CUSTOMER VALUE ANALYSIS REGRESSION PROCESS

Rishwanth Ravindran

Contents

Regression Case Study - Predict Customer Life-time Value for an Auto Insurance Company

Objective

For an Auto Insurance company, we have to predict the customer life time value (CLV). CLV is the total revenue the client will derive from their entire relationship with a customer. Because we don't know how long each customer relationship will be, we make a good estimate and state CLV as a periodic value

Step-1

Identifying the problem statement, our objective is to maximize revenue as the CLV is the total revenue the client will derive from their entire relationship with a customer

Step-2

Here we need to identify the target variable, in this dataset the target variable is Customer Lifetime Value

Target variable: Customer Lifetime Value

Step-3

Here we load the data into the R Studio and we look for missing values to treat them

In R, missing values are identified by "NA", we use functions to convert the blanks, extra spaces, cells containing the word NA or NULL in string columns to "NA"

In our data there are no missing values

Step-4

Here we explore the dataset using the exploratory commands which gives the descriptive statistics of the data

We identify what kind of variables are there? Here we spot Customer Lifetime Value, which is the Target Variable, which you want to fit

And then we remove useless columns in the data, and explore the rest, we also check if all the categorical variables are factor or not

Categorical column : (State, Response, Coverage, Education, Employment Status, Gender, Location Code, Marital Status,

Monthly Premium Auto, Months Since Last Claim, Months Since Policy Inception, Type of Open Complaints,

Type of Policies, Policy Type, Policy, Renew Offer Type, Sales Channel, Vehicle Class, Vehicle Size)

In that case we treat the column to change them into the suitable datatype

Step-5

Here we check whether it is a Regression problem or Classification? by looking at the Target variable we conclude the model

Here it is continuous variable, so our model will the Regression Algorithm

Step-6

We check for treating missing values, we need to treat missing values before we move ahead,

there are no missing values in the data, after removing the garbage columns and treating the missing values

Now we see for which of the columns the treatment of the outliers has to be carried out using boxplot

Here Customer Lifetime Value is having outliers based on the boxplot

For treating the missing values, we sort the Customer Lifetime Value column in ascending order

Here, there are some very high values in the end which are outliers for the Customer Lifetime Value column, we see what is last value in the data, which you won't consider as the outliers

We check which quantile it approximately corresponds to and all the values above that will be replaced and check the boxplot again and see whether outliers are removed or not

Step-7

Exploring the CONTINUOUS & CATEGORICAL features

Explore each "Potential" predictor for distribution and Quality

Here we understand whether these potential predictors are worthy of

selection or not- based on their distribution and we'll confirm with the help of different tests

Graphical representation involves histogram for continuous column and bar plot for categorical column

Library (RColorBrewer) is used to generate professional colours

We create looping to create the histograms for each column, for each of the columns, we are plotting the histogram by iteration

Step-8

Bivariate analysis

We generate visual Relationship between predictors and target variable, for Continuous Vs Continuous we use Scatter Plot for Continuous Vs Categorical we use Box Plot

Step-10

We check the strength of relationship between predictor and target variable, for Continuous Vs Continuous we use Correlation test for Continuous Vs Categorical we ANOVA test

We measure the strength of the relationship for Continuous Vs Continuous by Correlation analysis, correlation coefficient varies between -1 & 1

Here we use (use = "complete.obs") means use only those rows which are complete (No Missing values)

for perfect positive correlation: r=1

for perfect negative correlation: r=-1

for zero correlation: r=0

the following helps us to decide which col to keep and which one to reject

Final columns which have the high correlation with the target variable for which absolute correlation with Customer Lifetime Value is >0.2 are

Customer Lifetime Value is not a predictor, use the rest of the variables as potential continuous predictors

Based on the correlation test we choose Monthly Premium Auto, Total Claim Amount as the potential predictor

We measure the Continuous Vs Categorical correlation strength using the ANOVA test

Null Hypothesis H0: Variables are NOT correlated

Small P-Value <5%--> Variables are correlated (H0 is rejected)

Large P-Value--> Variables are NOT correlated (H0 is accepted)

Based on the ANOVA test we choose Coverage, Education, Employment Status, Marital Status, Type of Policies, Renew Offer Type, Vehicle Class, Vehicle Size as the potential predictor

Step-11

Here we try to bring the data in standardised format, to reduce coding effort

Generating the Data frame for machine learning

We choose the multiple Predictors which may have relation with Target Variable

based on the exploratory data analysis, select the useful variables

Target Variable Name: Customer Lifetime Value

Best Predictor Name: Monthly Premium Auto, Total Claim Amount, Coverage, Education,
Employment Status, Marital Status, Type of Policies, Renew Offer Type,
Vehicle Class, Vehicle Size

We extract the Target and Predictor variable columns from whole data to create a

generic dataset

By selecting all other columns as Predictors apart from target variable, we create the final data to be used for ML

Step-12

Here sampling process is done for splitting data into 70% for training 30% for testing

1: n row (Data For ML): It will give you the row no 1 to (total no. of records)

that is, the total set of rows

We want 70% of the total no. of rows of the original data for training and rest 30% of the rows are used for testing

We create Predictive models on training data to check the accuracy of each algorithm

Linear Regression

We predict TV based on all other variables in Data For ML Train

As a result Estimate column gives the intercept and the coefficient estimates associated to each predictor variable

the m1, m2,……c are what the algorithm is searching using the trial and error method. It will again find out the particular set of values which is giving me the minimum possible SSE which is the backend process

Residual standard error: 3429 on 6380 degrees of freedom

Multiple R-squared: 0.6901,   Adjusted R-squared: 0.6896

Multiple R-squared is higher than Adjusted R-squared and thus showing the model is having good fit

F-statistic: 1184 on 12 and 6380 DF, **p-value: < 2.2e-16**

Now we check for Accuracy of model on Testing data

Checking for accuracy

For final accuracy we take the mean/median of all the errors and subtract it from 100

Decision Tree

The DT algorithm function: ctree is present in the party library

ctree is going to see those minute fluctuations in Customer Lifetime Value and in each of these columns

Here DT will select only the columns which helps to bifurcate at each step.

You have the Target variable, predicted by all other predictors in Data For ML Train

The Decision will use only those variables which are useful in prediction, and it will simply ignore

other variables.

It is doing automatic feature selection

DT will choose those variables only which are helping to bifurcate

the algorithm will find out the bifurcating variables on its own

Checking Accuracy of model on Testing data

Then we generate the predictions using decision tree and arriving at the accuracy

For average accuracy we take the average of all the errors and subtract it from 100

Comparison of models

Linear Regression

Mean Accuracy of Linear Regression Model is: 79.0169856461506

Median Accuracy of Linear Regression Model is: 86.9119136041219

Decision Tree

Mean Accuracy of Decision tree Model is: 88.3197276515216

Median Accuracy of Decision tree Model is: 97.7292847875328

Based on accuracy Decision Tree is showing high accuracy

By looking at the p-value to conclude the result of the test

Test for homoskedasticity

Null Hypothesis

HO: there exists homoskedasticity: error variances are equal

p-value < 0.00000000000000022

Null Hypothesis is rejected

Test for serial correlation

Autocorrelation occurs when the residuals are not independent from each other

Null Hypothesis

HO: No autocorrelation

p- value > 0.05 means HO is accepted

p-value = 0.6269

Null Hypothesis is accepted

Test for normality

Null Hypothesis

HO: errors are normally distributed

p-value < 0.00000000000000022

Null Hypothesis is rejected

Test for multicollinearity

VIF values are close to 1, we are good to go with our regression model