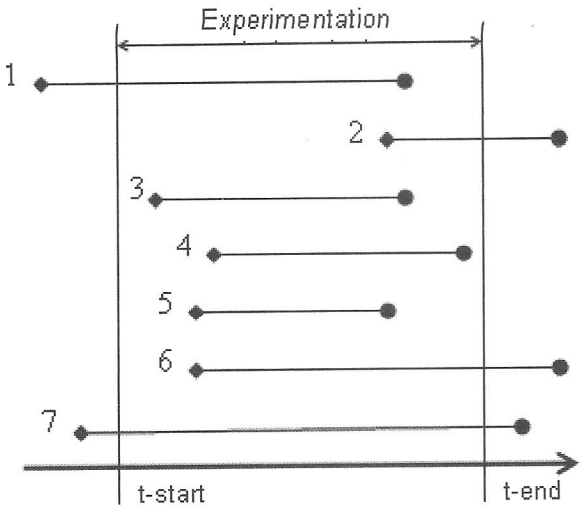
Appendix A—Survival Analysis

Survival Analysis is concerned with estimating the survival time, often the probability distribution function or the expected time, until an event occurs. Usually, but not necessarily, the event corresponds to some kind of a failure, such as the time between failures in reliability studies, the time until death of a patient in health care studies, or the time between claims in insurance studies. In our study, the failure corresponds to claims because of an accident or theft, and we are interested in estimating the mean time between claims (MTBC).

Often, survival analysis is conducted for a given period of time that we refer to as the experimentation period. The time scale could be measured in years, months or days, all of them are linear transformations and equivalent to each other. In the insurance industry, the common experimentation period is one year. Within this period several relevant events may occur, which are, in our case: opening up a policy, changing a policy (e.g., adding a young driver) and filing a claim. When a claim is filed or a new addition to the policy takes place, the previous period is closed and a new period is opened up, starting at that date. The period ends when any of the following occurs: the policy has ended, a new addition is issued, i.e. the previous one is closed, or a claim was filed and a new "experiment" starts.

In the ideal survival analysis experiment, one always knows the beginning period for each observation and has the time to wait until the event occurs. But life is not that simple. There are cases where the beginning time is not known and the only information known is when the observation entered the experimentation period. For instance, we may know exactly the time a patient entered the experiment but not the exact time the disease started. This case is referred as left censoring. Right censoring, on the other hand, occurs when an observation has reached the end of the experimentation period without the event actually happening. In fact, most of the observations in our car insurance case end without a claim and are therefore right-censored. Of course, both types of censoring can occur in the same observation, for example a policy that started out prior to the experimentation period (left censoring) and ended up with no claim during the experimentation period (right censoring). The different types of censoring are illustrated in the chart below for an experimentation period that starts at time "t-start" and ends at time "t-end." Observations 2-6 begin during the experimentation period. Observations 1 and 7 are left-censored as they begin before the experimentation period. Observations 2, 6 and 7 are right-censored ending up with a failure

Figure 2: Various Cases and the Experimentation Period



(e.g., claim) beyond the experimentation period. By definition, the starting time for left-censored events and the ending time of the right-censored events are usually not known to the researcher.

Clearly, the survival times are random variables with a given probability distribution. It is common in survival analysis to represent the survival process by means of two measures:

Survival function—the probability that the time until failure T will occur after time t:

$$F(t) = p(T \geq t)$$

Hazard function—the conditional failure rate at any given time. The hazard function is a non negative function defined by the density that a failure occurs at infinitesimally short period of time, given that a failure has not occurred earlier, i.e.:

$$h(t) = \lim_{\Delta \to 0+} \frac{p(t \leq T < t + \Delta \mid t \leq T)}{\Delta}$$

From a mathematical point of view, these two measures are equivalent in the sense that given one function, one can derive the other. Often one models the hazard function because it is easier to estimate by means of exogenous and time-dependent explanatory variables using, say, proportional hazard regression models (PHR).

A special case of survival analysis is when the hazard function is constant, i.e., the instantaneous failure rate is fixed over time. This distribution has no memory, implying that given that an observation (e.g., a failure) survived until time t, the probability of a failure in the next time period (t+1) is the same regardless of the time that elapses since the time of origin. It is common in survival analysis to denote the constant hazard rate by $h(t) = \rho$. In this case, one can show that the survival function has an exponanial distribution $F(t)=\exp(-\rho t)$ with a parameter $\rho = 1/\tau$ where τ equale to the MTBC.

In the car insurance problem we assume that the probability that an accident occurs in the next time period (say next month) does not depend on the time that elapsed since the previous claim was filed neither on the time since the beginning of the policy. This assumption implies that the process has no memory and that the rate of claims for a customer per unit of time is constant. This assumption holds especially as long as the characteristics of the driver or the car have not changed.

The fact that the process has no memory significantly facilitates the analysis as one can look at any snapshot of the process, without regards to what happened in the past. Thus one does not need to worry about left-censored events. Mathematically this implies that splitting a single period of time into sub-periods would not change the likelihood function. The fact that we can right censor all events with no claims during the observation period, makes the end of the observation period as a "natural" point to analyze the process at, and still make use of the full amount of information at our disposal.

To estimate the hazard function, we express it as a function of the various policy attributes, such as the age of the driver, the type of car, the claims filed in the past, etc. Denoting these attributes by $x_1,\ldots,x_n$, we obtain, in the case of a constant hazard function,

$$h(t,x_1,\ldots,x_n)=\rho(x_1,\ldots,x_n)\ .$$

where $\rho(x_1,\ldots,x_n)$ is some non negative function of $(x_1,\ldots,x_n)$.

The most common function used to for constant hazard function is $\rho=exp(\alpha+\Sigma\beta_i x_i)$ where α is a constant and β_i are the coefficients of the policy attributes.

The coefficients α and β's are estimated using the method of maximum likelihood. It is beyond the scope of this paper to describe the maximum likelihood method. Suffice it to say that the likelihood function account for all types of censored events (only right-censored events, in our case). Typically, censored observations carry some statistical information that is built into the likelihood function. For example, a policy that ended without any claim, i.e., a right-censored policy, has statistical value that will reduce the probability of a claim or equivalently increase the MTBC. For further details, one can refer to the literature (e.g., Cox and Oakes, 1984).

Equivalently, we note that one can model the number of accidents in a given time period, which in the case of exponential failure times is given by the Poisson distribution with a parameter $\lambda = \rho = 1/MTBC$. But this requiers that we aggregate customers into segments based on the different characteristics of the policy. In fact, this is the approach often used in practice to estimate the MTBC. But, as claimed above, because this process is based on segmentation, it may result in a loss of information.

Appendix B—Model Estimation Results and Analysis

In this appendix we analyze the coefficient estimates of our models. As mentioned above, claims due to accidents and claims due to theft are basically two different animals—the former is affected by the driver and claim history characteristics, and the latter by the vehicle characteristics. As a result, we developed a different model for each type of claim.

Accident Model. Table B.1 exhibits the coefficient estimates of the predictors in the final MTBC model for claims due to accidents (the first model in the two-stage model).

Table B.1: Coefficient Estimates in MTBC Model for Accidents

Predictor	Coefficient
Model year	-0.0274
Engine size	-0.000244
No-claim bonus 1	0.3473
No-claim bonus 2	0.4661
No-claim bonus 3	0.2067
Young driver	-0.2753
New driver	-0.4196
Safety devices	0.6544
Fleet car	0.1322
Claim frequency of the agent	-3.2745
Claim frequency for geographical area	-5.2740
Young driver co-payment	-0.1895
Third party insurance	1.2137

Table B.2: Coefficient Estimates in Claim Value Model for Accidents

Predictor	Coefficient
Model year	-108.88
Engine size below 1400cc	-4.6385
Engine size above 1400cc	3.0630
No young driver	-946.10
Safety devices	-429.74
Replacement care option	594.09
Extra premium at time of claim	-520.76
Window insurance	-460.69
Private insurance	872.61
Car category 2	-941.19
Car category 3	-582.94
Car category 5	-868.10
GMC, commercial	-1995.1
Toyota, commercial	-1717.0
Toyota, private	-757.1
BMW	7177.4
Mitsubishi, commercial	-1007.0
VW, commercial	1467.1
VW, private	-965.76
Citroen, commercial	-1748.1
Citroen, private	-1049.3
Daihatsu	-925.23
Hyundai	-1314.9
Nissan	2196.7
Seat	-2257.7
Renault	-884.40
Chevrolet	-1225.0
Claim frequency for agent	
$P < 0.0985$	-576.01
$0.0985 \leq P < 0.10774$	-995.05
$P \geq 0.10774$	-719.46
Claim frequency for geographical area	
$P < 0.0898$	-1058.0
$0.0898 \leq P < 0.09773$	-1070.3
$0.09773 \leq P < 0.10335$	-1107.3
$P \geq 0.10335$	-1605.4
No-claim bonus	663.98
No-Saturday driving discount	994.53
Third party insurance	-3099.3

To recall, the MTBC prediction problem was formulated in our paper as a constant-hazard survival model for which the MTBC is given by:

$$\text{MTBC} = \exp\left(\hat{\beta}_0 + \sum_i \hat{\beta}_i \, X_i\right)$$

where $\hat{\beta}_i, i = 0,1,2,\cdots$ are the coefficient estimates and $X_i, i = 1,2,\cdots$ the predictors.

A negative coefficient estimate implies that the MTBC decreases with an increase in the value of the corresponding predictor (holding all the other predictors constant). And with shorter MTBC, the frequency of filing claims increases, and so is the probability of filing claims. The contrary holds for positive coefficient estimates, where the claim probability increases as the value of the corresponding predictor increases.

In interpreting the coefficient estimates we need to distinguish between categorical and numerical variables. For example, being a young driver (coefficient = –0.2753) increases the probability of filing a claim as compared to being a "veteran" driver. Likewise, having certain safety devices in the car (coefficient = 0.6544) contribute to decreasing the claim probability versus not having these safety devices. The variable "claim frequency of the agent", on the other hand, is a numerical predictor. Its negative coefficient implies that the larger the number of claims filed by an agent in the past, the higher the probability of filing a future claim.

We note that most coefficients in Table B.1 agree with our prior hypothesis regarding the impact of the various predictors on the claim probabilities. For example, we expect higher risk for young and new drivers, and lower risk for drivers with good driving record (as reflected by the no-claim bonus variables). The impact of some predictors may not be as intuitive. For example, newer cars (exhibited by the model year variable) and bigger cars (represented by the car engine variable) have negative coefficients, thus increasing the claim probability for these types of cars. This is because newer and bigger cars are being used as the primary car of a household, spending more time on the road and driving longer mileage, and thus tend to be involved in accidents more than older and smaller cars.

It is interesting to note that we found a significant positive relationship between the probability of filing a claim and the record of the agent, as reflected by the frequency of claims filed by the agent. Also, certain geographical areas (identified by the two first digits of the zip code), which experience many claims in the past (i.e., are "bad"), contribute to increasing the risk. In all likelihood, these predictors may serve as "surrogate" to other predictors, which we do not know about, such as income, education level, socio-economic status, etc. Consequently, these predictors are very important in predicting risk.

Table B.2 presents the coefficient estimates of the predictors in the claim value model (the second model in the two-stage model). As expected, the influential predictors relate to the car characteristics (type, make, size). Positive coefficients in this model increase the claim value, and negative coefficients decrease it. Thus, very small engines were associated with lower claim value; a replacement car option (having a car from the insurer while the damaged car is in the garage) increases the claim value. The particular make of the car may be associated with particular characteristics of the drivers (income, education, etc.) and also with the price of spare parts. The contribution of the predictor: "claim frequency of the agent" and "claim frequency by geographical area" to the claim value is nonlinear. The various car makes have different impact on the claim value—some contribute more, some less; some result in an increase in the claim value, some in a decrease. There is no way to assess in advance how each type of car is going to affect the risk. It all depends on the profile of audience driving each car type and the pattern of use—which are not known to us.

Table B.3: Coefficient Estimates in MTBC Model for Theft

Predictor	Coefficient
Model year	0.037605
Engine size	-0.0006048
Safety devices	0.7686
Fleet car	0.7328
Theft frequency for agent	-30.86
Theft frequency for geographical area	-31.79
Single driver	0.3441

Theft Model. Table B.3 exhibits the explanatory variables in the MTBC model for theft. Note that the number of predictors in this model is smaller than in the MTBC model for accidents, because we had fewer observations to build the model with. Indeed, the car characteristics play a major role in this model. Newer cars (represented by the model year variable) are associated with lower probability of theft, probably because they have advanced safety devices that make them harder to steal. On the other hand, older cars are sometimes needed for spare parts which many increase the "demand" for these cars by thieves. Cars with single driver reduce the probability of theft, perhaps because s/he may be taking better care of the car. It is interesting to note that also in the theft model the predictors "theft frequency for the agent" and "theft frequency by geographical area" contribute to increasing the probability of theft.

Finally, we note that unlike in the accident model, there's no claim value model for theft. The reason is because we use the car value as a proxy for the damage incurred for stolen cars. The bias incurred by this, if any, is minimal because of the low recovery rate of stolen cars, for the particular company that we worked on.

References

Ai, Cheo Yeo, K.A.Smith, R.J. Wills, and M. Brooks (2001). "Clustering Techniques for Risk Classification and Prediction of Claim Costs in the Automobile Insurance Industry," *International Journal of Intelligent Systems in Accounting Finance & Management*, 10: 39-50.

Apte, C. Grossman, E. Pednault, E. Rosen, B. Tipu, F. and B. White (1999). "Insurance Risk Modeling Using Data Mining," *Proceedings of The Third International Conference on The Practical Applications of Knowledge Discovery and Data Mining*, IBM Research Division Technical Report RC-21314.

Apte, C. Grossman, E. Pednault, E. Rosen, B. Tipu, F. and B. White (1999). "Probabilistic Estimation Based Data Mining for Discovering Insurance Risks," *IEEE Intelligent Systems, Volume 14, Number 6, November/December 1999* IBM Research Report RC-21483.

Cox, D.R., and D. Oakes (1984). *Analysis of Survival Data*, London: Chapman and Hall.

Fayyad, U, G. Piatetsky-Shapiro, and P. Smyth (1996). "The KDD Process for Extracting Useful Knowledge From Volumes of Data," *Communications of the ACM*, 39: 27-34.

Hart, D.G., R.A.Buchanan, and B.A.Howe (1996.) *The Actuarial Practice of General Insurance:* Institute of Actuaries of Australia, Sydney.

Heckman, J.J. (1979). "Sample Selection Bias as a Specification Error," *Econometrica*, 47: 153-161.

Hossack, I.B., J.H. Pollard, and B. Zehnwirth (1983). *Introductory Statistics with Applications in General Insurance*, London: Cambridge University Press.

Johnson , P.D., and G.B. Hey (1971). "Statistical Studies in Motor Insurance," *Journal of the Institute of Actuaries*, 97: 199-249.

Kahane Y., and H. Levy (1975). "Regulation in the Insurance Industry: Determination of Premiums in Automobile Insurance," *Journal of Risk and Insurance*, 42: 117-132.

Kietz, J.U., U. Reimer, and M. Staudt (1997). "Mining Insurance Data at Swiss Life," *Proceedings of the 23rd VLDB Conference* (Athens, Greece): 562-566.

Lemaire, J. (1990). *Automobile Insurance: Actuarial Models Huebner Series in Insurance*, Boston: Kluwer Academic Publishing.

E.P.D. Pednault, B.K. Rosen, and C. Apte (2000). "Handling Imbalanced Data Sets in Insurance Risk Modeling," *IBM Research Report* RC-21731.

Shepard, D. (Ed.) (1995). *The New Direct Marketing*, New York: Irwin Professional Publishing.

Staudt, M., J.U. Kietz, and U. Reimer (1997). "ADLER: An Environment for Mining Insurance Data," *Proceedings, 4th KRDB Workshop*, Athens, Greece: 16.11-9.

Tobin, J. (1958). "Estimation of Relationships for Limited-Dependent Variables," *Econometrica*, 26: 24-36.

Viveros, M.S., J.P. Nearhos, and M.J. Rothman (1996). Applying Data Mining Techniques to a Health Insurance Information System," *Proceeding of the 22nd VLDB Conference* (Bombay, India): 286-294.

Williams, J. G., and Z. Huang (1996). "A Case Study in Knowledge Acquisition for Insurance Risk Assessment Using a KDD Methodology," *Pacific Rim Knowledge Acquisition Workshop* (Sydney, Australia).