|  |
| --- |
|  |
| Preliminary Project Report |
| All State Insurance |
|  |
|  |
| **11/15/2014** |

|  |
| --- |
| *Abstract: Allstate Insurance is looking to predict the propensity of a customer to purchase a given quote.* |

**Team Members:**

Ashish Mehta

Gilad Mail

Nikita Kumar

Priyanka Kapoor

Shravan Ravi

Vishal Kotecha

Table of Contents

[Executive Summary 2](#_Toc403763017)

[1. Key Summary Statistics 3](#_Toc403763018)

[2. ETL – Extract, Transform, Load and Enrichment of the data 4](#_Toc403763019)

[3. Correlation between Quantitative Predictors 4](#_Toc403763020)

[4. Primary Hypothesis & Research Question 5](#_Toc403763021)

[5. Tools being used to study the research question 5](#_Toc403763022)

[6. Preliminary Models & Key Findings 5](#_Toc403763023)

[Appendix 6](#_Toc403763024)

[Preliminary Model 6](#_Toc403763025)

## Executive Summary

Car insurance is, and has been for a long time, a lucrative business. In many states it is the law to carry car insurance if you are a driver, and that pays off to insurers. Although the premise behind having car insurance is to be conservative and be protected, the marketing budgets are anything but conservative. And for good reason – the industry is nearing $200 billion! Accordingly, companies are competing for market share by retaining, and more importantly, acquiring new customers. In 2013, 45% of those shoppers who were in the market for car insurance ultimately switched insurers; this is the highest rate seen since such data began being recorded. This means that, with the right mix of marketing deployment, an insurer may be able to poach a customer away from a competitor, thereby gaining both additional revenue and market share! To accomplish that, insurers spent over $4 billion in marketing alone.

Allstate Corporation is among the largest insurers in the United States, with about 10% market share, and a large marketing budget to suit. Although TV ads and online advertising are a larger majority of the marketing budget allocation, direct-mail is still a multimillion dollar activity that is used to lure potential new (or returning past) customers. The idea is that marketing material such as colorful brochures or letters with special offers, are sent to people’s homes in an effort to spark interest in becoming a car insurance customer. If they become a new customer, be it whether they currently do not have insurance or are switching from another company, Allstate stands to make $650 in revenue.; if they ignore the material, Allstate loses $5.50. These values are assumptions made based on average insurance price based on the sample data, and approximations in overhead (hiring marketing employees) and mailing costs. While the potential gain from an insurance sale is much larger than the associated costs, sending too much material to the incorrect target audience may prove to be an ultimate profit loss.

The premise of the analysis that this report tries to show is that using a known customer database, Allstate can better target its direct-mail recipients in hopes of attaining more customers than with a “blind” deployment. The sample data that we found was very large, comprising purchase history of people from every state in the US. We want to lessen the complexity and, therefore, shortened the data set to just residents of New York State.

In New York, there are about 12 million registered drivers that are potential targets. Our assumption is that the marketing budget for direct-mail is 13000 of those people. The data that we gathered displays information about people who received direct-mail marketing material for car insurance, and their decision whether to buy. This data will be used to learn the propensity to buy for the different categories/predictors, in effort to narrow down the targets to 13000 people. Then we will compare the lift we can achieve using domain expertise over simple sending randomly to any person.

The characteristics include shopping point, homeowner, car age, risk factor, oldest person in the household, youngest person in the household, whether the household is of a married couple, the duration of the insurance held previously, and whether the person ultimately purchased the new insurance due to the direct-mail offer. These predictors will be explained further in another section of this report.

To try and be consistent with the data sample size used for the training and testing models in the class mini cases, we used a similar fraction for the 13000 targets, making it 1,000 for training and 1,000 for testing.

## 1. Key Summary Statistics

|  |  |
| --- | --- |
| Main Variables | Description |
| Homeowner | 0 = No, 1 = Yes   (whether person owns home or not) |
| Car Age | Age of the person’s car |
| Car Value | How valuable was the person’s car when new  (based on   some internal identifier that Allstate has) |
| Married Couple | 0 = No, 1 = Yes  (whether person’s group contain a married couple) |
| Duration Previous | How long (years) person was covered by previous insurer |
| Success | 0 = No Purchase, 1 = Purchase Insurance |

The rest of the variables available in the data set are:

* Day
* Time
* Group Size (how many people will be covered under the policy)
* Risk Factor (an ordinal assessment of how risky the person is)
* Age Oldest (age of the oldest in the person’s group)
* Age Youngest (age of the youngest in the person’s group)

Homeowner: this is the most evenly distributed variable, with an almost 60/40 split in the sample between homeowners and non-homeowners, respectively for training. This seems to be a good variable to use, as our domain expertise says that married couples are more conservative and looking for better deals, and willing to decrease their financial costs.

Car Age: this variable has some outliers, but seems to be relatively even in distribution. This variable seemed to make sense to use because the older someone’s car, the less they would tend to want to spend on car insurance, and would look for a better alternative offer. There were a few outliers for much older cars.

Married Couple: as with homeowners, married couples are more likely to want to save on their monthly costs, and therefore are more inclined to shop for lower insurance options. Singles, on the other hand, are more inclined to pass on looking at direct-mail offers and stick with their current coverage. However, the distribution is a little skewed, with groups containing non-married couples comprising three-quarters representation in the sample for training.

Duration Previous: this variable is evenly distributed and domain expertise says that the longer time that a person has had car coverage with their previous insurer, the less likely they are to switch since they are probably already happy with their insurance.

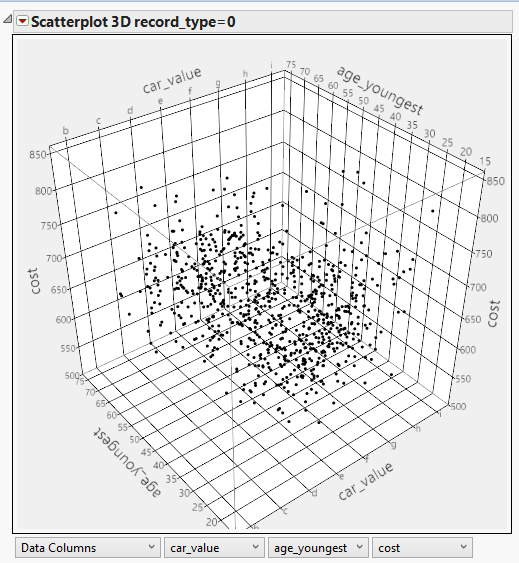
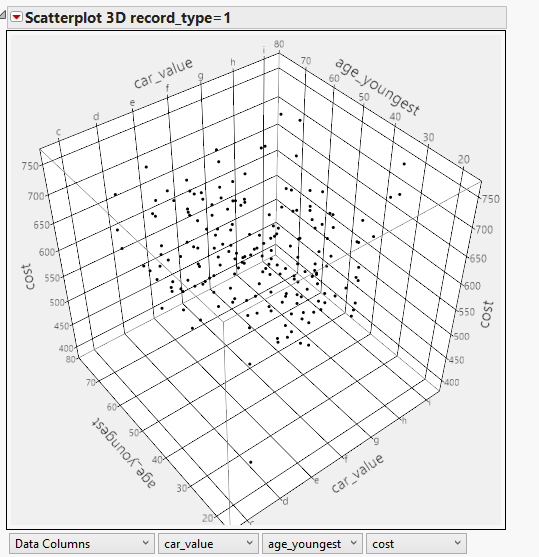
## 2. ETL – Extract, Transform, Load and Enrichment of the data

The data was obtained from Kaggle.com, which is a platform for predictive modelling, and analytics competitions on which companies and researchers post their data and statisticians and data miners from all over the world compete to produce the best models. The data that we found was for a competition hosted by Allstate Insurance to predict the propensity of a customer to purchase a given quote.

The original dataset was used to predict which car insurance policy a customer would end up choosing from amongst various options offered. Every customer was offered an average of 6 quotes from which he would end up choosing one quote. Thus, each tuple in the original dataset was a quote. This was not well suited to our problem as we are dealing with predicting whether a customer would purchase a quote, rather than predicting which quote the customer would purchase. To modify the data to suit our problem, we considered only one tuple belonging to each customer id. Thus, every tuple in our new dataset represented a unique customer. The initial data included purchase history for customers from all states in the United States. To reduce the complexity of the problem, we considered only residents from NY State. To achieve this, we filtered the initial data to only include information pertaining to NY residents.

Further, we excluded certain variables, which did not contribute towards the model. These variables were decided using a combination of tools like screening, scatter plots and histograms with domain knowledge.

## 3. Correlation between Quantitative Predictors



Our project makes use of regression techniques to predict the response variable using a combination of the predictor variables. Regression analysis is a statistical process for estimating the relationship among variables [1]. Since we had to do predict a variable, we thought it was essential to understand the correlation between the various predictors. We used a scatter plot to analyze the relation among the significant variables, namely car\_value, age\_youngest & cost. These variables were selected as significant because they each had a p-value of less than 0.001 which was safely below the threshold of 5%.

Upon inspection of the scatter plot for record\_type = 1, it is observed that a majority of the data points are scattered around the region where the values of the 3 significant variables are:

1. car\_value - between e and f
2. age\_youngest - uniformly distributed between 30 and 55 years
3. cost - between 625 and 700

Car\_value e and f on a scale of c to I indicate that the customer’s car was moderately priced when purchased; cost between 625-700 on a scale of 400 to 750 indicate that the cost of the insurance was quite high. Most of the people purchasing the insurance were people between the ages of 30 and 55 years, which could possibly indicate that people who were of considerable working age and who might be married are most likely to purchase the insurance.

## 4. Primary Hypothesis & Research Question

Null Hypothesis: the selected set of variables is not useful to predict the response variable (Success).

Alternate Hypothesis: the selected set of variables is useful to predict the response variable.

In this project, our goal is to predict the propensity of a customer to purchase a given quote. This is thus a **Prediction** problem.At the same time, we are dividing the customers into buyers and non-buyers. Hence, the problem can be viewed as a Classification based approach too.

## 5. Tools being used to study the research question

Various classification techniques were used to study the primary research question and business acumen was leveraged along with domain expertise to come up with viable solutions for the questions being answered. In total there are three classification techniques; decision tree, logistic regression and neural networks. We plan to use these different methods to carry out the data-mining task. We will then be deciding on a particular model, based on how well it fits our data and based on the quality of results obtained. To begin with, to find the propensity of a customer to purchase a quote, we primarily use Logistic Regression and complement it with Decision Trees to incorporate our domain expertise into the model.

## 6. Preliminary Models & Key Findings

We performed logistic regression on our model since we found it as a better fit for our dataset. From our domain expertise, we know that this method will give us a better model since it accounts for the variables both statistically and qualitatively. We found a couple of variables to be statistically significant. Day and duration\_previous were the variables that gave a low chi square value which meant that they were significant to our model.

We plan to improve our model by increasing the R-square value that we observed which was around 0.305. We also plan to enrich some variables by examining them a bit more deeply and maybe recoding some of the variables to be qualitative in nature so we can multiply them with quantitative variables to make business sense out of the new variables. Furthermore, we might even look at how it compares to have insurance in rural areas versus urban and suburban areas. Since most of the residents of New York City will not have cars, this relation will be very interesting since the market in that city will be limited to a select few especially to specialty categories for example tour operators, black car/limousine services and such.

## Appendix

### Preliminary Model

