Allstate Purchase Prediction

DSO 528: Final Project Submission

12/08/2014

Team Members:
Ashish Mehta
Gilad Mail
Nikita Kumar
Priyanka Kapoor
Shravan Ravi
Vishal Kotecha

Abstract: As a customer shops an insurance policy, he/she will receive a number of quotes with different coverage options before purchasing a plan. This is represented in this problem as a series of rows containing demographic and psychographic data of the customer. The task is to predict the probability of a customer buying the coverage options using a limited or enhanced subset of the total interaction history. If the eventual customer can be predicted sooner in the shopping window, the quoting process is shortened and the issuer is less likely to lose the customer's business.

Table of Contents

Executive Summary....2
Introduction....3
Problem Statement....4
Procedure....4
ETL – Extract Transform Load Enrich....5
Key Summary Statistics....7
Analysis....8
Second Best Model – Logistic Regression....8
Best Model – Logistic Regression....9
Performance Measure....10
For Second Best Model....10
For Our Best Model....11
Business Insights....11
Improvements....12
Conclusion....13
Appendix....14
1. Second Best Model:....14
2. Best Model:....17

Executive Summary

In addition to other forms to marketing, Allstate insurance spends millions in direct-mail advertising in attempt to gain new or "converted" customers. With a limited marketing budget, targeting the recipients randomly may not be the most effective way of ensuring likelihood of attaining the maximum potential customers, and the expenses/costs incurred may outweigh the profits gained.

The intent of this exercise was to produce a statistical model based on available data from past customers, in order to better predict which type of person will be a more likely potential customer via the direct-mail marketing strategy. The data set included various personal information for each candidate, which we used to narrow down to several key variables.

To find the best model, meaning largest predicted new customers and profit, several iterations of tools such as decision trees, regression analysis and variable distribution plot were created to determine the variables that have the greatest influence on the outcome. In addition, we used as much domain expertise as possible in our collective capabilities to further refine our models and predictors used. Those variables, in no particular order, are:

Variable	Type
Home Owner	Nominal
A,D,G	Continuous
Car_age	Continuous
Married_couple	Nominal
Has_kids	Nominal
Divorced/Wedlock	Nominal
Cost	Continuous
Compare_C	Nominal

This report will go in depth about the models used and the meaning of the results in order to convey the reasons used to decide on the key variables. We will also elaborate on some of the key insights that we discovered while working with the data set.

Overall, with a marketing budget allowing mailings to 10000 potential candidates (our team's assumption), we were able to show that using the model, we can predict a $753,646 profit on 10,080 new customers.

Introduction

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 of the insurance companies are anything but conservative. And for good reason – the industry is nearly $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 was 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 **$4.50**. 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 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, we siphoned our data and selected the Car Value E (as this subset of data was more promising and accurate).

Problem Statement

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. We want to target those customers who will have a proclivity to buy our insurance.

We got the dataset from Kaggle and modified the dataset to suit the above problem statement since the real world dataset has lot of entropy. We used decision tree and logistic regression to get our best model which we benchmarked against the best model provided by JMP.

Procedure

Null Hypothesis: there is no relationship between the selected set of variables and the dependent variable "Success".

Alternate Hypothesis: there is a relationship between the selected set of variables and the dependent variable "Success".

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.

For p-values lower than 5% (0.05) in the logistic regression analysis, we have considered the variable statistically significant, meaning the null hypothesis is rejected, and therefore conclude that there is a relationship that can used to predict the outcome between the predictors and the dependent variable.

Data

We went through the following processes to derive our current dataset:

Data Cleaning:

We filled in the missing values, replaced the N/A values with 0 (so that JMP is able to process our dataset). Since our original dataset contained tuples for each quote provided to an individual customer i.e. each customer had 7 tuples associated with him/her. We removed the duplicates using the customer
_id and record_type as the variables and then filtered the dataset based on the car_value= e. Later we randomly converted few customers around 10% as buyers and 90% as non-buyers.

Randomizing:

After cleaning the data and making it suitable for our purpose, we randomized the dataset by creating two columns of random variables and sorting our dataset in cycles of 3 using one random column at a time to randomize our dataset.

Training/Testing Dataset:

After performing the above steps our dataset consisted of 31,000 tuples. We funneled our dataset to build a model to 2000 tuples. We randomized this dataset of 2000 tuples to remove any bias in the dataset.

The data that we gathered displays information about people who previously received direct-mail marketing material for car insurance, and their decision whether to buy or not. This data will be used to learn the propensity to buy for the different categories/predictors, in effort to narrow down the targets.

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. For our model, we used 1,000 for training and 1,000 for testing.

ETL – Extract Transform Load Enrich

The variables given in the existing dataset are explained above. As we wanted to derive valuable business insights from our data, we created variables using the existing variables and our domain knowledge. We used some heuristic rules to determine the value of our new variables. These heuristic rules were applied to the existing variables. Enriched Variables are: *Divorce/Wedlock, has_kid, compare_c.*

The original data set that we got from Kaggle was not suited for our problem statement. The original problem statement was on re-targeting customers and was a sequential data-mining problem that would involve multiple regression, beyond the scope of this class. Therefore, we modified the dataset to suit our business problem and model it using the traditional classification method such as decision trees, logistic regression and neural network.
Here are some new variables that were created.

Has_Kid: We calculated the values for this variable using the age_oldest and age_youngest. The heuristic rule that we applied here is:

```
If age_oldest - age_youngest  > 15 then 1
                                Else 0
```

Where 1 means has a kid and 0 is doesn't have a kid.

Divorce/Wedlock: We used Has_Kid and married_couple to calculate whether a couple was divorced or a single parent. The heuristic rule that we applied here is:

```
If married_couple =1 && Has_Kid=1
then Divorce/Wedlock is 0
Elseif married_couple =0 && Has_kid =1
then Divorce/Wedlock is 1
Elseif married_couple=0 && Has_kid =0
then Divorce/Wedlock is 0
Else Divorce/Wedlock is 0
```

Compare_C: We used C_previous and Option C to check if the previous option that a customer had for option C and whether the Option offered to him/her in C are the same or not. The heuristic rule that we applied here is:

```
If C_previous = C then Compare_C =1
                Else Compare_C=0
```

We constructed these variables using our domain expertise. If someone has a kid there are more responsible and more likely to buy the car insurance. Divorce/Wedlock and has_kid are used in conjunction to predict whether a person will be a buyer or not.

By creating the above variables we are trying to create personas of our customers and trying to mine our data based on these variables and predicting the buyers of All-state insurance options.

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)
A, D, G	Coverage Options

The rest of the variables available in the data set are listed below, and will be shown in their entirety, with description, in the appendix:

- C-previous
- duration_previous
- A, B, C, E, F (coverage options)
- 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)

New Variables:

- has_kid
- divorced/wedlock
- Compare_C

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.

Analysis

Second Best Model – Logistic Regression

Input variables

- Homeowner
- Has_kids
- A
- D
- G
- Compare_C
- Car_age
- Married_couple
- Divorce/Wedlock

Output variable

= Record_type

Since our Y-variable is nominal and our X – variables are a mix of continuous and nominal variables, we decided to make use of Logistic Regression, the undisputed leader of traditional classification techniques.

We used logistic regression on the input variables and obtained the probability of record type = 1. We calculated the profit on these values using the Profit optimizer. The profit optimizer was modified to use our values.

*Please find all the appropriate diagrams in the appendix

Best Model – Logistic Regression

Input variables

- Home owner
- Has_kids
- A
- D
- G
- Compare_C
- Car_age
- Married_couple
- Divorced/Wedlocked
- Cost

Output variable

- Record_type

The results obtained from our second best model were satisfactory. However, we wanted to judge the effect of utilizing other variables. To this end, we maintained the base variables and tried making use of business insights to help choose the next variable. After trying out a few additional variables without any significant change, we finally included the '**cost'** variable. Initially, we were skeptical about using this feature as the results obtained from an analysis of the distribution of the variable showed the data to be highly skewed in a particular direction. We felt this would be inconclusive in predicting the probability.

On running the Logistic Regression model including the 'cost' variable, we obtained better results. Thus, cost did affect the prediction of record_type but in a positive way. Hence, we decided to include this variable too. From the analysis of the distribution of cost, it was noticed that the costs of the quotes being presented to the customers were heavily concentrated around the 550 – 750 range. This range could well represent the budget of the average customer, thereby lending credit to the notion that people tend to purchase insurance only if it lies within their specified budget range.

We ran the variables through a Neural Network model too. We however found that in our case, Logistic Regression tended to provide much more stable and meaningful results.

*Please find all the appropriate diagrams in the appendix

Performance Measure

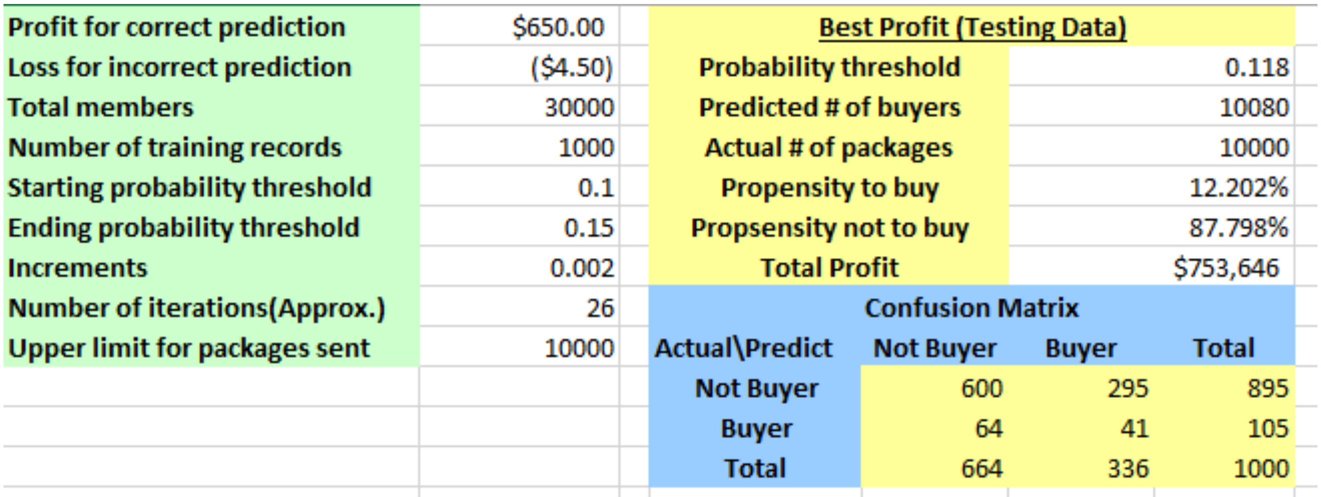
How did you rate your model, did you use confusion matrix, lift cure, R-sq, R-sq adj or others explain in detail. IF you tested your model on a new or "testing" data, what was the result? IF possible provide a "prediction" for you best model.

Repeatedly running different models with a slight change in the variables gave us a plethora of options to work with. These models however are of very limited use without an appropriate metric to quantify the performance of the model. We made use of several performance measures namely:

- Lift Curve
- R-square
- Adjusted R-square
- Confusion Matrix
- Accuracy (%)
- Misclassification Rate
- Profit (on Testing data)

The performance measures are as follows:

For Second Best Model

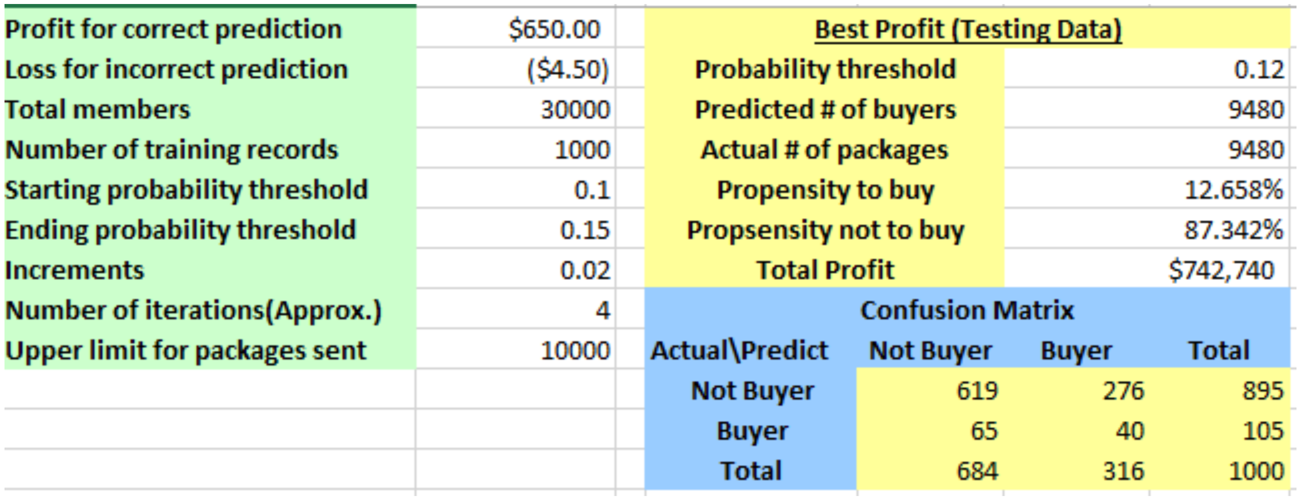


Profit for correct prediction	$650.00
Loss for incorrect prediction	($4.50)
Total members	30000
Number of training records	1000
Starting probability threshold	0.1
Ending probability threshold	0.15
Increments	0.002
Number of iterations(Approx.)	26
Upper limit for packages sent	10000

Best Profit (Testing Data)	
Probability threshold	0.118
Predicted # of buyers	10080
Actual # of packages	10000
Propensity to buy	12.202%
Propsensity not to buy	87.798%
Total Profit	$753,646

Confusion Matrix			
Actual\Predict	**Not Buyer**	**Buyer**	**Total**
Not Buyer	600	295	895
Buyer	64	41	105
Total	664	336	1000

For Our Best Model



Profit for correct prediction	$650.00
Loss for incorrect prediction	($4.50)
Total members	30000
Number of training records	1000
Starting probability threshold	0.1
Ending probability threshold	0.15
Increments	0.02
Number of iterations(Approx.)	4
Upper limit for packages sent	10000

Best Profit (Testing Data)	
Probability threshold	0.12
Predicted # of buyers	9480
Actual # of packages	9480
Propensity to buy	12.658%
Propsensity not to buy	87.342%
Total Profit	$742,740

Confusion Matrix			
Actual\Predict	**Not Buyer**	**Buyer**	**Total**
Not Buyer	619	276	895
Buyer	65	40	105
Total	684	316	1000

The other performance measures such as lift curve, R-square, Misclassification rate, profits etc. are included in the Appendix section.

Business Insights

The insurance business is very lucrative in nature with many complex computations that is a combination of a number of variables. The dataset that was provided to us was a small subset of the actual predictors that are needed to calculate a quote for a particular individual.

The obvious insight for a car insurance business is twofold, first is the insured person and the insured property. The insured person, younger the person more expensive the quote and it goes for the car as well, newer the car costlier the quote. This thinking was applied to our methodology to identify the propensity of buyers amongst the non-buyers and these variables were taken in to heavy consideration when making our technical model.

From our preliminary model, out of the 19% potential buyers, our models R-square value was very little around 3%, which did not make sense at all and contradicted with the model. Since R-square value is one of the indicators for the success of the model, we decided to go with other indicators like misclassification rate or the confusion matrix.

Some of the insights that we obtained that were not obvious were the inclusion of the variable C_previous in the original dataset. After applying domain expertise on this and getting multiple quotes from car insurance companies, we equated it as one of the major coverage's for a car (collision or bodily damage) which is very valuable for future quotes. In addition, car insurance re-targeting works like online advertisements, the potential buyers are constantly in kept in-touch with the quote they made to convert them in to actual buyers. This was only obvious to us after we looked at the dataset carefully, how there were multiple times for similar quotes for the same customer.

We then wanted to incorporate the variable “married_couple” along with the “age_oldest” and “age_youngest” variables provided. We wanted to differentiate single households with married/family households. This was done by enriching variables and figuring out if the household has a single person or is a family household and if the household is family, the age of the youngest driver and the cost would shoot up for them. We tried incorporating as much as domain expertise possible and making business inferences out of them.

Improvements

Some of the improvements that can be made to our project to achieve higher profit results are:

- Since the data is a sequential dataset, we would have been better with additional information about demographics. In addition, since this is a sequential data mining project, we could gain more knowledge about this topic and attempt the problem taking all the variables in to account i.e. containing all the tuples for the same customer for different quotes at different times
- We can leverage more psychographic information present in the dataset to arrive at our best model
- We hope to achieve more tribal knowledge by better analyzing the dataset and burning out variables and understanding the nuances of the problem
- Potentially make a star schema after accounting for all of the variables and understand if there is a one-to-one or one-to-many relationship
- Cluster the customer groups based on the states and create a heat map
- Create intersections between dependencies of individual data groups
- Use the quote just before the final quote as a testing measure (coverage options, day, time, cost)

Conclusion

Based on the model:

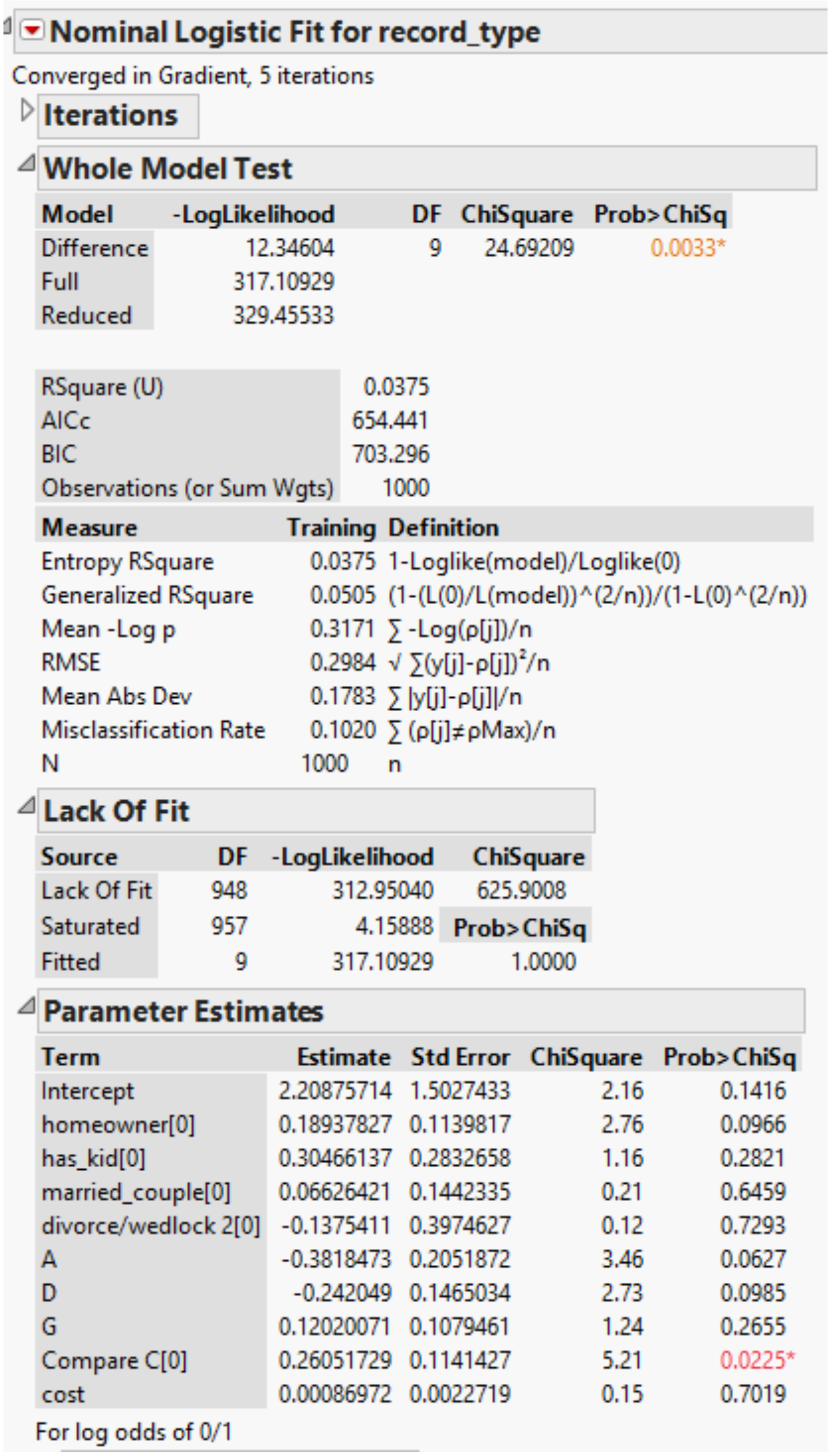
- R-square value = 3.75%
- Misclassification rate = 10.2%
- Compare_C is the most significant variable
- Base profit: $600,000 - Our profit: $753,000
- The jump in the overall profit was ~ 125%

From our results, we found out that most of the customers had a budget of $550-$750 to buy an insurance cover. We also achieved better results after continuous iterations with different models and variables chosen. We also concluded that real world data is hard to work with; we had to clean the data again and tweak the problem to bring it down to our scope. We took the ratios of some variables to enrich existing ones to shed light on other aspects of the dataset.

We also conclude that Compare_C is a very significant variable, customers who purchased a certain option of coverage C and had the same option of coverage C as their next purchase, were more likely to buy a quote that had the same option of coverage C. Our model has Rsquare of 3.75%, but this is statistically significant for our model since our dataset had high entropy and after cleaning the dataset were able to achieve this Rsquare. Overall we could attain a jump of 125% in the profit compared to the base profit.

Appendix

1. Second Best Model:



Nominal Logistic Fit for record_type

Converged in Gradient, 5 iterations

Iterations

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	12.34604	9	24.69209	0.0033*
Full	317.10929			
Reduced	329.45533			

RSquare (U)	0.0375
AICc	654.441
BIC	703.296
Observations (or Sum Wgts)	1000

Measure	Training	Definition
Entropy RSquare	0.0375	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.0505	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.3171	∑ -Log(ρ[j])/n
RMSE	0.2984	√ ∑(y[j]-ρ[j])^2/n
Mean Abs Dev	0.1783	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.1020	∑ (ρ[j]≠ρMax)/n
N	1000	n

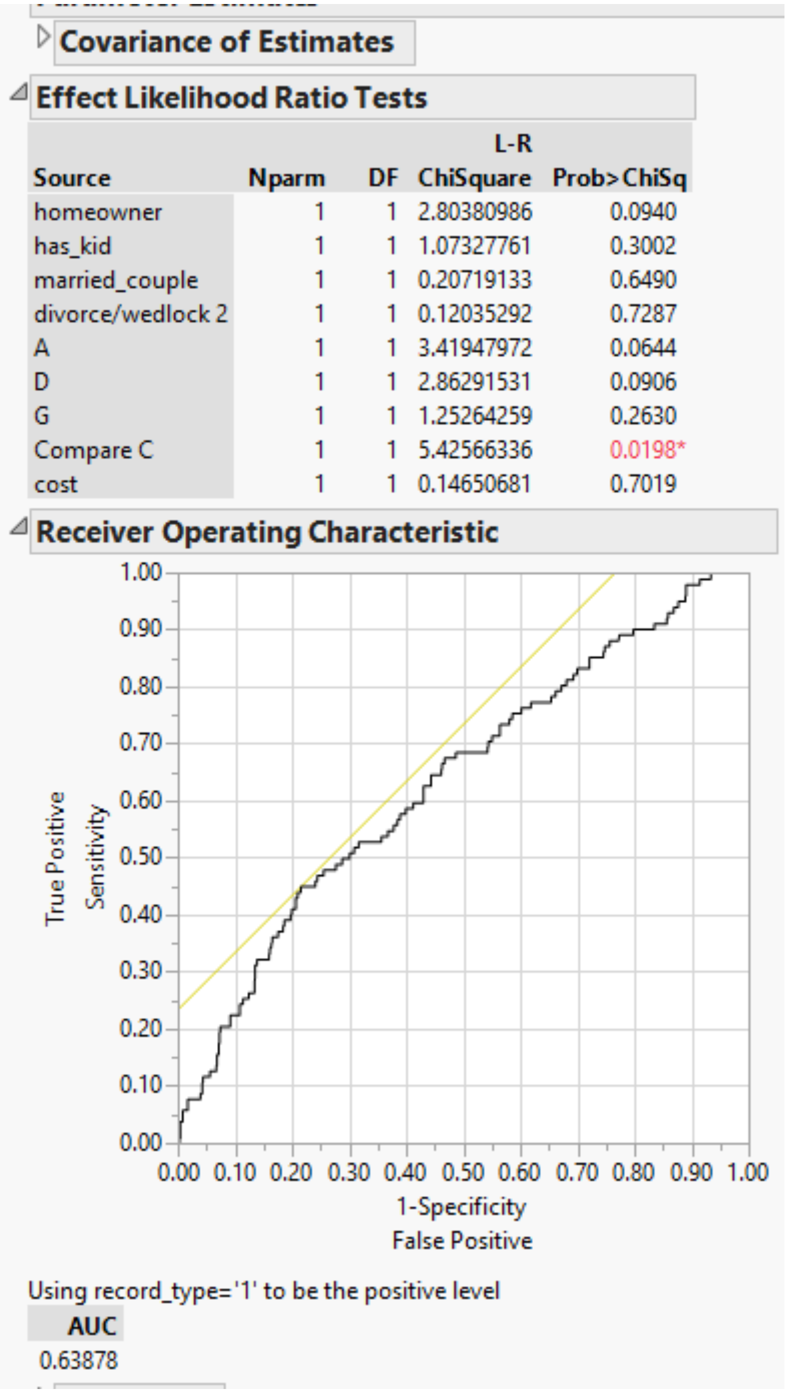
Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	948	312.95040	625.9008
Saturated	957	4.15888	Prob>ChiSq
Fitted	9	317.10929	1.0000

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	2.20875714	1.5027433	2.16	0.1416
homeowner[0]	0.18937827	0.1139817	2.76	0.0966
has_kid[0]	0.30466137	0.2832658	1.16	0.2821
married_couple[0]	0.06626421	0.1442335	0.21	0.6459
divorce/wedlock 2[0]	-0.1375411	0.3974627	0.12	0.7293
A	-0.3818473	0.2051872	3.46	0.0627
D	-0.242049	0.1465034	2.73	0.0985
G	0.12020071	0.1079461	1.24	0.2655
Compare C[0]	0.26051729	0.1141427	5.21	0.0225*
cost	0.00086972	0.0022719	0.15	0.7019

For log odds of 0/1



Covariance of Estimates

Effect Likelihood Ratio Tests

Source	Nparm	DF	L-R ChiSquare	Prob>ChiSq
homeowner	1	1	2.80380986	0.0940
has_kid	1	1	1.07327761	0.3002
married_couple	1	1	0.20719133	0.6490
divorce/wedlock 2	1	1	0.12035292	0.7287
A	1	1	3.41947972	0.0644
D	1	1	2.86291531	0.0906
G	1	1	1.25264259	0.2630
Compare C	1	1	5.42566336	0.0198*
cost	1	1	0.14650681	0.7019

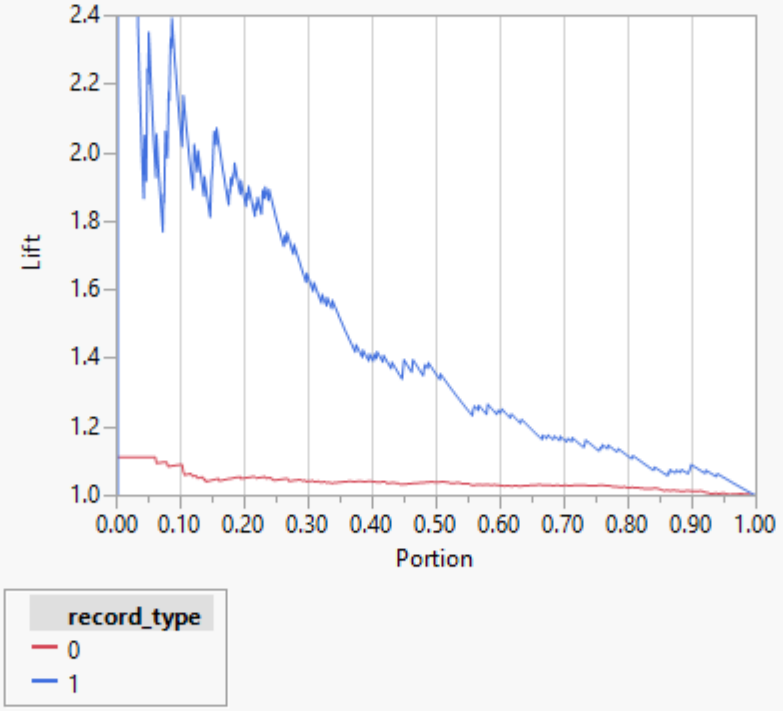
Receiver Operating Characteristic

Using record_type='1' to be the positive level

AUC
0.63878

ROC Table

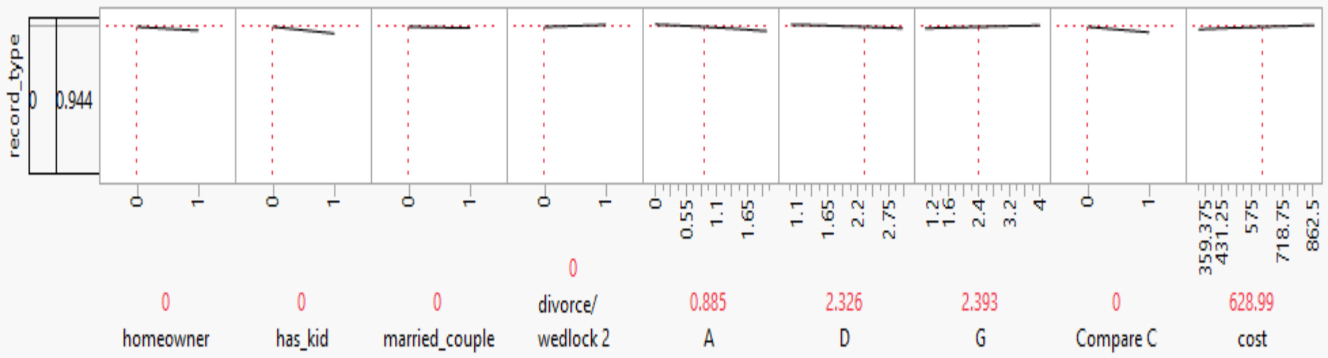
Lift Curve



Confusion Matrix

Actual		Predicted
Training	0	1
0	898	0
1	102	0

Prediction Profiler



homeowner 0

has_kid 0

married_couple 0

divorce/wedlock 2 0

A 0.885

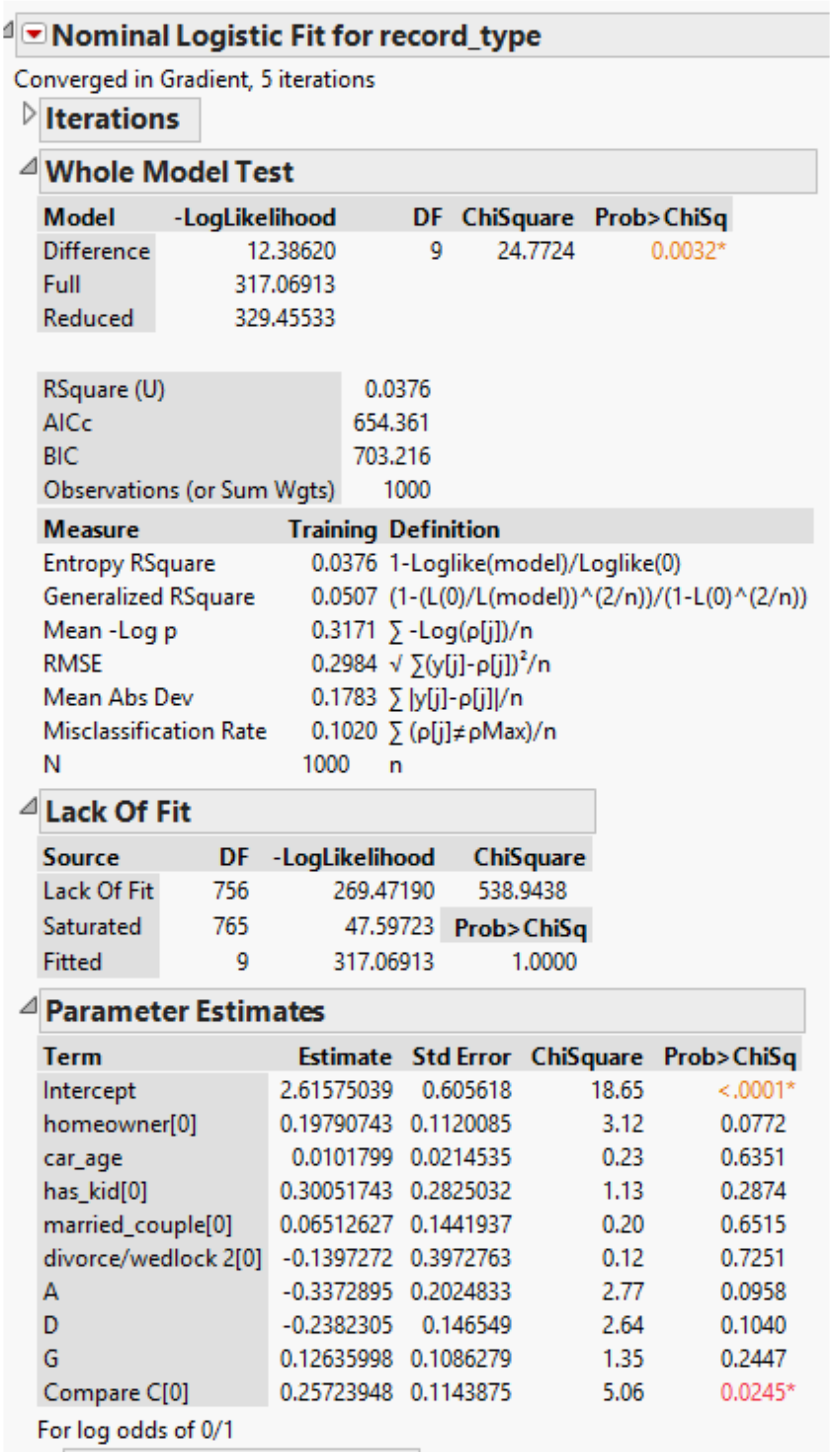
D 2.326

G 2.393

Compare C 0

cost 628.99

2. Best Model:



Nominal Logistic Fit for record_type

Converged in Gradient, 5 iterations

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	12.38620	9	24.7724	0.0032*
Full	317.06913			
Reduced	329.45533			

RSquare (U)	0.0376
AICc	654.361
BIC	703.216
Observations (or Sum Wgts)	1000

Measure	Training	Definition
Entropy RSquare	0.0376	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.0507	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.3171	∑ -Log(ρ[j])/n
RMSE	0.2984	√ ∑(y[j]-ρ[j])^2/n
Mean Abs Dev	0.1783	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.1020	∑ (ρ[j]≠ρMax)/n
N	1000	n

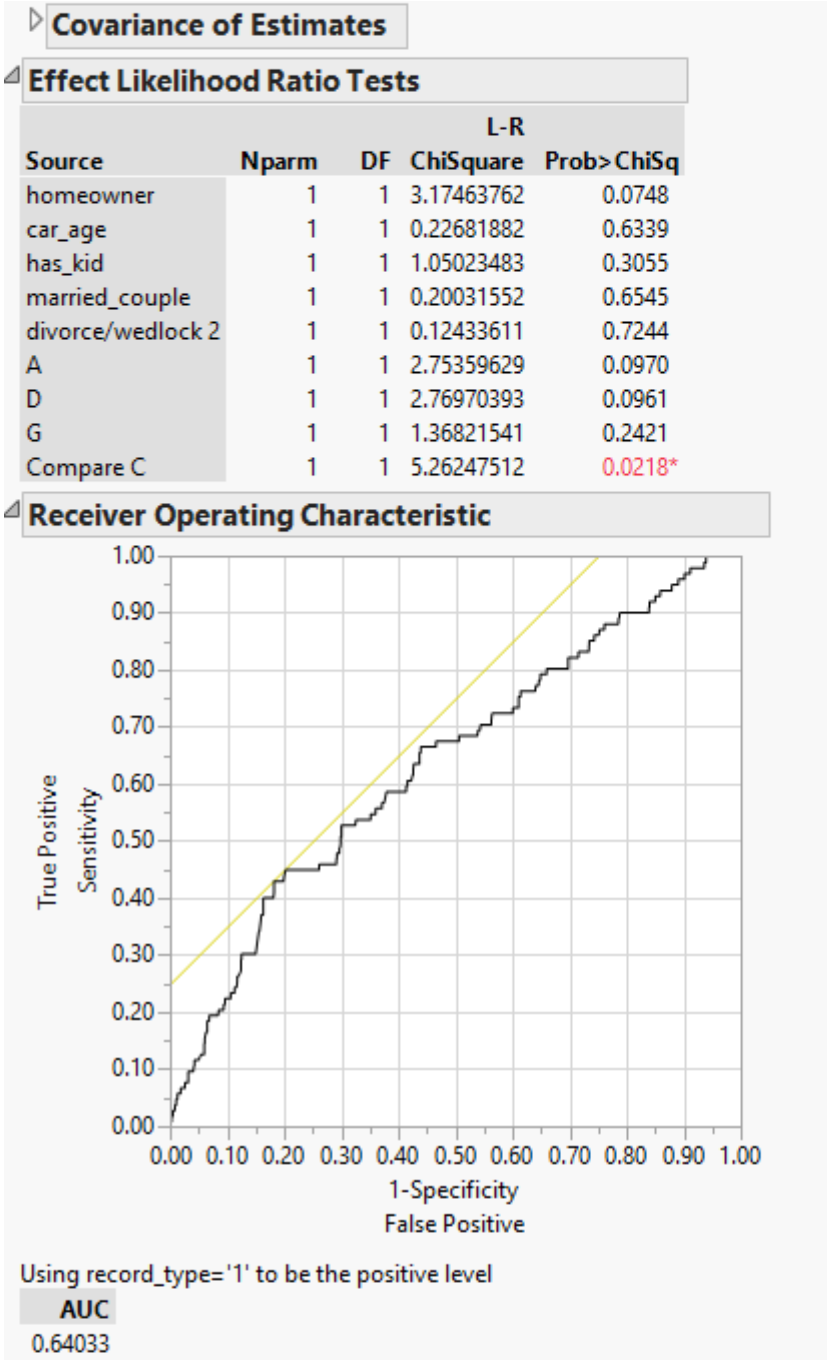
Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	756	269.47190	538.9438
Saturated	765	47.59723	Prob>ChiSq
Fitted	9	317.06913	1.0000

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	2.61575039	0.605618	18.65	<.0001*
homeowner[0]	0.19790743	0.1120085	3.12	0.0772
car_age	0.0101799	0.0214535	0.23	0.6351
has_kid[0]	0.30051743	0.2825032	1.13	0.2874
married_couple[0]	0.06512627	0.1441937	0.20	0.6515
divorce/wedlock 2[0]	-0.1397272	0.3972763	0.12	0.7251
A	-0.3372895	0.2024833	2.77	0.0958
D	-0.2382305	0.146549	2.64	0.1040
G	0.12635998	0.1086279	1.35	0.2447
Compare C[0]	0.25723948	0.1143875	5.06	0.0245*

For log odds of 0/1



Covariance of Estimates

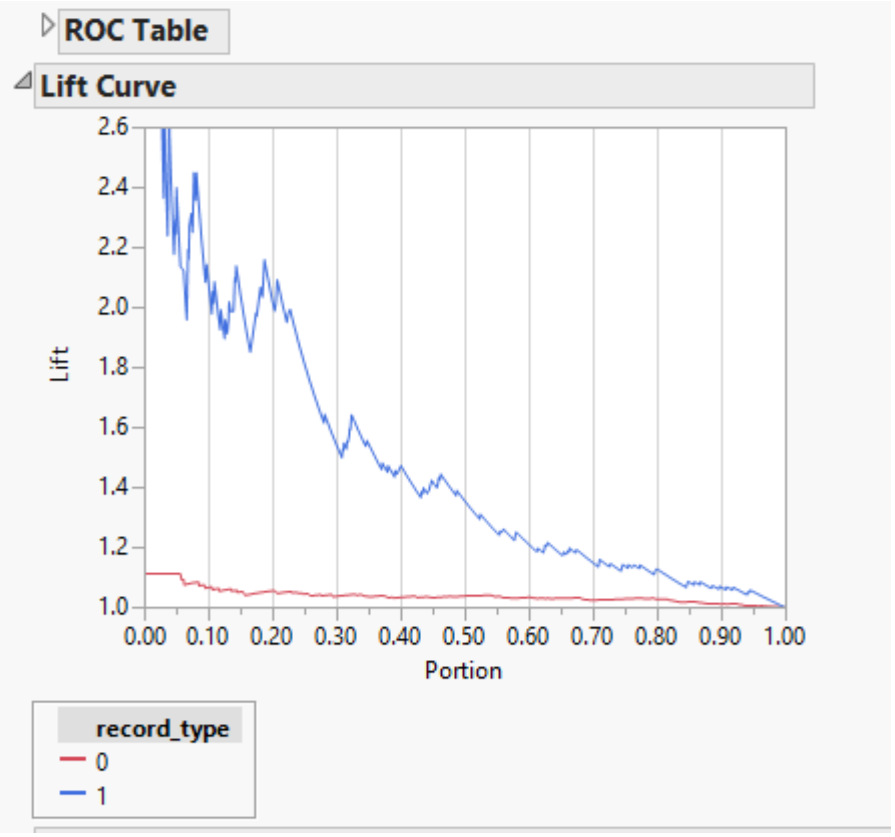
Effect Likelihood Ratio Tests

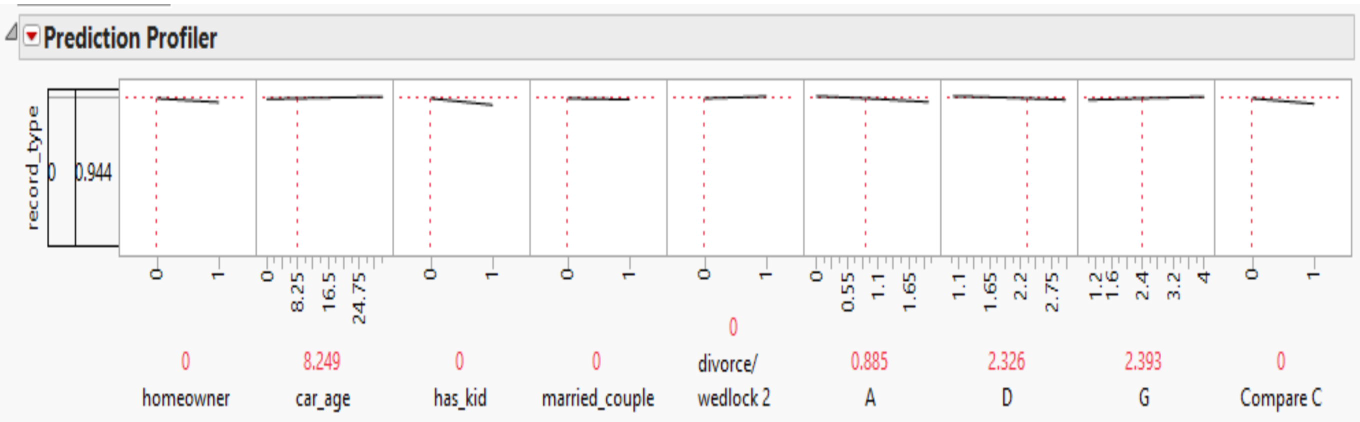
Source	Nparm	DF	L-R ChiSquare	Prob>ChiSq
homeowner	1	1	3.17463762	0.0748
car_age	1	1	0.22681882	0.6339
has_kid	1	1	1.05023483	0.3055
married_couple	1	1	0.20031552	0.6545
divorce/wedlock 2	1	1	0.12433611	0.7244
A	1	1	2.75359629	0.0970
D	1	1	2.76970393	0.0961
G	1	1	1.36821541	0.2421
Compare C	1	1	5.26247512	0.0218*

Receiver Operating Characteristic

Using record_type='1' to be the positive level

AUC
0.64033


ROC Table
Lift Curve
Lift
2.6
2.4
2.2
2.0
1.8
1.6
1.4
1.2
1.0
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00
Portion
record_type
0
1


Prediction Profiler
record_type
0
0.944
0
1
0
8.25
16.5
24.75
0
1
0
1
0
1
0
0.55
1.1
1.65
1.1
1.65
2.2
2.75
1.2
1.6
2.4
3.2
4
0
1
0
homeowner
8.249
car_age
0
has_kid
0
married_couple
0
divorce/
wedlock 2
0.885
A
2.326
D
2.393
G
0
Compare C