**Predicting Customer Life-time Value for an Auto Insurance Company**

**Objective**: To predict Customer Life-time values (CLV) for an Auto Insurance Company by framing a regression model, examining the factors that influence CLV, thus predicting the most profitable customers to the company. Customer lifetime value is the total revenue the client will derive from their entire relationship with a customer.

The data to be studied has 24 variables and 9134 observations. The variables provided are as follows:

Customer: Customer ID of the customer in Client’s company

State: State the customer belongs to

Customer Lifetime Value: Total revenue the client will derive from their entire relationship with the customer

Response: Customer response consisting of Yes or No

Coverage: Coverage provided to the customer by the insurance company. Can be Basic, Extended or Premium

Education: Customer’s educational qualification. Consists of five categories; Bachelor, College, Doctor, High School or below, Master

Effective To Date: The date when the insurance coverage goes into effect

EmploymentStatus: Employment status of the customer. Consists of five categories; Disabled, Employed, Medical Leave, Retired, Unemployed

Gender: Whether customer is male or female

Income: Customer’s income level

Location Code: Customer’s location type. Consists of three categories; Rural, Urban, Sub- urban

Marital Status: Customer’s marital status. Consists of three categories; Single, Married, Divorced

Monthly Premium Auto: The monthly amount paid by the customer for the automobile insurance

Months Since Last Claim: Months since the last claim or request made by customer for coverage or compensation.

Months Since Policy Inception: Months since inception of policy

Number of Open Complaints: Number of complaints made by the customer.

Number of Policies: Number of policies the customer holds

Policy Type: Type of policy the customer holds. It can be of 3 types, corporate, personal,special.

Policy: Policy sub- type, whether its L1, L2, or L3

Renew Offer Type: Offer to renew insurance. Is of 4 types

Sales Channel: The mode via which the customer acquired the insurance policy. It is of 4 types; Agent, Branch, Call Centre, Web

Total Claim Amount: Sum payable at the maturity of an insurance policy

Vehicle Class: Class of vehicle used by customer. It is of 6 types; Four door car, Luxury Car, Luxury SUV, SUV, Sports cart, Two door car.

Vehicle Size: Size of the vehicle used. Can be Large, Medium size or Small.

**Approach:**

The insurance company seeks to increase its revenue, determined by the Customer Lifetime Value. Our objective will be to find factors that will increase revenue.

CLV is a continuous variable. Hence, we will take the Linear Regression model approach.

We assume CLV as the dependent variable and the other variables as independent variables.

The purpose of regression is to understand the factors that affect CLV and to predict the value of dependent variable, given the values of independent variables.

MLRM can be represented as:

Y=B0+BiXi+e

CLV is set as the dependent variable(Y) and other variables in the data are set as independent variables (Xi)

Null hypothesis of the model is set as Bi=0 (unit change in independent variable causes no change in dependent variable

Alternative hypothesis of the model is Bi =/= 0 (unit change in independent variable causes change in dependent variable

**Steps:**

1. **Data exploration**

The data is imported into R, the structure of the data is studied. There are 24 variables and 9134 observations. Variables “Number of complaints” and “Number of Policies” are converted into factor variables to decipher the number of sub categories among them. Variable Customer lifetime value is renamed as CLV, and is considered as the dependent variable.

The dependent variable CLV is checked for outliers. Presence of outliers is found. Outlier treatment is done by capping till value 9119.

Next, independent variable, Monthly Premium Auto is checked for outliers and capped at value 140.

Independent Variable Income is capped at value 82412

Data is checked for missing values, none are found.

Next, variables deemed as irrelevant to our study are dropped off. These variables are Customer, State and Effective to date.

Then the final data frame is saved as a csv file named as “Projectdata”.

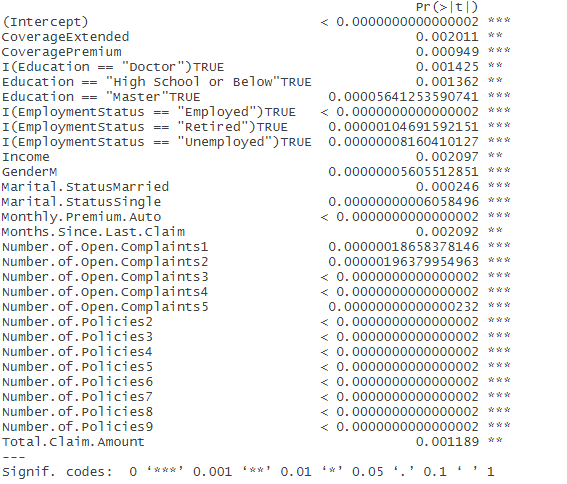
Next, the data frame is divided into training and validation data sets in a 70:30 format. The training data set is named as train.data and the validation data set is named as test.data.

Train.data has 4265 observations and 21 variables

Test.data has 1829 observations and 21 variables.

1. **Fitting the Model**

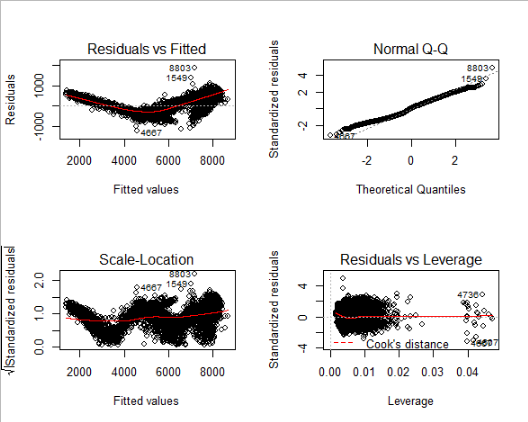
Next a linear regression model named LinearModel is built with the training data set, with CLV as the dependent variable and all other variables as independent variables. There is presence of insignificant variables. Iteration was run seven times, removing insignificant variables. The final model consists of CLV as the dependent variable. The independent variables consist of as follows:



R square value is 0.9621 and adjusted R square value is 0.9619, F statistic is 3842 on 28 and 4236 d.f. R square determines the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. Adjusted R-squared measures the variation in the dependent variable (or target), explained by only the features which are helpful in making predictions. Thus, it can be said that the model bult in the train data set explains about 96.2% of the variance in dependent variable.

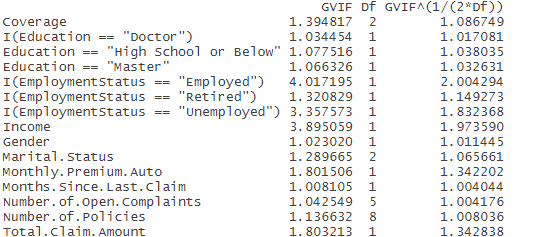
1. **Running diagnostic tests**

First, fitted values or y- hat values were found and included into the train.data as variable “pred”, and the Final model is plotted, giving the following charts:



Next MAPE is calculated, its value is 0.0726 or 7.26%, which is less than 10 %. Value of MAPE being less than 10% is preferred for a good model.

Next, presence of multicollinearity is checked using the Variable Inflation factor or vif function. Value of vif should be within 2 for multicollinearity to not exist. The vif values of the significant variables are as follows:



Vif values of variables are more or less within 2, so we can say multicollinearity is not present.

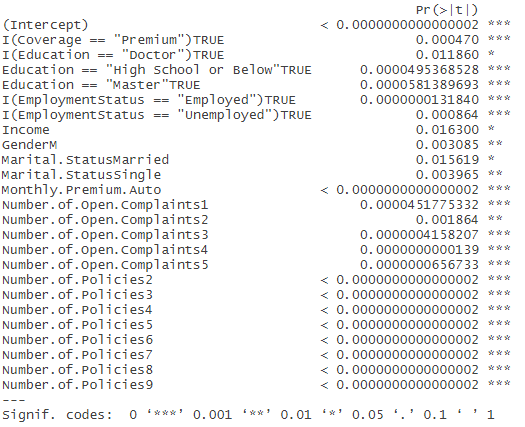
Next to test for autocorrelation, the Durbin -Watson test is done, D-W statistic was 2.01 and p value is 0.698, thus the null hypothesis (i.e. no autocorrelation present) is accepted.

Next to test for homoscedasticity, Breusch- Pagan test is done, p value came as significant, thus, null hypothesis, i.e. there is presence of constant variance in the model, is rejected. Thus, there is non-constant variance in the model, indicating heteroscedasticity.

Finally, to test normality in the model, Anderson- Darling test is conducted, where p value came as significant, thus the null hypothesis, i.e. the errors are normally distributed, is rejected. The alternate hypothesis is accepted, i.e. errors are not normally distributed.

1. **Model Validation**

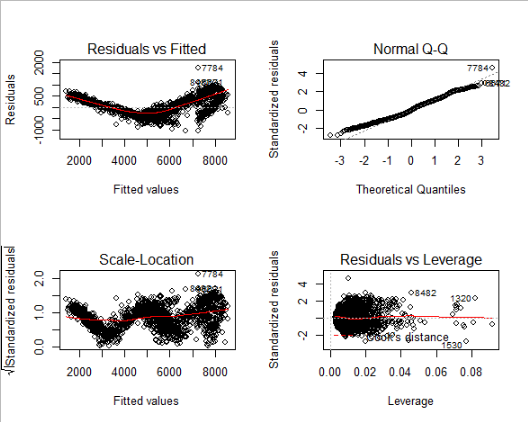
Next the model is validated in the test data. Model is run and few variables are found to be insignificant, the iteration is run four times, removing the insignificant variables, reaching the final model. The final model consists of CLV as the dependent variable, and the following as the significant independent variables:



Multiple R-squared is 0.9631, Adjusted R-squared is 0. 9626, F statistic is 1963 on 24 and 1804 DF.

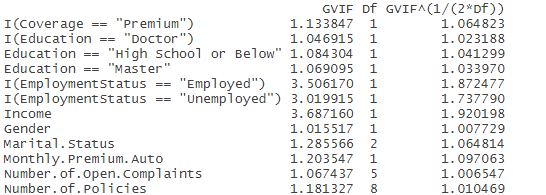
Thus, it can be said that the final model explains about 96.3% of the variance in the dependent variable.

Fitted values or y- hat values are found and included into the test.data as variable “pred”, and the model was plotted, giving the following results.



MAPE is calculated which is 0.07166 0r 7.16% which is less than 10%, thereby indicating that that are model is satisfactory.

Presence of multicollinearity is checked using the vif function. Following are the vif values for the significant variables:



Vif values for the variables are within 2, thus no multicollinearity present.

Next test for autocorrelation, the Durbin-Watson test is done, D-W statistic is 1.98, and p value is 0.68, thus, the null hypothesis is accepted, i.e. there is no presence of autocorrelation.

Next test for homoscedasticity, Breusch-Pagan test is done, p value came as significant, hence null hypothesis which is that there is presence of constant variance in the model, is rejected. Thus, there is non-constant variance in the model, indicating heteroscedasticity.

Finally, to test normality in the model, Anderson- Darling test is conducted, where p value came as significant, thus the null hypothesis, i.e. the errors are normally distributed, is rejected. The alternate hypothesis is accepted, i.e. errors are not normally distributed.

Coefficient values for the different significant independent variables in the final model are:

|  |  |
| --- | --- |
| **(Intercept)** | -1391.264174055 |
| **I(Coverage == "Premium")TRUE** | 147.641432857 |
| **I(Education == "Doctor")TRUE** | 114.583668210 |
| **Education == "High School or Below"TRUE** | 84.387001365 |
| **Education == "Master"TRUE** | 138.095244118 |
| **I(EmploymentStatus == "Employed")TRUE** | 196.397205343 |
| **I(EmploymentStatus == "Unemployed")TRUE** | -116.263025826 |
| **Income** | 0.001595689 |
| **GenderM** | -54.076311500 |
| **Marital.StatusMarried** | 64.575598670 |
| **Marital.StatusSingle** | -89.189743707 |
| **Monthly.Premium.Auto** | 52.737574483 |
| **Number.of.Open.Complaints1** | -124.107409062 |
| **Number.of.Open.Complaints2** | -134.167631475 |
| **Number.of.Open.Complaints3** | -260.730544756 |
| **Number.of.Open.Complaints4** | -465.735411398 |
| **Number.of.Open.Complaints5** | -547.953927309 |
| **Number.of.Policies2** | 5490.241064289 |
| **Number.of.Policies3** | 3192.956304350 |
| **Number.of.Policies4** | 3165.787550807 |
| **Number.of.Policies5** | 3165.447293397 |
| **Number.of.Policies6** | 3112.292427240 |
| **Number.of.Policies7** | 3085.888772366 |
| **Number.of.Policies8** | 3261.625055570 |
| **Number.of.Policies9** | 3197.429987130 |

**Business interpretation of significant independent Variables:**

For Continuous independent variables, coefficient value can be interpreted as the amount by which the dependent value will rise or fall (depending on the positive or negative value of the coefficient), with 1 unit increase in the value of the independent variable.

For Categorical independent variables, dummy variables represent various sub categories within the variable. For every categorical variable, a certain sub-category is not represented explicitly by a dummy variable, which is termed as the reference group. In analysis, each dummy variable is compared with the reference group. For example, a positive regression coefficient means value of the dependent variable (CLV for our example) is higher for the corresponding dummy variable, compared to the reference group of the concerned categorical variable. For a negative regression coefficient value, dependent variable (CLV for our example) is lower for the corresponding dummy variable, compared to the reference group of the concerned categorical variable.

Variables having positive coefficient value, thus having a positive relationship with CLV

* **Coverage** **Premium** (Reference group: Coverage Basic)
* **Education: Doctor, Education: High School or Below, Education:** Master (Reference group: Education Bachelor)
* **Employment Status: Employed** (Reference group: Employment Status: Disabled)
* **Income**
* **Marital Status:** **Married** (Reference group: Marital Status: Divorced)
* **Monthly Premium Auto**
* **Number of Policies** (Reference group: Number of Policies=1)

Variables having negative coefficient value, or a negative relationship with CLV

* **Employment Status: Unemployed** (Reference group: Employment Status: Disabled)
* **Gender Male** (Reference group: Gender Female)
* **Marital Status: Single** (Reference group: Marital Status Divorced)
* **Number of Open Complaints** (Reference group: Number of Complaints=0)

**Conclusion**:

**Thus, our final model can be deemed as a good model, explaining about 96.3% of the variance in the dependent variable; Customer Lifetime Value, with a MAPE of 7.16%. However, assumptions of the linear regression model like homoscedasticity and normality failed to hold true.**