Anatou Soumahoro

Data mining

Final project

Data source : <https://www.ibm.com/communities/analytics/watson-analytics-blog/marketing-customer-value-analysis/>

Synopsis:

AutoOne is car insurance company across the United States. The company commercializes different car insurance products through diverse channels (Call centers, agents …). The attached dataset contains the overall Customer Value (present value of future spends with ABC) of some of ABC’s customers throughout USA. Marketing seeks to identify the characteristics of customers responding to offers in order to better tailor which offers are made to which customers.

Business purpose: Determine which client respond to offer proposed by the company and allow a segmented marketing

Data dictionary

|  |  |  |
| --- | --- | --- |
| Variables | Type | Description |
| Customer ID | Categorical | Unique identifier |
| State | Categorical |  |
| Customer Lifetime Value | Numeric | The value generated by the customer for the insurance company |
| Response | Dependent Variable - Character | Whether the customer responded to the campaign and signed up for the re |
| Coverage | Categorical | Premium, Basic or extended coverage |
| Education | Categorical | The highest education level reached |
| Effective To Date | Date |  |
| Employment Status | Categorical | Unemployed, employed, Disabled, retired or medical leave |
| Gender | Categorical | Male or female |
| Income | Numeric | In $$$ |
| Location Code | Categorical | Suburban, urban or rural |
| Marital Status | Categorical | Divorced, married or single |
| Monthly Premium Auto | Numeric | In $$$ |
| Months Since Last Claim | Numeric |  |
| Months Since Policy Inception | Numeric |  |
| Number of Open Complaints | Numeric |  |
| Number of Policies | Numeric |  |
| Policy Type | Categorical | Personal or corporate |
| Policy | Categorical |  |
| Renew Offer Type | Categorical | The company proposes 3 different renewal offer |
| Sales Channel | Categorical | Sales conducted thru agent, call center, Branch or Web |
| Total Claim Amount | Numeric |  |
| Vehicle Class | Categorical |  |
| Vehicle Size | Categorical | SUV, Large or Medsize |

Data Description

'data.frame': 9134 obs. of 24 variables:

$ Customer : Factor w/ 9134 levels "AA10041","AA11235",..: 601 5947 97 8017 2489 4948 8434 756 1352 548 ...

$ State : Factor w/ 5 levels "Arizona","California",..: 5 1 3 2 5 4 4 1 4 4 ...

$ Customer.Lifetime.Value : num 2764 6980 12887 7646 2814 ...

$ Response : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 2 1 2 1 ...

$ Coverage : Factor w/ 3 levels "Basic","Extended",..: 1 2 3 1 1 1 1 3 1 2 ...

$ Education : Factor w/ 5 levels "Bachelor","College",..: 1 1 1 1 1 1 2 5 1 2 ...

$ Effective.To.Date : Factor w/ 59 levels "1/1/11","1/10/11",..: 48 25 42 13 53 18 48 10 19 40 ...

$ EmploymentStatus : Factor w/ 5 levels "Disabled","Employed",..: 2 5 2 5 2 2 2 5 3 2 ...

$ Gender : Factor w/ 2 levels "F","M": 1 1 1 2 2 1 1 2 2 1 ...

$ Income : int 56274 0 48767 0 43836 62902 55350 0 14072 28812 ...

$ Location.Code : Factor w/ 3 levels "Rural","Suburban",..: 2 2 2 2 1 1 2 3 2 3 ...

$ Marital.Status : Factor w/ 3 levels "Divorced","Married",..: 2 3 2 2 3 2 2 3 1 2 ...

$ Monthly.Premium.Auto : int 69 94 108 106 73 69 67 101 71 93 ...

$ Months.Since.Last.Claim : int 32 13 18 18 12 14 0 0 13 17 ...

$ Months.Since.Policy.Inception: int 5 42 38 65 44 94 13 68 3 7 ...

$ Number.of.Open.Complaints : int 0 0 0 0 0 0 0 0 0 0 ...

$ Number.of.Policies : int 1 8 2 7 1 2 9 4 2 8 ...

$ Policy.Type : Factor w/ 3 levels "Corporate Auto",..: 1 2 2 1 2 2 1 1 1 3 ...

$ Policy : Factor w/ 9 levels "Corporate L1",..: 3 6 6 2 4 6 3 3 3 8 ...

$ Renew.Offer.Type : Factor w/ 4 levels "Offer1","Offer2",..: 1 3 1 1 1 2 1 1 1 2 ...

$ Sales.Channel : Factor w/ 4 levels "Agent","Branch",..: 1 1 1 3 1 4 1 1 1 2 ...

$ Total.Claim.Amount : num 385 1131 566 530 138 ...

$ Vehicle.Class : Factor w/ 6 levels "Four-Door Car",..: 6 1 6 5 1 6 1 1 1 1 ...

$ Vehicle.Size : Factor w/ 3 levels "Large","Medsize",..: 2 2 2 2 2 2 2 2 2 2 ...

* 9134 observations with a total of 24 variables with information about existing clients
* No missing values
* No duplicates based on customer ID
* Pretty consistent response, no outliers.

How much response do we actually have?

> table(Target.df$Response)/nrow(Target.df)\*100

No Yes

85.67988 14.32012 (in %)

*Let's see how much response we have by offer*

> table(Target.df$Response,Target.df$Renew.Offer.Type)

Offer1 Offer2 Offer3 Offer4

No 3158 2242 1402 1024

Yes 594 684 30 0

* We get the most positive response for Offer 2

Let’s See how much response we get depending of the number of policy a client has.

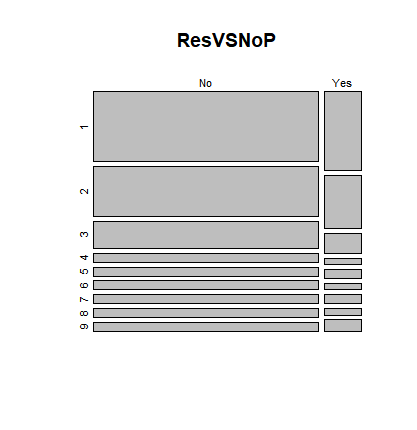
> table(Target.df$Response,Target.df$Number.of.Policies)

1 2 3 4 5 6 7 8 9

No 2735 1952 1036 367 347 330 373 342 344

Yes 516 342 132 42 60 42 60 42 72

* We get the most positive response with client with only one policy with company.



Let’s see how much response we get depending of the sales channel

Agent Branch Call Center Web

No 2811 2273 1573 1169

Yes 666 294 192 156

We get the most positive response when the sale is through agent – Kuddos to agent !!!

What about the marital status

Divorced Married Single

No 1045 4602 2179

Yes 324 696 288

Married people have the most positive answer – most of our customer are married let’s try to see percentage

Divorced Married Single

No 11.440771 50.383184 23.855923

Yes 3.547186 7.619882 3.153055

What about the location code

> table(Target.df$Response,Target.df$Location.Code)/nrow(Target.df)\*100

Rural Suburban Urban

No 17.637399 52.233414 15.809065

Yes 1.773593 11.035691 1.510839

We get the most positive response from client living in the suburban

### For our analysis we created a new variable that we will call engagement – as a dummy variable, if the client answered to an offer (Yes =1 , No = 0)

Let see how much engagement we have

0 1

Count 7826.00000 1308.00000

Percentage 85.67988 14.32012

Training and validation set created.

target2.df<-Target.df

train.index<-sample(c(1:dim(target2.df)[1]),dim(target2.df)[1]\*0.7)

train.df<-target2.df[train.index,selected.var]

valid.df<-target2.df[-train.index,selected.var]

Model

The data were first partitioned into training (70%) and validation (30%) sets, and then Naïve bayes classifier was applied to the training set ( we use package e1071)

we want to use Naive Bayes classifier to study the engagement subject to the type of coverage, Marital status, employment status, Location, Renew offer and sales channel

Naïve Bayes results

> target2.nb<-naiveBayes(Engaged~., data = train.df)

> target2.nb

Naive Bayes Classifier for Discrete Predictors

Call:

naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:

Y

0 1

0.8568747 0.1431253

Conditional probabilities:

Coverage

Y Basic Extended Premium

0 0.60843373 0.30138737 0.09017890

1 0.61311475 0.29071038 0.09617486

Employment

Y Disabled Employed Medical Leave Retired Unemployed

0 0.041255933 0.636728733 0.047097481 0.009492516 0.265425338

1 0.059016393 0.590163934 0.054644809 0.142076503 0.154098361

Location

Y Rural Suburban Urban

0 0.2077401 0.6060606 0.1861993

1 0.1267760 0.7693989 0.1038251

Marital\_Status

Y Divorced Married Single

0 0.1323476 0.5881709 0.2794816

1 0.2513661 0.5245902 0.2240437

Renew\_Offer

Y Offer1 Offer2 Offer3 Offer4

0 0.40051114 0.29061701 0.17999270 0.12887915

1 0.46338798 0.51256831 0.02404372 0.00000000

Sales\_Channel

Y Agent Branch Call Center Web

0 0.3501278 0.2968237 0.2000730 0.1529755

1 0.5071038 0.2218579 0.1519126 0.1191257

> confusionMatrix(table(pred.class1, train.df$Engaged))

Confusion Matrix and Statistics

pred.class1 0 1

0 5432 784

1 46 131

Accuracy : 0.8702

95% CI : (0.8617, 0.8783)

No Information Rate : 0.8569

P-Value [Acc > NIR] : 0.00113

Kappa : 0.2029

Mcnemar's Test P-Value : < 2e-16

Sensitivity : 0.9916

Specificity : 0.1432

Pos Pred Value : 0.8739

Neg Pred Value : 0.7401

Prevalence : 0.8569

Detection Rate : 0.8497

Detection Prevalence : 0.9723

Balanced Accuracy : 0.5674

'Positive' Class : 0

Our model does a good job at predict negative engagement ( no response) but it is not effective to predict positive engagement ( yes response)

Let’s run on validation dataset

> #validation

> pre.class2<-predict(target2.nb, newdata = valid.df)

> confusionMatrix(table(pred.class2, valid.df$Engaged))

Confusion Matrix and Statistics

pred.class2 0 1

0 2335 320

1 13 73

Accuracy : 0.8785

95% CI : (0.8657, 0.8905)

No Information Rate : 0.8566

P-Value [Acc > NIR] : 0.0004677

Kappa : 0.2671

Mcnemar's Test P-Value : < 2.2e-16

Sensitivity : 0.9945

Specificity : 0.1858

Pos Pred Value : 0.8795

Neg Pred Value : 0.8488

Prevalence : 0.8566

Detection Rate : 0.8519

Detection Prevalence : 0.9686

Balanced Accuracy : 0.5901

'Positive' Class : 0

The accuracy is pretty much the same ~ 87% on both training dataset and validation dataset.

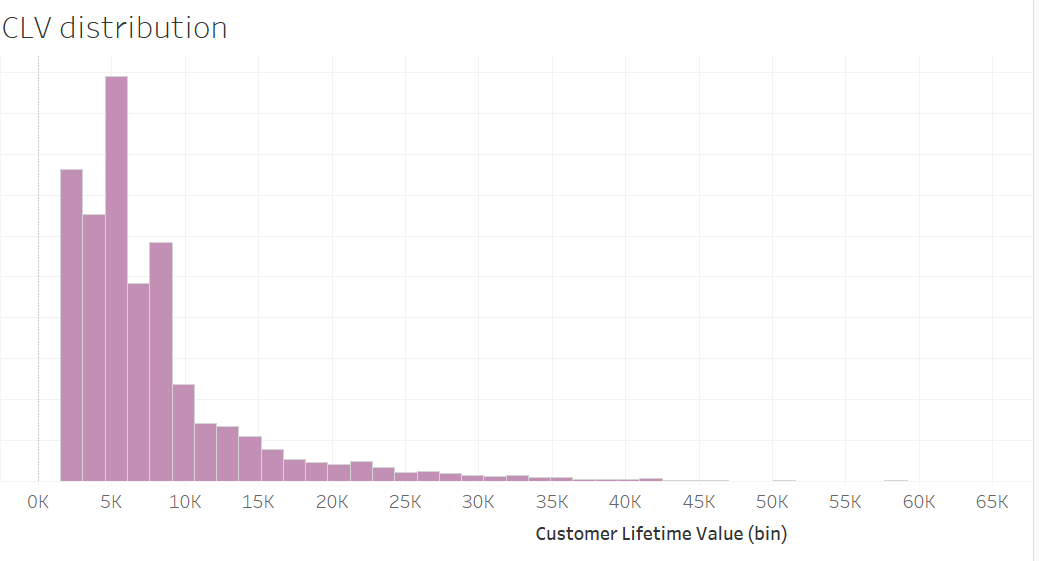
2nd Model

Exploratory analysis: use logistic regression to determine which variable will influence the target of customer lifetime value

> summary(Customer.Lifetime.Value)

Min. 1st Qu. Median Mean 3rd Qu. Max.

1898 3994 5780 8005 8962 83325



CLV is skewed to the right

We set the target CLV at $8000 because it is close to the mean

Model

> summary(Regression\_Test)

Call:

glm(formula = Customer.Lifetime.Value ~ ., family = "binomial",

data = TargetDem.df)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.8892 -0.7449 -0.6693 0.9814 1.9919

Pr(>|z|)

(Intercept) 1.81e-13 \*\*\*

StateCalifornia 0.56045

StateNevada 0.61999

StateOregon 0.42310

StateWashington 0.65033

EducationCollege 0.93951

EducationDoctor 0.75412

EducationHigh School or Below 0.06030 .

EducationMaster 0.25252

EmploymentStatusEmployed 0.90995

EmploymentStatusMedical Leave 0.50547

EmploymentStatusRetired 0.32143

EmploymentStatusUnemployed 0.19688

GenderM 0.38505

Income 0.17996

Location.CodeSuburban 0.33043

Location.CodeUrban 0.30940

Marital.StatusMarried 0.06318 .

Marital.StatusSingle 0.00162 \*\*

Policy.TypePersonal Auto 0.54944

Policy.TypeSpecial Auto 0.01920 \*

Vehicle.ClassLuxury Car < 2e-16 \*\*\*

Vehicle.ClassLuxury SUV < 2e-16 \*\*\*

Vehicle.ClassSports Car < 2e-16 \*\*\*

Vehicle.ClassSUV < 2e-16 \*\*\*

Vehicle.ClassTwo-Door Car 0.86728

Vehicle.SizeMedsize 0.03107 \*

Vehicle.SizeSmall 0.13629

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 11631 on 9133 degrees of freedom

Residual deviance: 10265 on 9106 degrees of freedom

AIC: 10321

Number of Fisher Scoring iterations: 4

We re – run the regression with only the variables that showed significance

glm(formula = Customer.Lifetime.Value ~ Education + Location.Code +

Marital.Status + Vehicle.Class, family = "binomial", data = TargetDem.df)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.7888 -0.7375 -0.6835 0.9969 1.8739

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -1.039318 0.088380 -11.760 < 2e-16 \*\*\*

EducationCollege -0.003813 0.062952 -0.061 0.95170

EducationDoctor -0.037650 0.133051 -0.283 0.77720

EducationHigh School or Below 0.121115 0.063138 1.918 0.05508 .

EducationMaster 0.111800 0.094432 1.184 0.23645

Location.CodeSuburban -0.172017 0.063549 -2.707 0.00679 \*\*

Location.CodeUrban -0.075371 0.078722 -0.957 0.33834

Marital.StatusMarried -0.120055 0.068948 -1.741 0.08164 .

Marital.StatusSingle -0.317054 0.078241 -4.052 5.07e-05 \*\*\*

Vehicle.ClassLuxury Car 2.412618 0.184755 13.058 < 2e-16 \*\*\*

Vehicle.ClassLuxury SUV 2.391079 0.174580 13.696 < 2e-16 \*\*\*

Vehicle.ClassSports Car 1.682724 0.099790 16.863 < 2e-16 \*\*\*

Vehicle.ClassSUV 1.651016 0.060044 27.497 < 2e-16 \*\*\*

Vehicle.ClassTwo-Door Car 0.010612 0.066056 0.161 0.87237

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

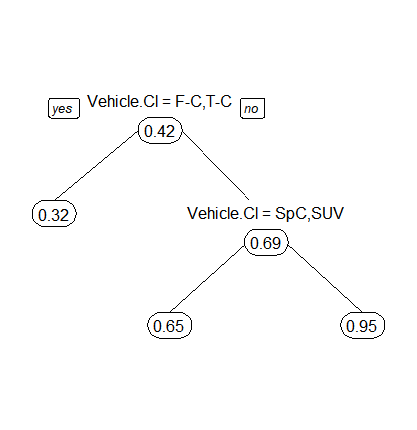
(Dispersion parameter for binomial family taken to be 1)

Null deviance: 11631 on 9133 degrees of freedom

Residual deviance: 10295 on 9120 degrees of freedom

AIC: 10323

Verification with decision tree



Conclusion: When looking at demographic variables, the vehicle class is the main indicator that can determine if a future client will have a Customer lifetime value of +$8000. AutoOne should create a marketing campaign focused on owner of luxury cars and SUV.