December 2007

Price and Revenue Optimization

Daniel Ossa Rodriguez
Worcester Polytechnic Institute

Esteban Ramiro Paez
Worcester Polytechnic Institute

Fernando Miguel Olloqui
Worcester Polytechnic Institute

Project Number: **JPA0702**

PRICE AND REVENUE OPTIMIZATION

A Major Qualifying Project Report:

submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

Degree of Bachelor of Science

by

____

Fernando Olloqui

____

Daniel Ossa

____

Esteban Paez

Date: January 10, 2007

Approved:

____

Professor Jon Abraham

Abstract

This project applies mathematical techniques to design a process by which an insurance company can optimize revenue. The main objective is to categorize client portfolios using K-means algorithm to segment the market. After clusters are obtained, a logistic regression is applied to determine the optimal premium increase/decrease that maximizes the revenue for each cluster, based its specific characteristics. Applying the optimal premium change to each customer subgroup, the firm will increase its overall revenue.

Executive Summary

The goal of this project was to design a process by which an insurance company can develop a dynamic pricing strategy that optimizes revenue through effective client segmentation. A dynamic pricing strategy or price and revenue optimization is the science of determining what products and services to offer to which customer segments, through which channels and at what prices in order for a company to maximize profit and meet strategic objectives.

Attempting to design and apply such strategy within the insurance industry could be quite of a challenge. Essentially, the objective is to speculate what will the customer do when the price to renew his/her insurance increases or decreases? What is the optimal increase/decrease? How many customers is the insurance firm willing to lose by an increment in premium that would ultimately become an increment in profits? What reaction will the competition have in response to this new strategy? The team will try to predict customer behavior patterns, and their potential effects in revenue and market share that could help experts answer these questions.

The team has outlined the Price and Revenue Optimization process as a five stage process:

- Data Cleanup
- Customer Indicators - variables to be used
- Customer Segmentation – Clustering
- Demand Estimation - Logistic Regression
- Revenue Optimization

The data provided needed some adjustments. The team had to fix, clean up and even delete a major portion of the data because of incorrect entries and the possible consequences of using defective data. For example, using a driver's age less than sixteen or seventeen or greater than 100 would be using insignificant data. However, in the real world, perfect data does not exist, and such errors due occur. Therefore, it was crucial to take a good look at the entire population and make sure that the data that the team will use would provide good results.

"Customer Indicators" such as the type of fuel used in the car , the sex, the hometown, the profession, the driver rating, the premium, etc. of the policyholder, are several variables the team analyzed in order to identify key indicators that would determine which variables will be used in the segmentation. The team focused on identifying those variables that showed independence in renewal rates among categories within the variable.

For instance, if a variable such as sex showed that 85% of the males renewed the policies offered, but only 45% of the females renewed their policies, then this variable becomes a good indicator since it shows a significant degree of variation between its categories. Now, if a variable such as fuel type is analyzed, and of the two types of fuel, diesel and gas, diesel users have a 70% renewal rate and gas users have a 68% renewal rate, this variable is not a good indicator of variation or independence, therefore it was ignored.

Customer segmentation or customer differentiation is the process by which the data are grouped based on the similarities of the indicators. Each policyholder has different reasons upon which the renewal of the policy is based on. The objective of customer segmentation is to identify those customers, who have different characteristics, but this "difference" is small enough to treat both customers as one, this is known as "Data Clustering."

The data clustering process was conducted with the help of the statistical software SAS. This software comes with a package of procedures that provided essential help throughout the project. One of these procedures is called FASTCLUST. This procedure clusters data points using the "K-means" algorithm. The K-means algorithm computes the smallest Euclidean distance between data points, and merges or groups that most similar ones, leaving only a certain number of groups containing variables with similar characteristics (smallest Euclidean distances). Consider the following example:

Indicators	**Sample Data Point 1**	**Sample Data Point 2**	**Scale**
Sex:	Male	Male	M,F
Age:	65	68	20-70
Fuel Type:	Gas	Gas	Gas, Diesel
Rating:	1	2	1-15
Maximum Coverage:	1,000,000	1,000,000	0.5 to 2 million
Vehicle's Age:	2	2	0-5
Years insured:	5	6	1-7

Table 1 - Sample data points

Looking briefly at the characteristics between the data points, it could be inferred that both of them are quite similar. There are minor differences, but if the data are standardized, those differences become quite small. This example shows two data points, with slightly different characteristics. When the FASTCLUST compares these points, the Euclidean distance among the categories would be very small, hence merging both data points as one. Iterating this process through all the data points using the specified variables would produce a certain number of clusters (10 for this project) that could be further used in the Demand Estimation stage. The team iterated the process for three techniques, each with different variables, in order to see whether or not some indicators provided better results than others.

Demand estimation is the use of statistical techniques to determine how will demand be affected by a change in the price of a given product. The laws of demand and supply establish that, and increment in price will result in a decrement in the quantity demanded. The real question is, how sensitive is the demand for this product relative to price changes? The team used the logistic distribution as an approximation to the distribution of the market's demand. The logistic distribution is commonly used to model binary response variables; in this project the response variable is the final decision of the customer of whether or not to renew the policy at the proposed premium. The logistic regression model uses "predictor" variables (sex, age, profession.. etc) to model the "response" variable or the renewal decision. Again, using a procedure called LOGISTIC within SAS, the team was able to approximate the demand for each of the clusters produced using the parameters outputted by this procedure. With the estimated demand in place and the use of the proposed premiums, the team was ready for the Revenue Optimization stage of the process.

Revenue optimization, as its name suggests, is the process of optimizing or "maximizing" the revenue. This optimization strategy, simply put, determines which customers are sensitive to price changes and which ones are not. The model identifies the optimal combination of premium changes and sensitive customers in order to maximize revenue. Using the model parameters, the team was able to compute the renewal retention at any given increment or decrement on the spectrum. The students narrowed the increment/decrement interval as -15% to 15%, with 0.5% bins. Using the mean premium for each cluster, and the retention rate at each increment/decrement, the students computed the maximum revenue for each cluster, for each of the three techniques used.

The following two tables summarize the results for Technique 1:

Technique 1									
Cluster	**Renewal**	**Cancel**	**Total**	**Premium**	**Coverage**	**Age**	**Acc**	**Yrs**	**Age (c)**
1	39937	2418	42355	295	1,696,123	69	0	1	10
2	3987	352	4339	401	4,723,230	52	0	1	6
3	46573	4385	50958	330	2,089,283	52	0	1	3
4	24179	3876	28055	447	1,666,005	56	2	2	7
5	3985	473	4458	336	1,449,659	43	0	2	10
6	30416	4106	34522	593	1,656,216	41	0	2	5
7	6424	1385	7809	331	1,763,391	60	1	1	6
8	7489	607	8096	307	1,174,040	59	0	7	10
9	34152	1846	35998	991	1,864,934	46	1	1	6
10	53823	5975	59798	529	1,718,117	42	1	2	6
Total	**250965**	**25423**	**276388**						

Table 2 – Cluster Summary for Technique 1

Table 2 shows the characteristics for each cluster under this technique. Cluster 1 had 39,937 renewals, and a total frequency of 42,355 people. The average premium for this cluster is 295 dollars. This means that, on average, people with a premium close to this one are allocated in Cluster 1. A similar analysis is used for the Maximum Coverage, the age of the driver, the number of accidents, the number of years insured and the age of the car.

Technique 1						
	Actual			Prediction		
Cluster	**Δ**	**ρ**	**Revenue**	**Δ**	**ρ**	**Revenue**
1	-1.6%	90.9%	11,406,151	-2.0%	94.2%	11,767,829
2	-2.5%	92.5%	2,398,598	12.0%	85.7%	2,551,566
3	-1.5%	93.9%	17,491,447	0.0%	94.3%	17,828,537
4	4.1%	91.9%	4,696,954	123.0%	68.4%	7,489,889
5	-2.8%	89.1%	13,873,691	-3.0%	91.8%	14,255,718
6	-4.9%	87.0%	14,773,160	-6.0%	90.1%	15,124,401
7	0.5%	95.6%	7,531,396	30.0%	84.3%	8,599,116
8	-1.3%	87.9%	8,745,360	-3.0%	92.7%	9,059,811
9	-5.5%	80.6%	5,258,937	72.0%	59.0%	7,008,391
10	-0.9%	93.6%	8,295,992	96.0%	72.0%	12,615,683
Total			**94,471,686**			**106,300,939**

Table 3 – Revenue and Retention figures for Technique 1

Table 3 shows the revenue and retention figures. The delta symbol Δ represents the optimal percentage increment or decrement that should be offered to a policyholder based on the previous year’s premium. The ro symbol ρ represents the optimal retention rate linked to the Δ% in premium. As you can see, the model is not perfect. It suggests premium increments of 123% or 96% which in real life will never occur. However, looking at Clusters 1,3,5,6,8 the results show figures very similar to the ones used by the insurance firm. Ultimately, our goal to optimize and hence maximize revenue designing a systematic approach was achieved. However, there is plenty of room for improvements and even more questions to be answered.

The purpose of this project was to develop a general model by which an insurance company could adapt to PRO and remain competitive in the field. The mathematical model developed produced solid results that could potentially be used by any firm who decides to pursue the study of revenue optimization. The model does answer many of the questions asked at the beginning of this summary, however it also lead to more. Although the goal was achieved, the group and the overseeing advisors still asked themselves, how accurate is the model? Could the model be used for other insurance fields? how can the model integrate competitor’s reactions? Are there any other factors that have been overlooked?

The students believe there is plenty of room for improvement in this model. However, this project was a series of major first steps toward innovation in the insurance world. The industry is ever-changing, and the availability of information is much greater than it was before. From an academic standpoint, the success of the project is shown not by the completion of the objectives, but through the unanswered questions that the achievement of the goal produced.

Acknowledgements

Table of Contents

Table of Figures

Table of Tables

1 Introduction

The insurance industry is one that constantly has to change and adapt to meet both client satisfaction and local regulations while remaining profitable. One of the newest tools that are currently being discussed in the industry to meet these criteria is the use of Price and Revenue Optimization (PRO). “PRO is the science of determining what products and services to offer to which customer segments, through which channels and at what prices in order for a company to maximize profit and meet strategic objectives” **(Krikler, Dolberger, Eckel, 2004).** The key phrase in this sentence is meeting “strategic objectives”, as this strategy can be used to increase profitability, market share or any other goal a company may want to achieve.

Even though this concept is new when applying it to insurance products, PRO has been used for over 20 years in other fields. The first one to apply this concept was the airline industry **(Krikler, Dolberger, Eckel, 2004).** Fierce competition from no-frills airlines lead the main players to develop sophisticated analytical strategies to match the market’s demand to their available supply. In this case, it involved adapting ticket prices to different demand characteristics, obtained from diverse customer segments. The great success of PRO in the airline industry was soon discovered by other industries and was quickly adopted by car rentals, hotels, cargo companies, retails and automotives. Now this trend has extended to the Insurance industry, where each of the companies will have to implement PRO, in order to remain competitive in the field. “Expertise in price optimization will become a core competency of all insurers in the market**.”(Towers Perrin, 2007)**

Implementing PRO requires there four main steps: data collection, demand estimation, price optimization and monitoring **(Krikler, Dolberger, Eckel, 2004).** Briefly, this strategy utilizes client segmentation to calculate demand elasticity and this way understand how a specific price change will affect different client types. The authors worked with data from an Italian car insurance company, which was provided by Towers Perrin, to simulate and construct a standard process by which any insurance company could achieve an optimal pricing strategy. The conclusions and recommendations from this paper will aid the current efforts to implement and adapt the insurance industry to more precise pricing strategies.

2 Background

The main purpose of this background section is to research about many of the main concepts of the project. The research focused on topics such as price optimization, price-elasticity. More specifically how they are measured, and how they relate to the insurance world. This section will also show some basic examples that will help understand this topic in the in the insurance world context.

Pricing strategies are extremely important to any business, company, or corporation that intends to be profitable. From a small town market with twenty customers to huge corporations such as Coca-Cola, Microsoft, or Google, the way they price their products or services makes a difference in their income at the end of the year. However, pricing might not be that easy in some types of businesses. For example, airlines use complex ticket pricing techniques which take into account hundreds of factors such as crude oil prices, origin and destination of the flight, time of the year, type of seat, and so on. There might be a significant price difference if you book a flight for a random Saturday afternoon in February, than if you book that same flight for the Wednesday before thanksgiving; even if it's the same route and the same carrier.

The insurance world is as or perhaps more complex. The main reason is that insurers have no factual way of predicting if, when, or how a claim will occur. Every year, insurance firms around the world put enormous effort "into setting prices according to very fine-grained segments- age, driver history, car type, location, etc.- and some of their interactions" **(Orlay, Davey & Howard, 2004).** The problem is that many of these insurers do not know these different prices on their profits. Answers to questions such as:

- Will the company make more or less profits if you increase prices in a certain segment?
- Will it be better off by increasing prices and thus losing some customers, or lowering them to attract more new customers?
- Which specific client-segment is more/less sensible to a price increase/decrease?

The answers to these questions are certainly not easy to obtain; but certainly, they would be a lot easier to figure out if insurance companies knew exactly how elastic (or inelastic) each of these segments are. In other words, if they were able to accurately predict how each niche population would react to a sudden price change. With this information insurance companies could multiply their profits by charging clients as much as they are willing to pay.

This is precisely what price optimization is. It is defined as "the integration of demand-side pricing (a customer's willingness to pay) into an overall pricing strategy" **(Sanche, Towers Perrin, 2007)**. The idea is to determine prices by considering not only supply-side factors (what the service/product costs to be provided/produced, plus a profit margin) but also by combining these with demand-side techniques (pushing customers to the limit). This strategy ultimately seeks to provide the insurer exact information so it can exploit a particular strategic objective, generally customer volume or profitability, while adapting to the changing business environment **(Sanche, Towers Perrin, 2007).** To have a more thorough understanding of the tradeoff between volume and profitability, the example below **(Orlay, Davey & Howard, 2004).** quantifies these two factors showing how each can contribute to a company meeting its financial objectives.

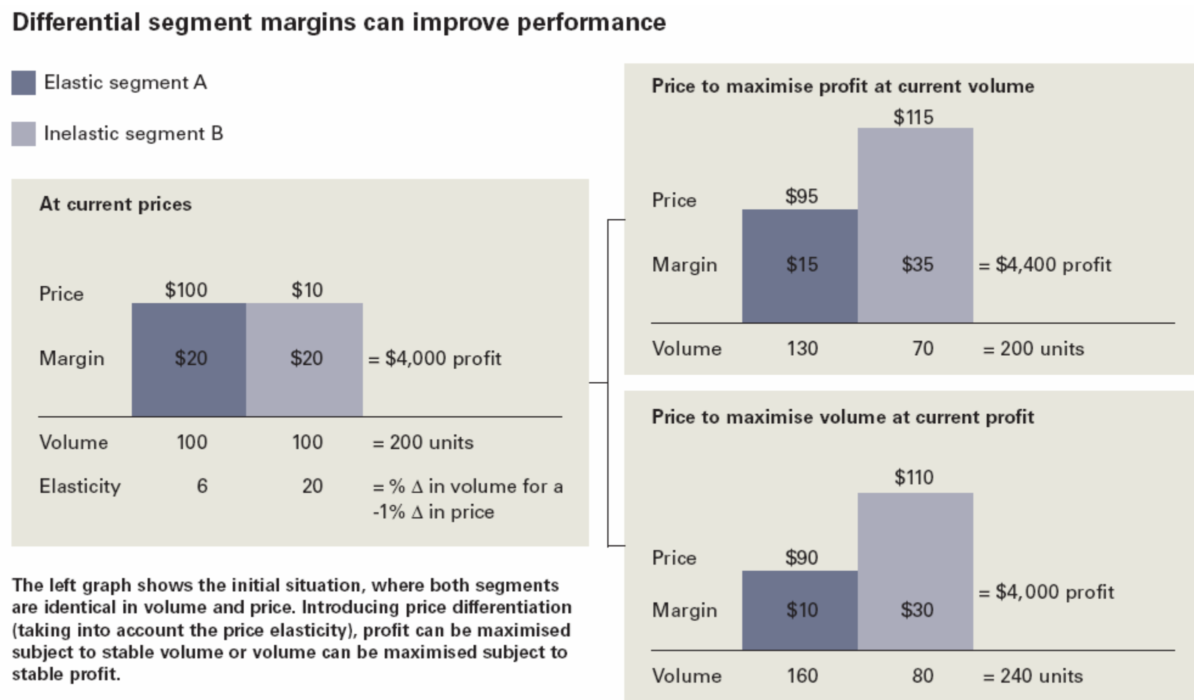

Figure 2-1 - Volume vs Profitability

There are three main components to a price optimization program **(Towers Perrin, 2007)**:

3. **Claim propensity models:** These models express how particular customer attributes are predictive of their tendency to report a claim. They are used to develop rating plans or customer scoring systems for underwriting.
4. **Market situation models:** These express how the company's position among the competition and how the market's competitive intensity will vary by segment or niche within the market.
5. **Customer behavior models**: These models convey how client's attributes combined with the market's situation are predictive of behavior.

This Major Qualifying Project will deal with the first of these three components, as our model will only be a mathematical prediction. Other factors such as competition, economy, and customer's actual behavior are mucho more complex to analyze. The outcome of this project will thus be only the first step in a long process. From our results experts can then consider external factors using market data and good judgment.

3 Methodology

The team has outlined the Price-Revenue Optimization process as a five stage process:

- Data Cleanup
- Customer Indicators - variables to be used
- Customer Segmentation – Clustering
- Demand Estimation - Logistic Regression
- Revenue Optimization

The data provided to the group, though massive, needed some adjustments. The team had to fix, clean up and even delete a major portion of the data because of its condition and the way it was going to be used. As it was assumed, working with real data would imply dealing with such an "imperfect" data set.

"Customer Indicators" such as the type of fuel used in the car , the sex, the hometown, the profession, the driver rating, the premium, etc. of the policyholder, are several variables the team analyzed in order to identify the key indicators that would determine which variables will be used in the segmentation.

Customer segmentation or customer differentiation is the process by which the data are grouped based on the similarities of the indicators. Each policyholder has different reasons upon which the renewal of the policy is based on. The objective of customer segmentation is to identify those customers, who have different characteristics, but this "difference" is small enough to treat both customers as one, this is known as "Data Clustering."

As discussed in our background, demand estimation for renewal policies will be determined through a logistic regression, fitting a logistic distribution. The logistic distribution is commonly used to model binary response variables; in this project the response variable is the final decision of the customer of whether or not to renew the policy at the proposed premium. This tool will enable the team to model the demand for each of the segments identified in the clustering.

Revenue optimization, as its name suggests, is the process of optimizing or “maximizing” the revenue. This optimization strategy, simply put, determines which customers are sensitive to price changes and which ones are not. The model identifies the optimal combination of premium changes and sensitive customers in order to maximize revenue.

3.1 Data Clean-up

The data set given included three kinds of policies: New, Renewed, and Cancelled policies. This study focuses only on retention, hence every policy classified as “New” had to be eliminated from the data set. The initial data set contained 760,233 policies, after the deletion of 249,558 “New” policies, the team was left with 510,675 data points to analyze.

One of the main variables in this study is the “Age” of the policy holder. Because this variable is so important, the team had to make sure there were no errors linked to this variable. There were two main problems related to this variable were: the existence of negative ages and positive ages ranging from 0 to 17. Italy’s legal age to drive is 18 years, therefore it did not make sense to have minors paying insurance premiums for their cars. Furthermore, the existence of negative ages suggests manual input mistakes, therefore the team determined that any values under 20, including missing and negative values had to be deleted. The deletion of these 32,392 entries, left the students with 478,283 data points.

The students created a variable, “Number of Accidents” that counted the number of accidents for the last 6 years for each policy holder. This variable was a potential indicator of renewal trends among the data, so it was important to exclude any errors from this variable. Furthermore, we merged the categories for 4 and over accidents for each policyholder into one category of 4 accidents. This variable had a few missing entries, and negative values that had to be deleted. In the same step, the variables “Previous Premium” and “Proposed “Premium” were also analyzed, and any missing entries were deleted. This entire step deleted 72,826 entries.

It was brought to the team’s attention that the variable “Age of the Car” had a default value for its missing entries. This default value, age 3, inflated the frequency of this variable at that point. There were to options as to how this could be fixed. The first option suggested that the team could delete every entry with age 3, and infer from the results of ages 2 and 4. The second option was to interpolate between ages 2 and 4, and fit a distribution based on the frequency of this variable to determine the value of age 3. Implementing the first option implied deleting a major portion of the data set, 101,509 data points. On the other hand, implementing the second option mean deleting less entries but it could have potentially contaminated the data set. Since the initial data set was quite massive, the group decided it was best to delete every entry with age 3, rather than to contaminate the data. The team also decided to merge car ages 14 and 15 in one category as age 14 and ages 16 and over in one bin of age 15. Figures 3.1 and 3.2 show the frequency of the variable “Age of the Car” before and after the deletion of all values when age 3.

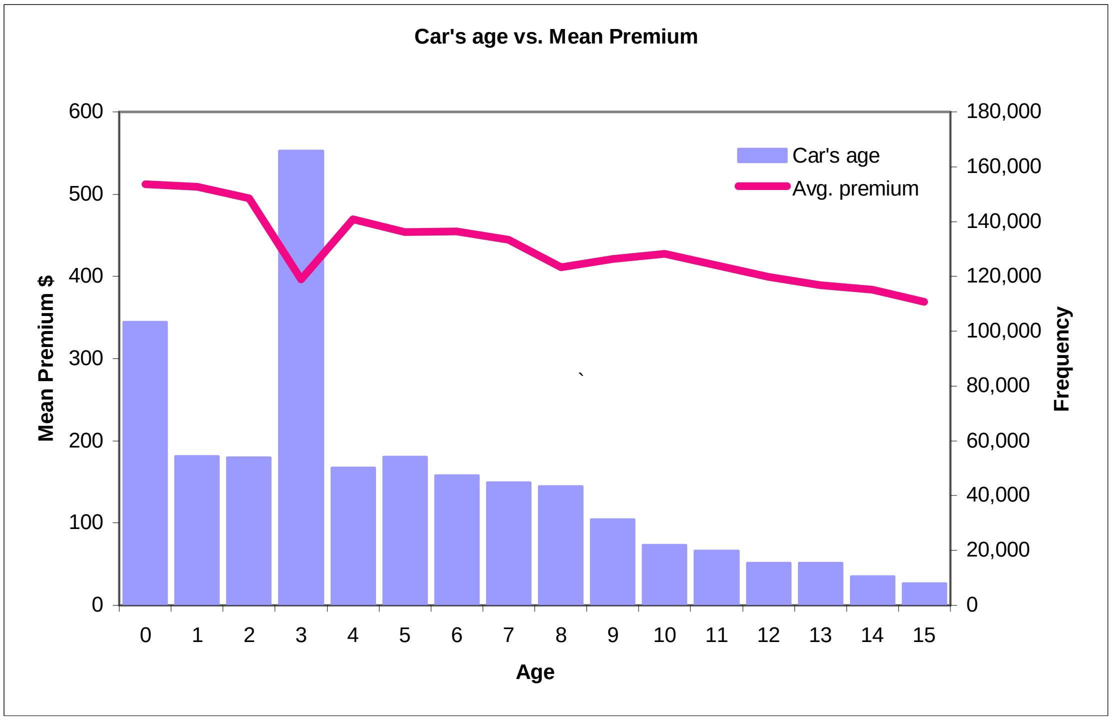

Figure 3-1 - Car's age vs Mean Premium (Before)

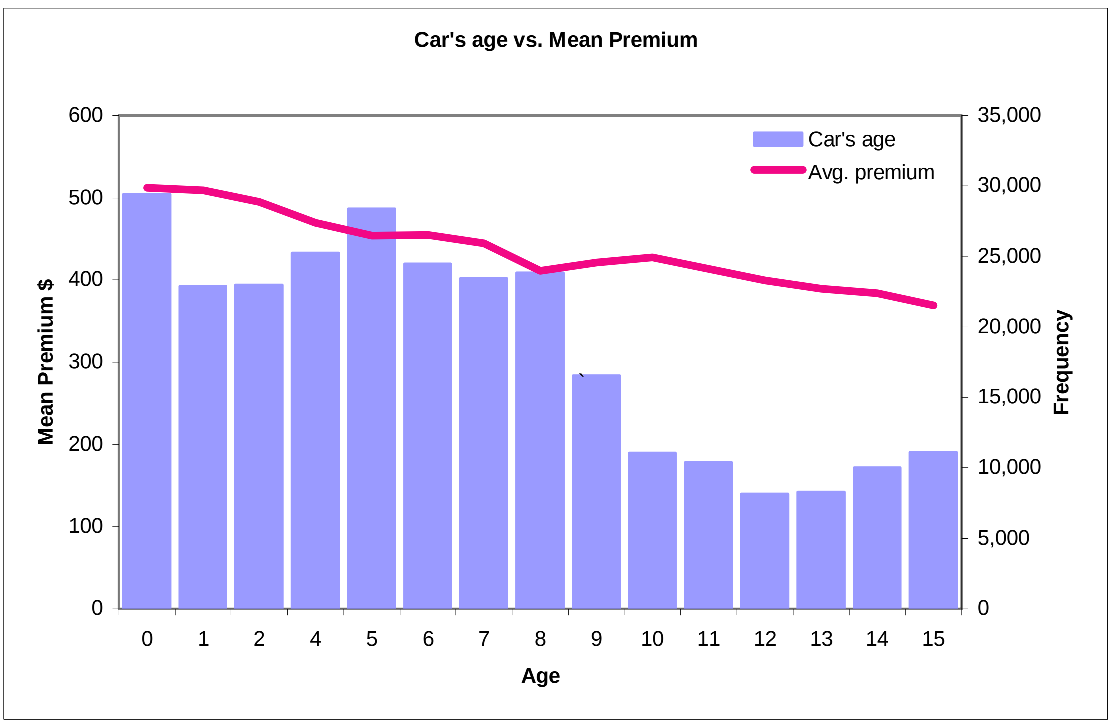

Figure 3-2 - Car's age vs Mean Premium (After)

The variable "Change" was created to show the percentage increase or decrease from the previous premium to the proposed premium. Because this variable is directly related to price sensitivity, the response we are trying to estimate, it seemed illogical to include it as an indicator variable. However, in the latter stages, this variable will help us determine the optimal increase/decrease that should be offered to each segmented group.

The team developed a frequency table, classifying each percentage change into bins. The width of each bin was 0.5%, and it started at -15%, ending at +15%. There was an abnormal amount of entries allocated in the bins -3% to -2.5% and -.5% to 0%. This abnormal frequency was dropping the renewal percentage, leaving two outliers in the distribution. Further analysis demonstrated that the renewal frequency was not the issue, but the number of cancellations was so large that it caused the renewal percentage to drop from approx 95% to values in the range of 45-50%. Figures 3.3 and 3.4 show the increment in frequency and the decrement in renewal percentage before and after the modification of the data.

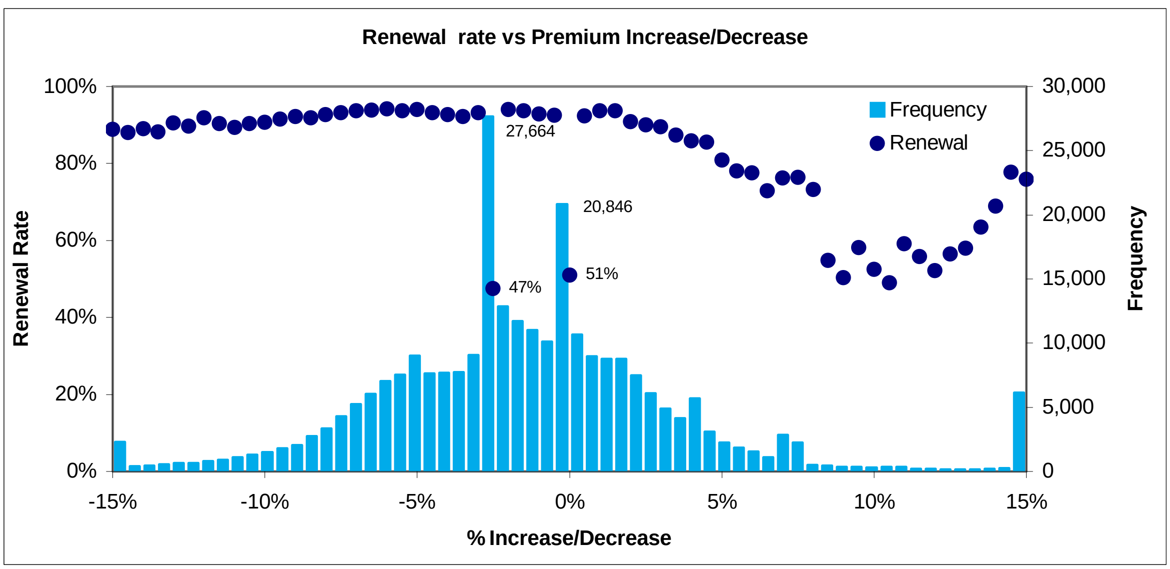

Figure 3-3 – Renewal Rate vs Premium Increase/Decrease (Before)

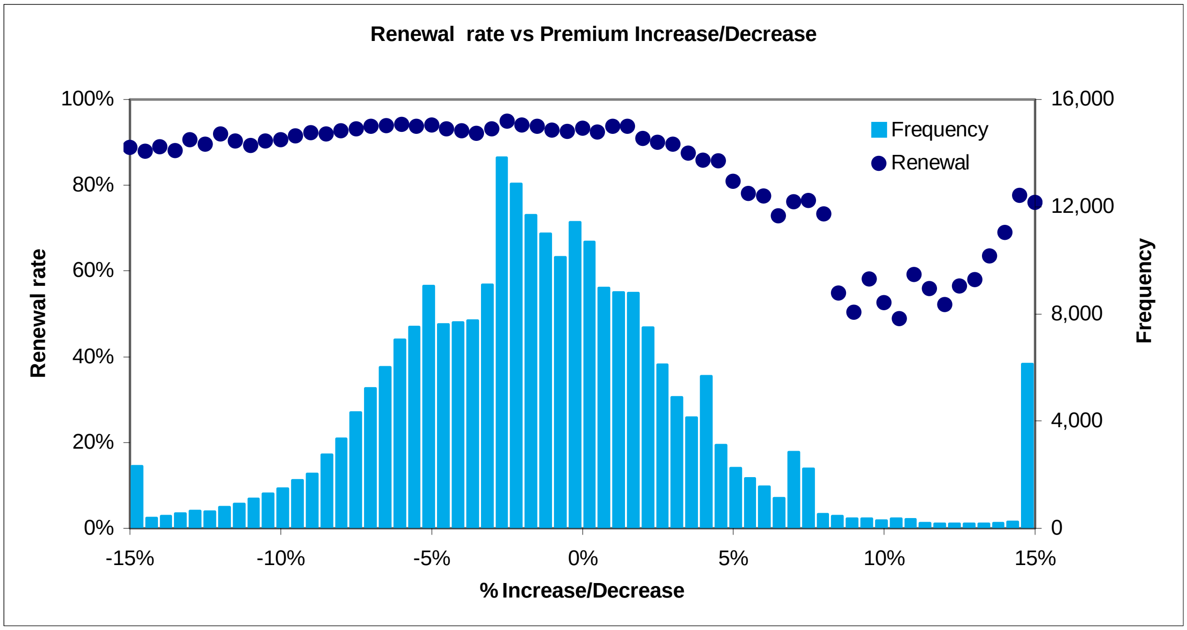

Figure 3-4 - Renewal Rate vs Premium Increase/Decrease (After)

Even before conducting any analysis, the data has to be checked and fixed for each erroneous entry. Once the data set is clean, and the variables that will be analyzed have no missing entries or default values, then the next step is to determine which variables will be useful for the clustering process. Below you will find a summarizing table of the deleted entries.

Description	**Deleted Entries**	**Total Entries**
Initial Data Set		760,233
- Newly signed	249,558	510,675
- Age<20	32,392	478,283
- Missing premio_ante & premio_post	2	478,281
- # accidents<0 & missing, merged 4+ into 4	72,826	405,455
-Auto_eta=3, merged 14&15 into 14, and 16+ into 15	101,509	327,722
- S and U for -2.5% & 0% (error)	51,334	**276,388**

Table 3-1 – Summary of Deletions

3.2 Customer Indicators – Variables to be used

Once the team eliminated all the errors, they proceeded to choose six variables to continue examining throughout the remainder of the project. It was agreed along with the sponsor liaison that the analysis would be done on only six variables, as it was a manageable due to time and computer power constraints. Initially the group studied eleven potential variables that were characteristic to the client's policies. These included the driver's general living area (rural or urban), the maximum policy coverage, last year's premium, the driver's profession, the driver's sex, the driver's age, the car's age, the number of accidents, the number of years the client has been subscribed with the company, the driver's driving record, and the type of fuel the car uses.

The criterion used to determine which of these variables were going to be used for the clustering and further analysis was based on how the renewal rate was impacted by the variable. In order for a variable to be chosen, the renewal rate needed to be significantly different for each of the categories it was subdivided into. For example, for sex, the difference in the renewal rate between males and females was the factor that was taken into account. At this stage some subjective analysis was also needed.

The first couple of variables that were discarded from the remainder of the study were sex and the type of fuel of the car. Even though subjective reasoning would say that neither sex nor the type of fuel are reasonable variables to impact the driver's renewal decision, they were still analyzed to confirm that in fact this was the trend. Additionally each of these two variables only has two categories, which would make the clustering unnecessary. Figures 3.5 and 3.6 show that the renewal rate for each of the categories in these two variables is not much different from each other.

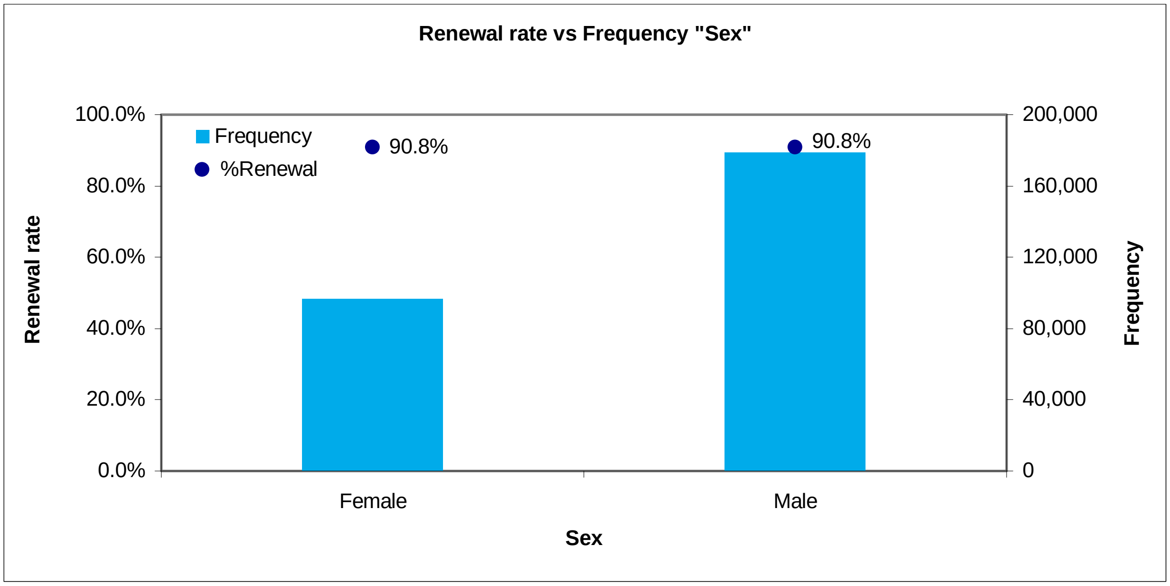

Figure 3-5 – Renewal Rate for the variable "Sex"

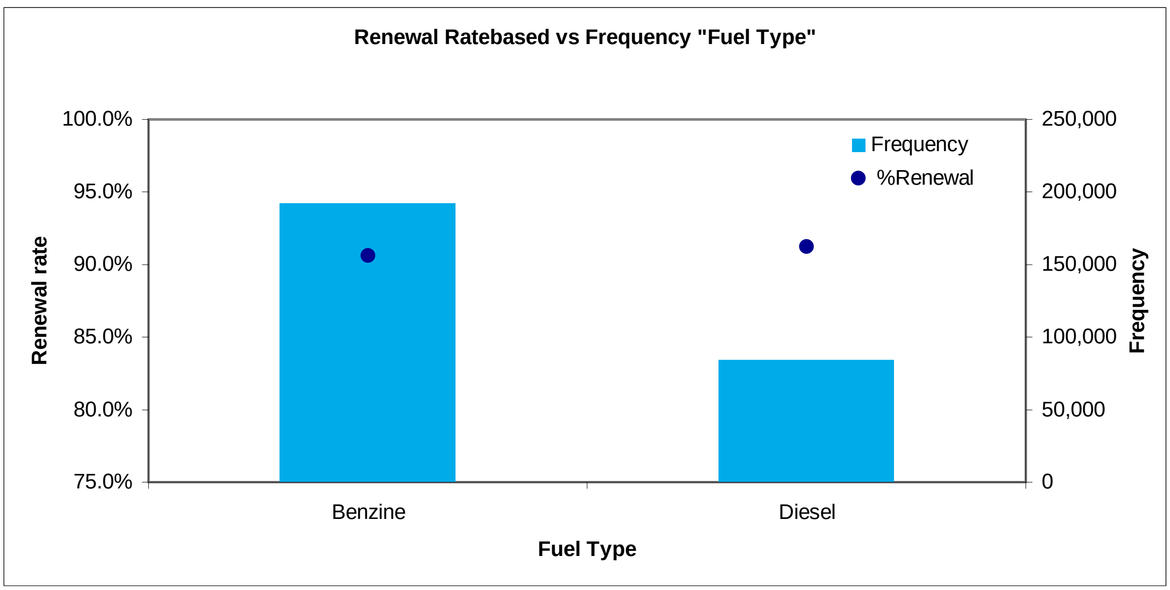

Figure 3-6 - Renewal Rate for the variable "Fuel Type"

Another variable that behaved similar to the above two was the region where the driver lived. The group wanted to see if the renewal rate was impacted by whether the insured lived in a rural or an urban area. Since the data showed there was not much difference in renewal percentages between these two categories, this variable was also disregarded from the remainder of the study. For a graphical representation of this variable's behavior refer to **Appendix A**.

The final two variables that we removed from the study were the driver's profession and the insured's driving record. These two variables required some modification from their original formats in the initial database. Unfortunately, due to time constraints and the fact that it was the first time the students used SAS, these variables had to be discarded as well. The driving record variable had a special format in the database that the group was unable to decipher in order to do some calculations with it. For profession however, the students were able to group the different job titles into 13 general categories, with similar sizes. Nevertheless since "profession" was categorized in "alpha-characters" groups rather than "numeric-characters", this presented a problem when trying to include this variable into the clustering. The reason being the system could not compute distances between "alpha-characters". For more information regarding analysis on "profession" and "insured's driving record" refer to **Appendix A**. After the initial analysis on the preliminary variables, six of them were finally chosen to continue the study. These included the maximum policy coverage, the previous year's premium, the driver's age, the car's age, the number of accidents, and the number of years the client has been signed with the insurance company. All of these six variables showed a more significant difference in renewal rate between categories. For example, Figure 3.7 shows this trend for the maximum policy coverage.

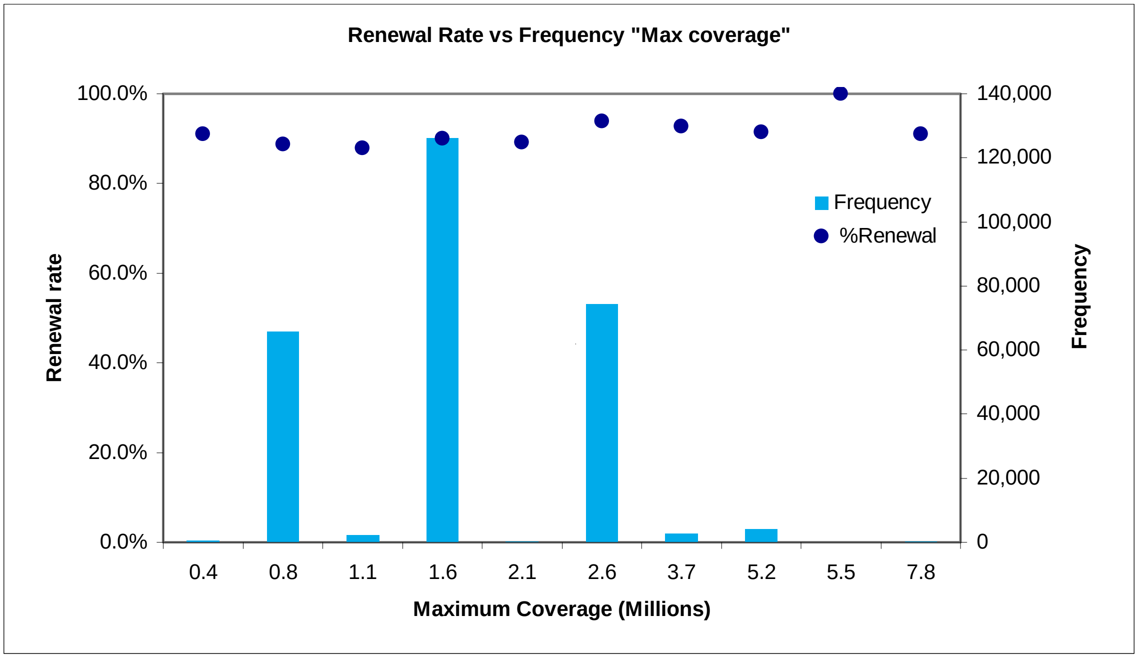

Figure 3-7 Renewal rate for the variable "Maximum Coverage"

For a similar analysis of the remaining five variables refer to **Appendix A**.

3.3 Customer Segmentation – Data Clustering

As most marketing courses or seminars have discussed, customer segmentation is "is the practice of dividing a customer base into groups of individuals that are similar in specific such as age, gender, interests, spending habits, and so on." (Marketing Textbook, 2007) Segmentation is an extremely important process by which companies attempt to target different customer groups effectively, in order to allocate marketing resources effectively.

Ideally, as Ian Turvill suggests in his article "Marketing: The New Policy for Insurers," insurance firms should treat each customer as an individual. Modeling or predicting customer behavior in real life, could be a very expensive process if the model treats each policyholder as an individual. The alternative is to identify similar characteristics among customers, and potentially merge customers who have very similar traits. Although this may sound simple, the mathematical algorithm behind it is very sophisticated.

The clustering algorithm is an iterative process by which the Euclidean distance between each pair of data points is computed. Specifically, this process is known as "K-means clustering." The k-means algorithm is an algorithm to cluster objects based on attributes into k partitions. The objective it tries to achieve is to minimize total intra-cluster variance, or, the squared error function:

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - \mu_i)^2$$

where there are $\boldsymbol{k}$ clusters $\boldsymbol{i = 1,2, \dots k}$ and $\boldsymbol{\mu_i}$ is the centroid or mean point of all the points $\boldsymbol{X_j}$. (SAS, 2007).

The team used the procedure "FASTCLUS" which is a feature of the statistical software SAS. This procedure applies the k-means algorithm to a specified data set, using specified variables by the user. As it was mentioned in the previous section, the variables considered for the clustering are:

Variable	Translation
Massim	Policy Coverage
Eta	Age
Auto Eta	Car's Age
Premio Ante	Previous Premium
Number of accidents	Number of accidents
Number of years insured	Number of years insured

Table 3-2 – Variables used in SAS

Before applying the FASTCLUS procedure to the data, it was imperative to standardize the variables in order to maintain a homogeneous range among them. For example, the variable "Massim" has a range of 400,000 to 7,800,000 units. If we compare this range to the range of Car's Age, which is 0-15 units, and the algorithm attempts to compute the variance of the distances between variables, the result will not only be incorrect but also huge. Therefore, the standardization of each variable was crucial in order to get accurate results. The variables were standardized using the "STANDARD" procedure in the SAS library.

The team decided to perform three different variable combinations to determine if a particular combination would produce significantly better results that another one. The following combinations were used:

Combination 1	**Combination 2**	**Combination 3**
1. Massim	1. Massim	1. Massim
2. Eta	2. Auto Eta	2. Premio Ante
3. Auto Eta	3. Premio Ante	3. Number of accidents
4. Premio Ante	4. Number of accidents	4. Number of years insured
5. Number of accidents	5. Number of years insured	
6. Number of years insured		

Table 3-3 – Clustering Combinations

The customer segmentation process allowed the team to differentiate customers using the different variables from the three combinations established. The next step in the process is to construct a model that would allow the team to estimate the demand for each cluster in the output file.

3.4 Demand Estimation

Demand estimation is the use of statistical techniques to determine how will demand be affected by a change in the price of a given product. The laws of demand and supply establish that, and increment in price will result in a decrement in the quantity demanded. The real question is, how sensitive is the demand for this product relative to price changes?

The *elasticity* of a product, or the *price elasticity of demand,* is an indicator of the sensitivity of a product's demand in relation to a product's price increase/decrease. In this study, if the insurance firm increases premiums to a certain policy group by a given percentage, how will this affect the renewal rate? The implementation of a mathematical model that will resemble the behavior of the demand for this study will help the team answer this question.

3.4.1 The Logistic Regression Model

As the background section mentions, the team decided that the model that would best simulate demand for this study would be the Logistic Regression Model. Logistic regression is commonly used when the response variable being modeled is binary. For instance, if your insurance firm increases the next year's premium, would you renew with this firm or not? Modeling this "answer" using a binary response variable, 1 for "Renewing" and 0 for "Not Renewing" is the objective pursued in the estimation of the demand.

The logistic regression model uses "predictor" variables (sex, age, profession.. etc) to model the "response" variable or the renewal decision. (Phillips, 2005) The model analyzes binomially distributed data of the form:

$$Y_i \sim B(p_i, n_i), \text{ for } i = 1, \ldots, m,$$

Where $\boldsymbol{Y_i}$ represents the response variable with $\boldsymbol{n}$ know Bernoulli trials with probability of success, in our case retention, $\boldsymbol{p_i}$ which is unknown. The combination of all predictor variables for each data point, results in the vector $\boldsymbol{X_i}$ which contains the corresponding numerical values of each predictor variable. With these tools in place, the model computes the probability of success for each data point as:

$$p_i = \mathrm{E}\left(\left.\frac{Y_i}{n_i}\right| X_i\right).$$

The equation above shows that the probability of renewal is equal to the expected number of successes (***Y***) divided by the number of trials (***n***), given the predictor variables grouped in the vector ***X***.

The logits of the unknown probabilities $\boldsymbol{p_i}$ are modeled as a linear function of $\boldsymbol{X_i}$:

$$\operatorname{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i}.$$

Where β_0, β_1,.. β_k are the estimated parameters of the logistic regression model.

Solving for $\boldsymbol{p_i}$ using the right hand equations, leads to the common solution of the logistic regression model:

$$p_i = \frac{1}{1+e^{-(\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i})}}.$$

Which can be expressed as:

$$y = \frac{1}{1+e^{-f(X)}}$$

Where $f(x)$ is the linear function:

$$f(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

The team used the LOGISTICREGRESSION procedure in the SAS library to perform the regression on each cluster based on the predictor variable “Change.” This variable represents the increase or decrease in premium offered by the insurance firm for each policyholder. This variable is the main predictor of the sensitivity of each customer given that customers have already been segmented based on the similarities in their characteristics. The model outputs two parameters, β_0, and β_1. The first parameter is the intercept of the linear function, and the second parameter is associated to the predictor variable "Change." With the model outputs, the team proceeded to construct the demand curve estimated by the linear regression model.

With the use of the clusters and the demand estimation curve, the students were able to predict the potential revenue points at each percentage change. The following steps involved optimizing the revenue by maximizing dollars and customer retention.

3.5 Revenue Optimization

Once the demand curves were fitted for each of the clusters in the three different clustering strategies, the team then proceeded to focus on the project's main goal. As mentioned in the Introduction, Price Revenue Optimization can be adapted to the specific goals of the company and in this case, the team focused in developing a strategy to maximize profit. The expected revenue was computed for each cluster in each of the three strategies, and compared to the original revenue for each of these same clusters. The original revenue was computed by separating the clients who renewed and those who didn't. Then, the average premium that was actually offered to those who renewed for each cluster was obtained from the initial dataset and multiplied by the number of people who renewed within the cluster. This essentially shows the revenue obtained from the company's original pricing strategy. The specific formula used to compute the original revenue was:

$$R_{total} = \sum_{c=1}^{n} R_c = N_c \times \mu_c$$

R_c = Revenue for each cluster

N_c = Number of observations that renewed in each cluster

μ_c = Mean Premium for each cluster

Table 3.4 summarizes the process for the calculations mentioned above.

Cluster	**Renewal Frequency**	**Mean Premium**
1	N_1	μ_1
2	N_2	μ_2
3	N_3	μ_3
.	.	.
.	.	.
.	.	.
n	N_n	μ_n
Total	$\sum N_c$	$R_{total} = \sum R_c$

Table 3-4 - Sample Cluster Summary

Following a similar approach, the team calculated the expected revenue for each of the clusters for each of their clustering strategies. One of the outputs calculated by the students with SAS, was the average previous year's premium for each of the clusters. This was obtained since one of the variables that were chosen to perform the segmentation, was the previous year's premium.

Using the estimated demand performed in the previous stage for each cluster, the authors were able to determine how much the premium needed to be increased or decreased to maximize the revenue. The number of people who would renew for each cluster was also directly related to the premium increase/decrease. By graphing the percentage increase/decrease in premium as the independent variable and the expected profit as the dependent variable, the students were able to determine what would be the optimal percentage change in premium to maximize revenue. The following graph is an example of the expected results:

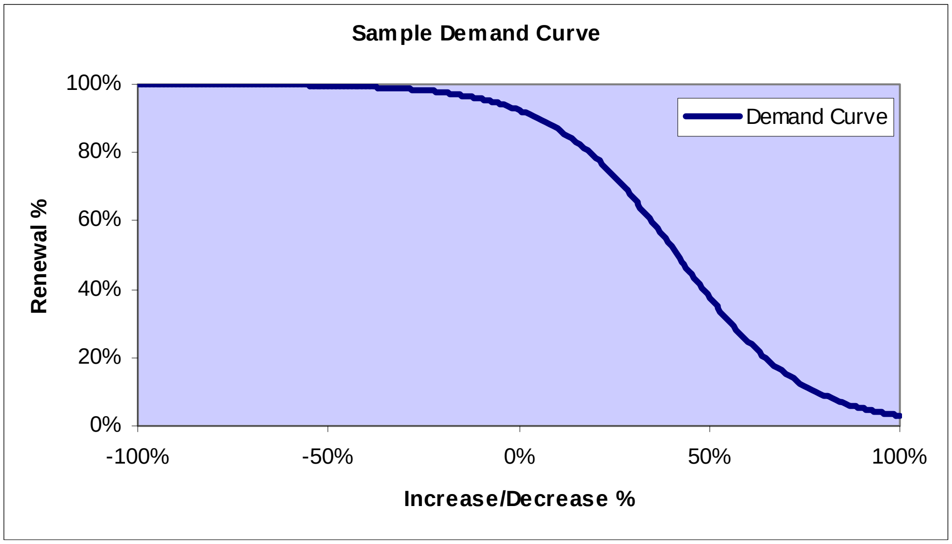

Figure 3-8 – Sample Demand Curve

$$R_{total} = \sum_{c=1}^{n} R_{\max} = (1 + \Delta_c\%) \times \rho_c \times \mu_c \times N_c$$

R_{max} = Maximized Revenue

%optimal = Optimal percentage change

Roptimal = Optimal retention percentage

μ_c = Mean Premium for each cluster

N_c = Number of total observations in each cluster

This process was repeated for each of the clustering strategies and compared to determine the best strategy. Table 3.5 summarizes the estimated number of clients who would have renewed for each cluster in each of the strategies and the average premium charged in each group to maximize revenue.

Technique 1-3		
Cluster	**Renewal Frequency**	**Max Revenue**
1	$N_1 \times \rho_1$	$\mu_1 \times \Delta_1$
2	$N_2 \times \rho_2$	$\mu_2 \times \Delta_2$
3	$N_3 \times \rho_3$	$\mu_3 \times \Delta_3$
.	.	.
.	.	.
n	N_n	μ_n
Total	$R_{total} = \sum_{c=1}^{n} R_{\max} = (1 + \Delta_c \%) \times \rho_c \times \mu_c \times N_c$	

Table 3-5 – Sample Revenue Summary

By completing the five stages outlined at the beginning of this chapter, the team of students was able to develop a general process by which a company can follow to develop an appropriate pricing strategy. Even though the data cleanup and the variable identification are tedious tasks, they are extremely necessary when working with real-world data. The customer segmentation (clustering) and demand estimation are the key steps to developing a demand-driven price optimization strategy. Finally, the Price and Revenue Optimization model should be adapted to company's goals, which in this case were maximizing revenue, but can also be applied for example to increasing market share. The next chapter will discuss the findings that provided for support in developing the recommendations and conclusions.

4 Findings and Results

This section outlines the main findings of the group after applying the model on a database of nearly one million entries for an Italian auto insurance. The first part discusses what variables where chosen and the criteria used to chose those variables out of many possibilities. Then, this section will show the results obtained after using the FASTCLUS command in SAS with six, five, and four variables respectively. It will finally compare the results obtained when applying price optimization to maximize revenues for each of the ten clusters.

4.1 Customer Indicators

The database given to the group had over ten variables to choose from, besides the ones that could be created with the given data. Thus, the group had to decide which and how many variables to use for the clustering procedure. In order to do this the group used to specific criteria:

1. Identified variables that had the greatest impact on retention rate. In other words, looked at those with the greatest variation in retention percentage at each frequency.

2. Searched for independence in renewal rates between categories within each variable.

After conducting the analysis described in the previous section, the team decided to use the following variables:

1. Car’s Age
2. Driver’s Age
3. Maximum Policy Coverage

4. Number of accidents
5. Previous Premium
6. Years Insured

4.2 Customer Segmentation – Data Clustering

The next step was to segment the market or perform the clustering technique. In a business model, customer segmentation is the practice of dividing a customer base into groups of individuals that share similar characteristics. In a mathematical context, this procedure is called K-Means Clustering, which groups observations by computing the smallest Euclidean distance between each one of them. The team used the three clustering techniques discussed in the previous chapter in order to be able to compare the results and the impact of using different number of variables. The three techniques produced the following results:

Technique 1									
Cluster	**Renewal**	**Cancel**	**Total**	**Premium**	**Coverage**	**Age**	**Acc**	**Yrs**	**Age (c)**
1	39937	2418	42355	295	1,696,123	69	0	1	10
2	3987	352	4339	401	4,723,230	52	0	1	6
3	46573	4385	50958	330	2,089,283	52	0	1	3
4	24179	3876	28055	447	1,666,005	56	2	2	7
5	3985	473	4458	336	1,449,659	43	0	2	10
6	30416	4106	34522	593	1,656,216	41	0	2	5
7	6424	1385	7809	331	1,763,391	60	1	1	6
8	7489	607	8096	307	1,174,040	59	0	7	10
9	34152	1846	35998	991	1,864,934	46	1	1	6
10	53823	5975	59798	529	1,718,117	42	1	2	6
Total	**250965**	**25423**	**276388**						

Table 4-1 – Cluster Summary for Technique 1

Technique 2								
Cluster	**Renewal**	**Cancel**	**Total**	**Premium**	**Coverage**	**Age**	**Acc**	**Yrs**
1	39236	3919	43155	351	2,659,757	49	0	1
2	6130	495	6625	414	5,261,905	53	0	1
3	53769	3499	57268	297	1,655,067	70	0	1
4	10101	892	10993	622	1,587,183	41	0	2
5	42550	5179	47729	705	1,854,330	47	2	1
6	26212	3911	30123	307	1,185,479	58	0	7
7	22671	1045	23716	950	1,904,047	46	0	1
8	28862	3966	32828	366	1,590,353	58	2	2
9	5617	1350	6967	376	1,745,125	53	1	1
10	15817	1089	16906	334	1,336,690	44	0	2
Total	**250965**	**25345**	**276310**					

Table 4-2 - Cluster Summary for Technique 2

Technique 3							
Cluster	**Renewal**	**Cancel**	**Total**	**Premium**	**Coverage**	**Acc**	**Yrs**
1	3683	370	4053	370	1,137,552	2	4
2	6151	509	6660	399	4,716,789	0	1
3	27165	3708	30873	585	1,962,712	0	1
4	33059	4543	37602	313	1,208,388	0	7
5	30594	1552	32146	334	1,798,703	1	1
6	5322	1469	6791	991	1,790,429	0	1
7	7409	624	8033	511	1,964,895	2	1
8	49626	2775	52401	287	2,407,785	0	1
9	11558	974	12532	624	1,463,964	1	2
10	76398	8820	85218	334	1,308,639	0	2
Total	**250965**	**25344**	**276309**				

Table 4-3 - Cluster Summary for Technique 3

4.3 Demand Estimation

The group used the Logistic Model to estimate the demand curve for each of the clusters. Renewal probability was used as the response variable, while price change and cluster where our predictor variables. All ten curves were estimated, obtaining different shapes and results. The main reason for this was that data was not spread equally over the entire range. In other words, not many customers were offered a large premium increase or decrease so there are very few data points at the ends and many towards the center of the graphs.

Figures 4.1 and 4.2 show the demand curves for Clusters 1 and 6 using Technique 1. These curves are good examples of how customers would react to specific changes in their policy renewals. They are good examples of a good S-shaped curve because they reflect customer’s high sensitivity close to 0% and insensitivity towards extreme changes in policy premiums.

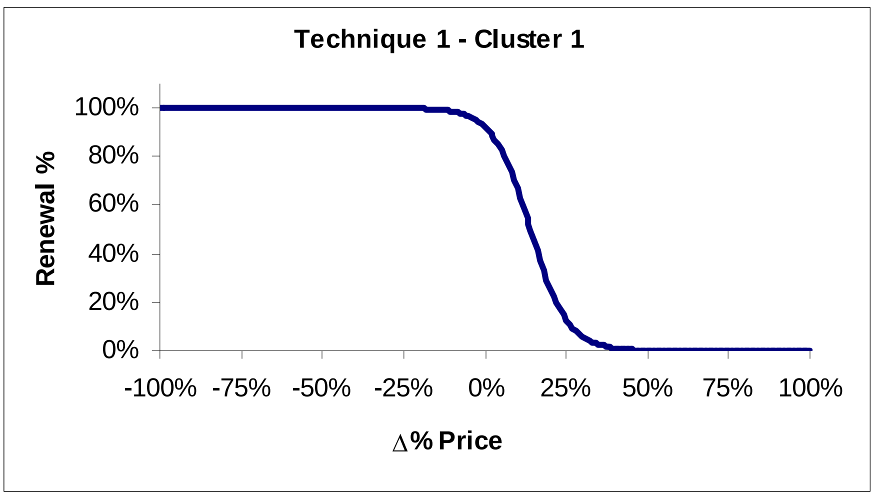

Figure 4-1 – Demand Curve for Technique 1- Cluster 1

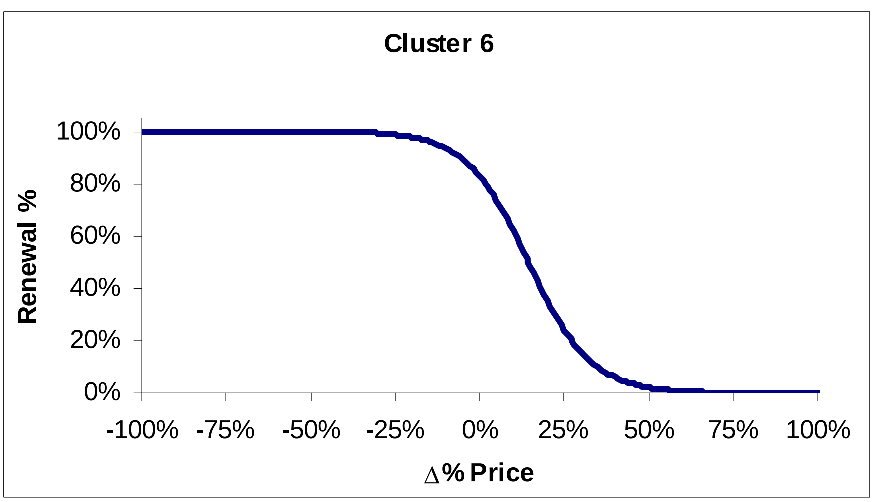

Figure 4-2 - Demand Curve for Technique 1- Cluster 6

In contrast, clusters 2 and 10 especially show very insensitive populations which would seem very odd in real life. Mathematically this is how the model predicts that these two sets of customers would react, but in reality this reveals the lack of data mentioned previously.

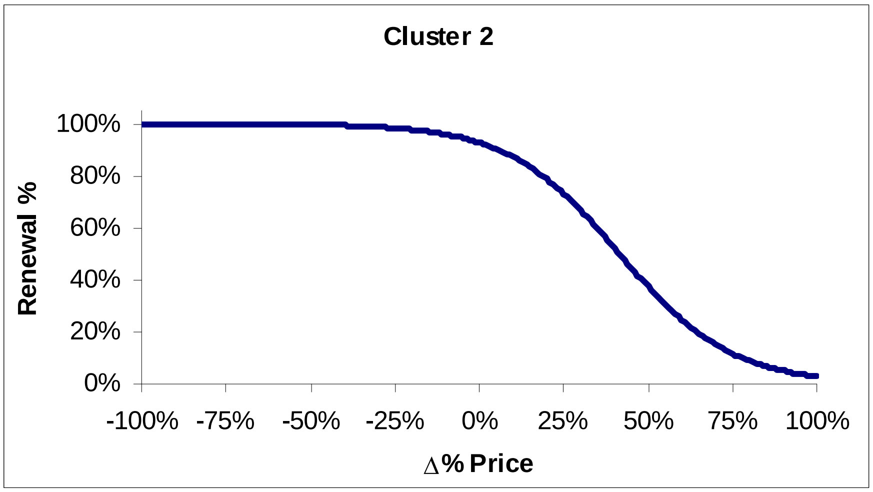

Figure 4-3 - Demand Curve for Technique 1- Cluster 2

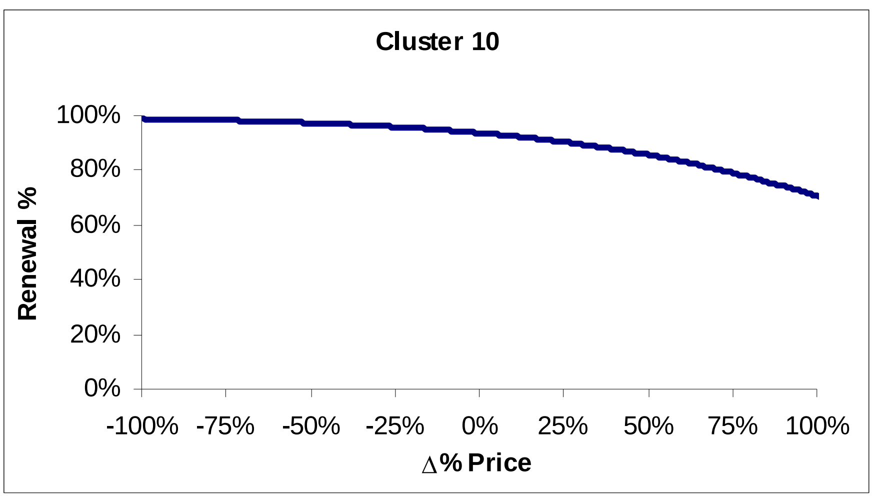

Figure 4-4 - Demand Curve for Technique 1- Cluster 10

4.4 Revenue Optimization

The last step in the model was finding the optimal percentage change in premium to ensure the maximum possible revenue per cluster. Then it was just a matter of adding the optimal revenues for each of the ten clusters to obtain the overall optimal revenue for each of the three techniques that the group used. The results for Technique 1, 2, and 3 are as follows:

Technique 1						
	Actual			**Prediction**		
Cluster	Δ	ρ	**Revenue**	Δ	ρ	**Revenue**
1	-1.6%	90.9%	11,406,151	-2.0%	94.2%	11,767,829
2	-2.5%	92.5%	2,398,598	12.0%	85.7%	2,551,566
3	-1.5%	93.9%	17,491,447	0.0%	94.3%	17,828,537
4	4.1%	91.9%	4,696,954	123.0%	68.4%	7,489,889
5	-2.8%	89.1%	13,873,691	-3.0%	91.8%	14,255,718
6	-4.9%	87.0%	14,773,160	-6.0%	90.1%	15,124,401
7	0.5%	95.6%	7,531,396	30.0%	84.3%	8,599,116
8	-1.3%	87.9%	8,745,360	-3.0%	92.7%	9,059,811
9	-5.5%	80.6%	5,258,937	72.0%	59.0%	7,008,391
10	-0.9%	93.6%	8,295,992	96.0%	72.0%	12,615,683
Total			**94,471,686**			**106,300,939**

Figure 4-5 – Revenue Optimization Summary for Technique 1

Technique 2						
	Actual			**Prediction**		
Cluster	Δ	ρ	**Revenue**	Δ	ρ	**Revenue**
1	-2.2%	94.3%	13,728,343	2.0%	91.9%	14,008,979
2	-2.3%	91.9%	1,611,922	11.0%	85.4%	1,702,320
3	-1.3%	91.4%	13,648,888	-1.5%	94.2%	14,037,265
4	-4.6%	86.2%	14,343,547	2.7%	81.9%	14,676,487
5	3.2%	89.4%	2,897,145	281.5%	59.1%	7,081,078
6	-1.3%	88.1%	9,225,146	-2.8%	92.7%	9,563,808
7	-6.1%	82.3%	5,732,147	45.8%	64.3%	6,955,806
8	4.3%	92.5%	2,861,423	80.3%	73.2%	3,912,444
9	0.0%	94.9%	12,834,010	42.7%	80.7%	15,573,562
10	-2.2%	90.0%	17,589,113	-3.2%	93.8%	18,148,376
Total			**94,471,686**			**105,660,125**

Figure 4-6 - Revenue Optimization Summary for Technique 2

Technique 3						
	Actual			Prediction		
Cluster	Δ	ρ	Revenue	Δ	ρ	Revenue
1	3.5%	90.9%	1,411,008	71.0%	72.2%	1,851,131
2	-2.5%	92.4%	2,393,657	12.0%	85.3%	2,538,519
3	-4.4%	88.0%	15,193,645	-4.0%	89.0%	15,437,237
4	-1.7%	87.9%	10,173,669	-3.0%	92.4%	10,551,243
5	0.6%	95.2%	10,287,972	37.0%	99.0%	12,093,693
6	-5.8%	78.4%	4,965,741	27.0%	64.7%	5,527,100
7	3.9%	92.2%	3,931,448	100.0%	81.6%	6,693,713
8	-1.1%	94.7%	14,057,483	2.0%	93.3%	14,284,403
9	-2.1%	92.2%	7,059,951	79.0%	71.3%	9,977,449
10	-2.0%	89.7%	24,997,113	-3.0%	94.1%	25,981,376
Total			**94,471,686**			**104,935,864**

Figure 4-7 - Revenue Optimization Summary for Technique 3

The next section will discuss the team’s conclusions on the project and the recommendations for areas on improvement.

5 Conclusions and Recommendations

After completing the analysis on the data from an Italian insurer, the group of students was able to develop a general process by which and insurance company can implement Price and Revenue Optimization. The first step was to clean up the data and decide which variables to use, then the students were able to determine a suitable method for clustering the data. By using three different clustering techniques, which were differentiated by the number of variables that were used, it became evident that a better clustering technique was that with a greater amount of variables. This conclusion came from analyzing the different optimal portfolios for each technique and choosing the one that maximized revenue. Nevertheless, in order to calculate the optimal percentage change in premium to maximize revenue, the group had to perform demand estimation using logistic regression first, for each of the ten clusters. Then, the students managed to conclude how sensitive each cluster was to a small or large premium change based on their renewal rate. This is information is useful for an insurer to know how each of its different types of customers react to a different premium change, and thus price its products accordingly for each group of people.

There are some issues that need to be considered when performing such a study with real-world data. First of all, the data clean up is a very important step as there might be several potential errors in the data, and it is imperative that the study is representative of the population that is being worked with. In this case, even though it reduced the amount of observations that were being analyzed to less than half, the students felt it was still significant to continue the study with this reduced number. Second, one should take into consideration that this project provides conclusions based on a mathematical model that may not necessarily be optimal in the real world. For example, in some clusters, the mathematical model showed that the most favorable (revenue wise) percentage change for that group of people was a 30% increase or more. Even though such an increase in premium would substantially reduce the market share for that group, a large premium would produce larger revenue than the others. Nevertheless, one needs to take into account that in the real world such a large increase in premium will probably not provide the results one was looking for. Also, the mathematical model in this project shows that for certain negative percentages changes (i.e. -100%), there will be a 100% retention rate. However, this will never happen in the real world, since there will be people who will cancel their policies no matter what they are offered (i.e. if they sell their car). Therefore, it is imperative that the model be adapted to the actual market that is being work with, and not just implement the mathematical model.

The purpose of this project was to develop a general model by which an insurance company could adapt to PRO and remain competitive in the field. Nevertheless, the study did not take into account several additional complexities that exist in the real world (due to time constraints), which should definitely be considered when expanding this process. For example, the students did not consider how the pricing of products would be impacted by the competitor's actions. It is very probable that in the real world, changes in policy prices would entice a reaction from other insurance companies, which most definitely be trying to adapt to this new competition. Thus, these responses from the rest of the market can be anticipated and included into a deeper analysis of this process. Aside from the competition, it is also very important to consider the economic cycles within the industry as well as the specific characteristics of the population that is being worked with. Also, for practical purposes, the research group only worked with six variables in the clustering and grouped the observations only in ten clusters. Nevertheless, as the study proves, a greater the number of variables used for clustering and a greater the number of clusters (in other words, a more complex clustering strategy) will produce a more precise pricing strategy based on the customer's demand; therefore providing a bigger margin to increase revenue.

6 Appendix

7 References

Balcombe, K, A Bailey, A Chalak, and I Fraser. "Bayesian Estimation of Willingness to Pay Where Respondance Mis-Report Their Preferences." Oxford Bulletin of Economics and Statistics (2007). JSTOR. WPI. Oct.-Nov. 2007.

Barone, Guglielmo, and Mariano Bella. "Price Elasticity Based Customer Segmentation in the Italian Auto Insurance Market." Journal of Targeting, Measurement and Analysis for Marketing 13 (2004): 21-31. JSTOR. WPI. 1 Oct.-Nov. 2007.

Boucek, Charles H., and Thomas E. Conway. "Dynamic Pricing Analysis." JSTOR. WPI. 1 Oct.-Nov. 2007.

Brumelle, Shelby, and Darius Walczak. "Dynamic Airline Revenue Management with Multiple Semi-Markoff Demand." Journal of Operations Research 51 (2003): 137-148. JSTOR. WPI. Oct.-Nov. 2007.

Cramer, J S. Logit Models From Economics and Other Fields. 1st ed. Cambridge: Cambridge UP, 2003. Oct.-Nov. 2007.

Crozet, Jb. "Revenue Management and Insurance Cycle." GIRO (2005). JSTOR. WPI. 1 Oct.-Nov. 2007.

Krikler, Samuel, Dan Dolberger, and Jacob Eckel. "Method and Tools for Insurance Price and Revenue Optimization." Journal of Financial Services Marketing 9 (2004): 68-79. JSTOR. WPI. 1 Oct.-Nov. 2007.

Moridaira, Soichiro, Jorge L. Urrutia, and Robert C. Witt. "The Equilibrium Insurance Price and Underwriting Return in a Capital Market Setting." The Journal of Risk and Insurance 59 (1992): 291-300. JSTOR. WPI. Oct.-Nov. 2007.

Murphy, Karl P., Michael J. Brockman, and Peter K. Lee. "Using Generalized Linear Modules to Build Dynamic Pricing Systems." JSTOR. WPI. 1 Oct.-Nov. 2007.

Orlay, Peter, Bruce Davey, and Jeremy Howard. "Setting Optimize Prices." K Forum Fall 2004: 15-28. JSTOR. WPI. 1 Oct.-Nov. 2007.

Philips, Robert. Pricing and Revenue Optimization. United States: Stanford UP, 2005. 1-384.

Price Optimization: a Potent Weapon for Inovative Insurers. Towers Perrin. Towers Perrin, 2007. 1-16.

Ridgeway, Greg. "Maximum Likelihood and Logistic Regression." (2004). JSTOR. WPI. Oct.-Nov. 2007.

Rodriguez, G. "Logit Models for Binary Data." (2007). JSTOR. WPI. Oct.-Nov. 2007.

Schlesinger, Harris. "Non Linear Pricing Strategies for Competitive and Monopolistic Insurance Markets." The Journal of Risk and Insurance 50 (1983): 61-83. JSTOR. WPI. Oct.-Nov. 2007.

Silverstone, H. "Estimating the Logistic Curve." Journal of the American Statistical Association 52 (1957): 567-577. JSTOR. WPI. Oct.-Nov. 2007.

Tsanakas, Andreas, and Evangelia Desli. "Measurement and Pricing of Risk in Insurance Markets." Journal of Risk Analysis 25 (2005). JSTOR. WPI. Oct.-Nov. 2007.

Tsiatis, Anastasios A. "A Note on a Goodness of Fit Test for the Logistic Regression Model." Biometrika 67 (1980): 250-251. JSTOR. WPI. Oct.-Nov. 2007.

Yen, Wendy M. "Obtaining Maximum Likelihood Trait Estimates." The Journal of Educational Measurement 21 (1984): 93-111. JSTOR. WPI. Oct.-Nov. 2007.