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 2 million 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 2 million 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 (2000 tuples used for 100,000 targets), we used a similar fraction for the 2 million targets, making it 25,000 for training and 25,000 for testing.

4. **Hypotheses**:

**Null Hypothesis:** the selected set of variables isn’t 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.

xbuyers and non-buyers. Hence, it can be viewed as a Classification based approach too.

5. We plan to use 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.

2.

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 wasn’t 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 didn’t contribute towards the model. These variables were decided using a combination of tools like screening, scatter plots and histograms with domain knowledge.