**Approach:**

This is an Insurance company data and we are intended to frame a regression model that can give us a regressed prediction of Customer Lifetime Value which is our Target variable such that we can predict which of the customers out of 9134 customers are the most profitable ones to this insurance firm. For this first we have to check the data i.e. data type of each variable. Then if we found any categorical variable then we need to factor it first. After that we need to check whether outlier is present or not. If outlier is present then we have to remove it. After removing certain outliers, we have to split the data into two parts, one is train data and other one is test data. Then we build the model and check the accuracy by running some tests on this model. Finally, we predict the values and suggest some business recommendations.

Certain steps are to be needed to build the model.

**Step - 1: Data manipulation:** After importing the csv file we need to check the structure and dimension of the data and also replace the target variable name “Customer Lifetime Value” by “clv”. We convert two variables number of open complaints and no of policies into factors.

**Code:** colnames(dg)[which(names(dg)=="Customer.Lifetime.Value")]="clv"

str(dg) and dim(dg).

dg$Number.of.Open.Complaints <- as.factor(dg$Number.of.Open.Complaints)

dg$Number.of.Policies <- as.factor(dg$Number.of.Policies)

Outcome & interpretation: Here we have 9134 observations and 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 ...

$ clv : 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 : Factor w/ 6 levels "0","1","2","3",..: 1 1 1 1 1 1 1 1 1 1 ...

$ Number.of.Policies : Factor w/ 9 levels "1","2","3","4",..: 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 ...

Customer Lifetime Value is the total revenue the client will derive from their entire relationship with a customer. This is the dependent variable in our dataset.

**Step 2:** Now we need to check whether there exist any missing values or not.

Code: as.data.frame(colSums(is.na(dg)))

Outcome & interpretation:

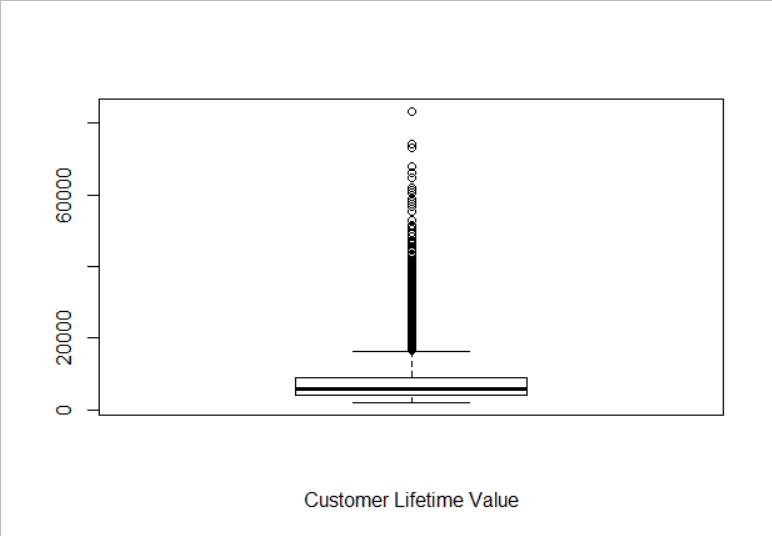
|  |
| --- |
| colSums(is.na(data3))  Customer 0  State 0  clv 0  Response 0  Coverage 0  Education 0  Effective.To.Date 0  EmploymentStatus 0  Gender 0  Income 0  Location.Code 0  Marital.Status 0  Monthly.Premium.Auto 0  Months.Since.Last.Claim 0  Months.Since.Policy.Inception 0  Number.of.Open.Complaints 0  Number.of.Policies 0  Policy.Type 0  Policy 0  Renew.Offer.Type 0  Sales.Channel 0  Total.Claim.Amount 0  Vehicle.Class 0  Vehicle.Size 0 |
|  |

There are no null values in the data. So, there is no need to remove it or replace it by mean, median or mode.

**Data exploration: Step – 3:** Now we need to check outliers.

Code: quantile(dg$clv, seq(0,1,.05))

boxplot(dg$clv, xlab="Customer Lifetime Value")

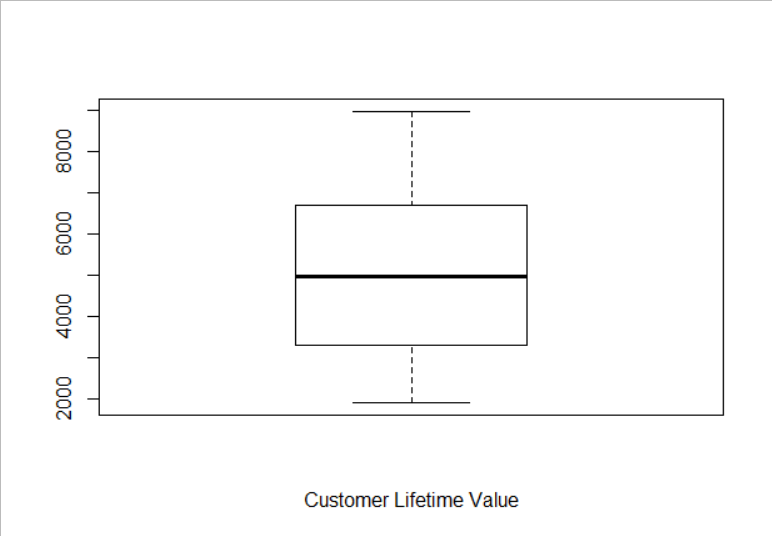


**Outcome & interpretation**: Here we can see that after 75% the value of clv is increasing and at 100% the value is extremely high, also from the boxplot we confidently say that outlier exists. So, we need to remove them. Otherwise it may create some distortion when we build the model.

Code: dg1=dg[dg$clv <9000, ]

quantile(dg1$clv, seq(0,1,0.05))

boxplot(dg1$clv, xlab = "Customer Lifetime Value")



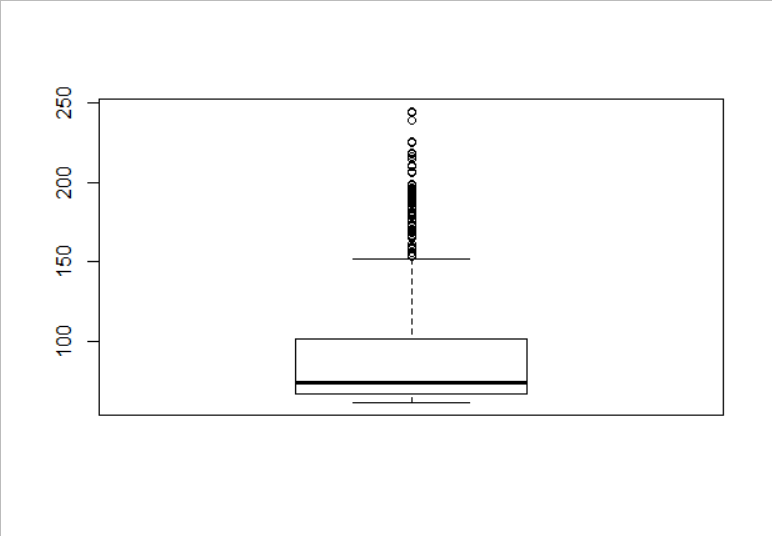
Here we can see that outliers are removed totally. Basically, some of the clv values are vertically distant from the others, that’s why it creates outliers. If we keep the outliers then it will create an impact on the data and the model will not be the best model.

**Step – 4:** We also check outliers for the independent variable and we notice that monthly premium auto has outliers so we need to remove it.

Code: quantile(dg1$Monthly.Premium.Auto, seq(0,1,.05))

boxplot(dg1$Monthly.Premium.Auto)

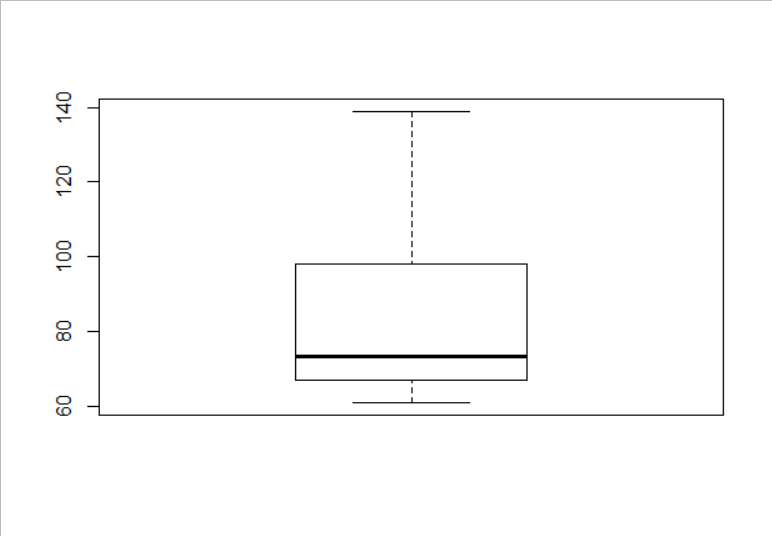
**Outcome & interpretation:**



But here we see that outliers still exists. So, we need to remove them again.

Code: dg2 <- dg1[dg1$Monthly.Premium.Auto<140,]

boxplot(dg2$Monthly.Premium.Auto)

Here we can see that outliers are removed totally. Basically, some of the monthly premium values are vertically distant from the others, that’s why it creates outliers. If we keep the outliers then it will create an impact on the data and the model will not be the best model.

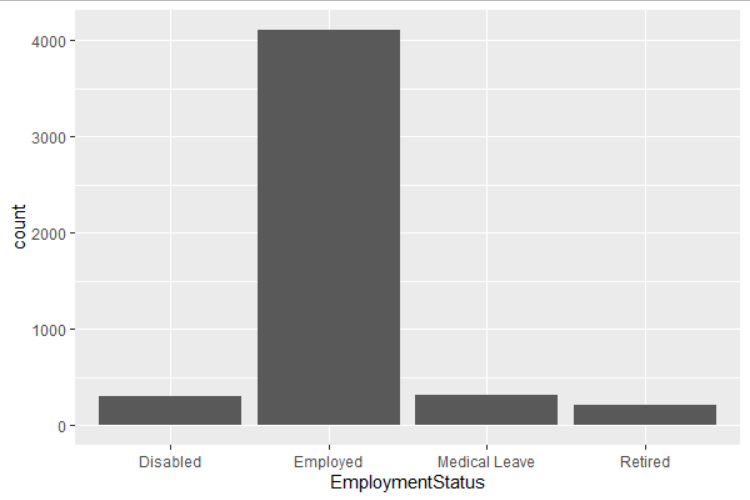
**Step – 5:** Here we also noticed that lot of customers are unemployed and their income is zero. If the earning for a customer is zero then it is meaningless to keep them in the model as they don’t have the ability to buy a policy. So, we filter out income greater than zero.

Code: dg3 <- filter(dg2, Income>0)

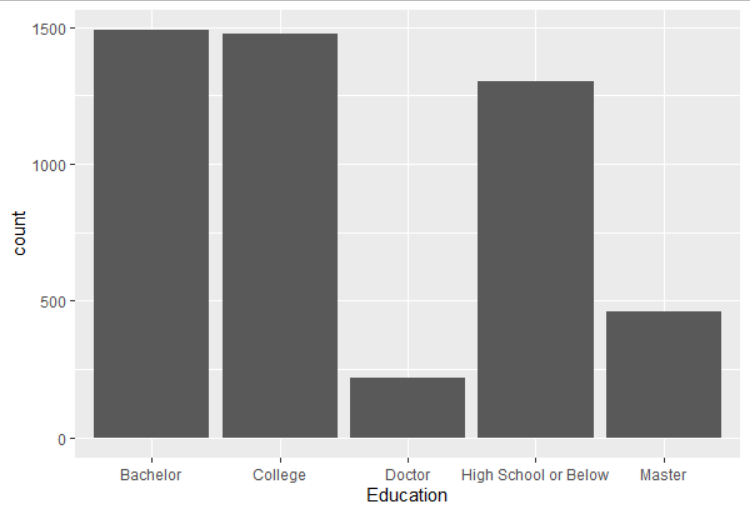
**Step – 6:** At the very first step we drop those variables which doesn’t play a crucial role in the model analysis. So, we drop customer id, state and date column.

Code: dg4= dg3[, !(colnames(dg) %in% c("Customer","State","Effective.To.Date"))]

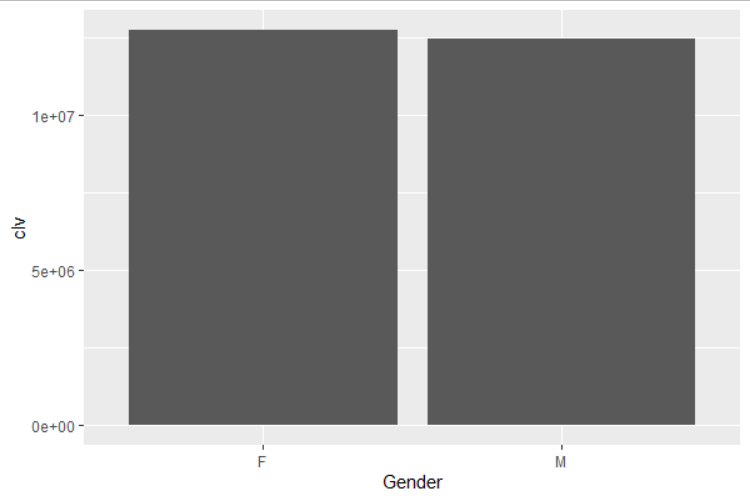
Outcome and interpretation: After removing certain outliers and dropping 3 columns we have 21 variables 4941 observation. We observed that these three variables have no impact on the target variable.



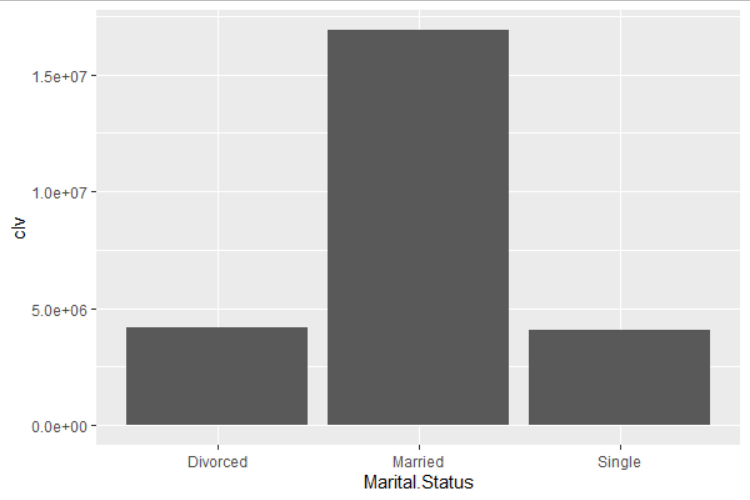
In this figure we plot employment status on x axis and the graph shows that there are higher number of employed people contributing to clv.



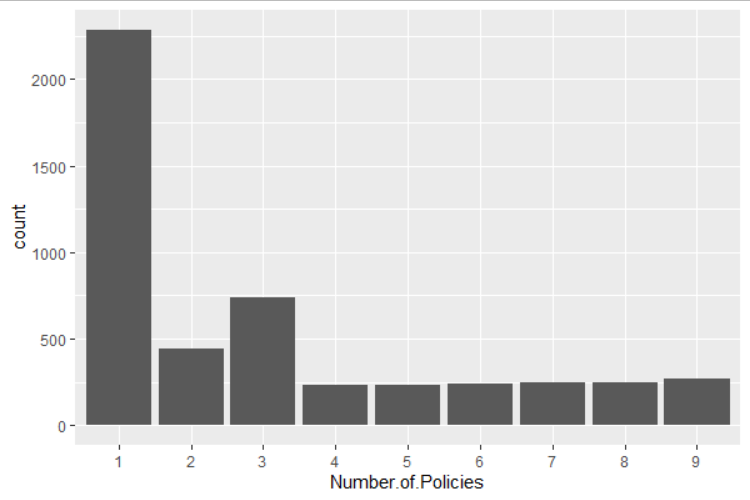
In this figure we plot the education on x axis. Here no of people in bachelor and college are almost same followed by the number of people in higher secondary contributing to clv.



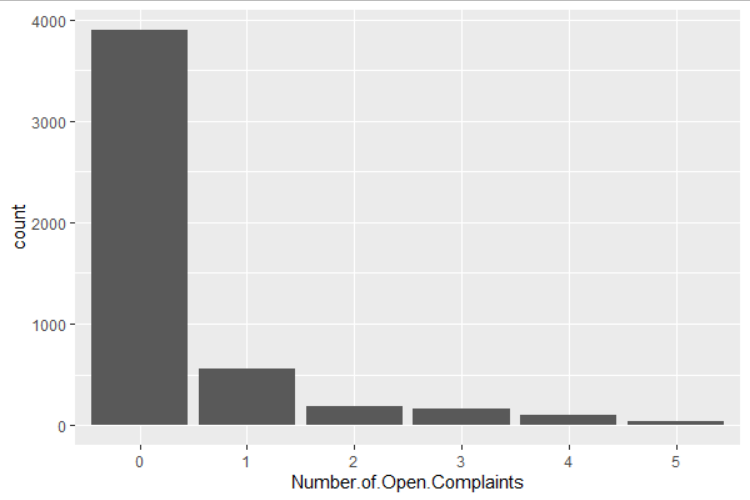
In this figure we plot gender on the x axis and customer Lifetime Value on the y axis. From this bar diagram it is clearly shown that the average customer lifetime value for both male and female are almost same.



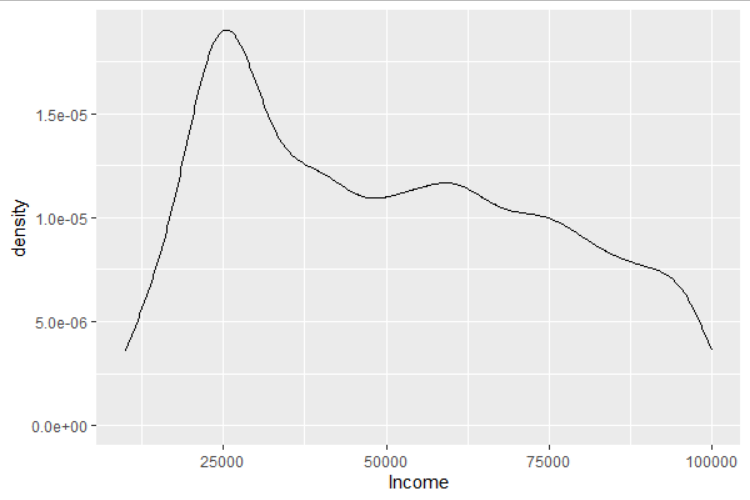
In this figure we plot marital status on x axis and Customer Lifetime value on y axis. Here the average customer lifetime value for married people is high. Single or bachelor people have a slightly high average lifetime value than the divorced ones.



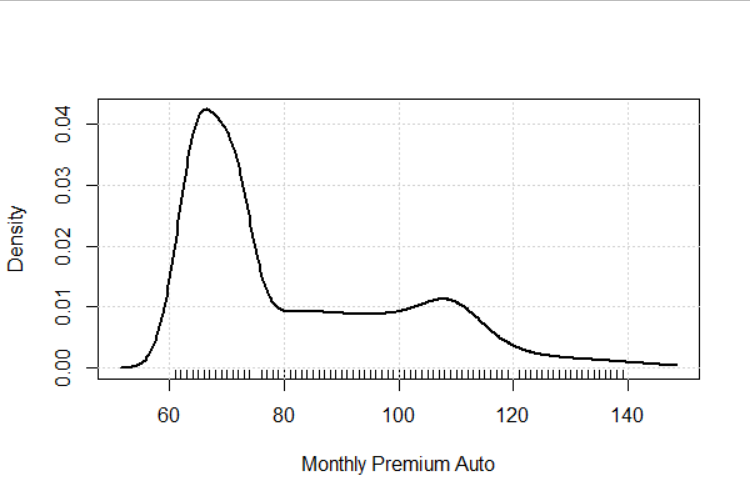
This figure shows the graphical display of policy variable. Also, we can see that there are higher number of people who take lesser number of policies.



This figure shows that number of open complaints is decreasing. There is maximum amount of people who have no complaints. And it is good for the company having very few complaints.



In this figure we plot income on x axis and frequency density on the y axis. Here we can see a high pick around 25000. after that the income levels are falling. From this we can conclude that the distribution is not symmetric.



In the above figure we plot monthly premium on the x axis and frequency density on the y axis. At the beginning it is showing a high peak after that the premium amount is falling. Also, the distribution of monthly premium is not symmetric. It means lesser premium attracts more people than the higher one. Also, a large no. of people gives lesser amount of premium.

**Model building:**

Here the target variable is continuous & we run a multiple linear regression model and want to predict which of the customers out of 4941 customers are the most profitable ones for the company. For this we first split the dataset into two parts. One is train data and other one is test data. The train data carry 70% of the data and the remaining 30% is the test data. After splitting the data, we build a model and to check the accuracy we run this model on the test dataset.

Code: sample = sample.split(dg4$clv,SplitRatio = 0.70)

train4 =subset(dg4, sample ==TRUE)

str(train4)

dim(train4)

test4=subset(dg4, sample==FALSE)

str(test4)

dim(test4)

Outcome & interpretation: here in the train data we have 3458 observations and in the test data we have 1438 observations. We split the dataset for better understanding of the analysis. In train dataset we build the model and in test dataset we make predictions based on the model. To get the final model we use trial & error method.

**Hypothesis:**

**H0: the variable has no effect on clv.**

**H1: the variable has effect(positive or negative) on clv.**

**If p-value less than significant level, i.e. if p-value is close to zero we should reject null hypothesis and accept alternative one.**

**1st model:**

Code: L0<- lm(clv~., data=train4)

summary(L0)

**Outcome & interpretation:** Call:

lm(formula = clv ~ ., data = train4)

Residuals:

Min 1Q Median 3Q Max

-958.06 -301.46 -1.41 250.03 1036.80

Coefficients: (2 not defined because of singularities)

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

(Intercept) -1.294e+03 9.988e+01 -12.951 < 2e-16 \*\*\*

ResponseYes -3.419e+01 1.867e+01 -1.831 0.067163 .

CoverageExtended -1.773e+01 2.856e+01 -0.621 0.534913

CoveragePremium 1.099e+02 5.847e+01 1.879 0.060364 .

EducationCollege 2.233e-01 1.611e+01 0.014 0.988943

EducationDoctor 4.841e+01 3.210e+01 1.508 0.131575

EducationHigh School or Below 5.394e+01 1.659e+01 3.251 0.001162 \*\*

EducationMaster 6.730e+01 2.358e+01 2.854 0.004348 \*\*

EmploymentStatusEmployed 1.407e+02 2.892e+01 4.865 1.20e-06 \*\*\*

EmploymentStatusMedical Leave -6.868e+01 3.573e+01 -1.922 0.054662 .

EmploymentStatusRetired -1.619e+02 4.199e+01 -3.856 0.000118 \*\*\*

GenderM -1.943e+01 1.253e+01 -1.551 0.121012

Income 2.092e-03 3.145e-04 6.652 3.36e-11 \*\*\*

Location.CodeSuburban -4.314e+01 2.888e+01 -1.494 0.135365

Location.CodeUrban -3.370e+01 2.296e+01 -1.468 0.142282

Marital.StatusMarried 4.224e+01 1.725e+01 2.448 0.014397 \*

Marital.StatusSingle -9.166e+01 2.144e+01 -4.275 1.96e-05 \*\*\*

Monthly.Premium.Auto 5.317e+01 1.268e+00 41.929 < 2e-16 \*\*\*

Months.Since.Last.Claim -1.538e+00 6.319e-01 -2.434 0.014969 \*

Months.Since.Policy.Inception 9.552e-02 2.247e-01 0.425 0.670814

Number.of.Open.Complaints1 -3.560e+01 2.002e+01 -1.778 0.075414 .

Number.of.Open.Complaints2 -6.772e+01 3.321e+01 -2.039 0.041502 \*

Number.of.Open.Complaints3 -1.382e+02 3.454e+01 -4.001 6.45e-05 \*\*\*

Number.of.Open.Complaints4 -3.447e+02 4.555e+01 -7.567 4.87e-14 \*\*\*

Number.of.Open.Complaints5 -2.713e+02 8.658e+01 -3.134 0.001740 \*\*

Number.of.Policies2 5.542e+03 2.357e+01 235.161 < 2e-16 \*\*\*

Number.of.Policies3 3.233e+03 1.879e+01 172.082 < 2e-16 \*\*\*

Number.of.Policies4 3.211e+03 3.008e+01 106.737 < 2e-16 \*\*\*

Number.of.Policies5 3.138e+03 3.040e+01 103.226 < 2e-16 \*\*\*

Number.of.Policies6 3.168e+03 3.004e+01 105.456 < 2e-16 \*\*\*

Number.of.Policies7 3.264e+03 2.986e+01 109.309 < 2e-16 \*\*\*

Number.of.Policies8 3.243e+03 2.943e+01 110.174 < 2e-16 \*\*\*

Number.of.Policies9 3.258e+03 2.821e+01 115.492 < 2e-16 \*\*\*

Policy.TypePersonal Auto -7.343e+01 3.356e+01 -2.188 0.028710 \*

Policy.TypeSpecial Auto -1.038e+02 6.028e+01 -1.722 0.085114 .

PolicyCorporate L2 -6.272e+01 3.983e+01 -1.575 0.115415

PolicyCorporate L3 -6.646e+01 3.676e+01 -1.808 0.070701 .

PolicyPersonal L1 -6.294e+00 1.967e+01 -0.320 0.749044

PolicyPersonal L2 5.941e+00 1.645e+01 0.361 0.718000

PolicyPersonal L3 NA NA NA NA

PolicySpecial L1 1.148e+02 8.683e+01 1.322 0.186152

PolicySpecial L2 3.583e+01 7.069e+01 0.507 0.612300

PolicySpecial L3 NA NA NA NA

Renew.Offer.TypeOffer2 1.187e+01 1.600e+01 0.742 0.457993

Renew.Offer.TypeOffer3 -2.582e-02 1.961e+01 -0.001 0.998949

Renew.Offer.TypeOffer4 -1.134e+01 2.189e+01 -0.518 0.604488

Sales.ChannelBranch 4.229e+00 1.561e+01 0.271 0.786512

Sales.ChannelCall Center -1.459e+01 1.766e+01 -0.826 0.408764

Sales.ChannelWeb 1.019e+01 1.964e+01 0.519 0.603983

Total.Claim.Amount -1.874e-02 7.098e-02 -0.264 0.791774

Vehicle.ClassSports Car -6.465e+01 6.338e+01 -1.020 0.307800

Vehicle.ClassSUV -1.082e+01 5.484e+01 -0.197 0.843574

Vehicle.ClassTwo-Door Car -4.097e-02 1.517e+01 -0.003 0.997845

Vehicle.SizeMedsize 1.041e+01 2.086e+01 0.499 0.617937

Vehicle.SizeSmall 1.425e+01 2.450e+01 0.582 0.560849

---

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

Residual standard error: 362.9 on 3405 degrees of freedom

Multiple R-squared: 0.9667, Adjusted R-squared: 0.9662

F-statistic: 1903 on 52 and 3405 DF, p-value: < 2.2e-16

Here maximum variables have a p-value which is greater than significant level i.e. 95% or 99%. So, we have to accept the null hypothesis for those variables and remove them one by one as they have no impact on clv. Here we can see that R2 is 0.96 it means the explanatory variables can explain 96% of the dependent variable. Also, there are lots of variables that are insignificant. we need to drop them one by one so that we can get the best fitted model.

**2nd model:**

Code: L1<- lm(clv~Response+ Coverage+ Education+ EmploymentStatus+ Gender+ Income+Location.Code+Marital.Status+Monthly.Premium.Auto+Months.Since.Last.Claim+ Months.Since.Policy.Inception+Number.of.Open.Complaints+ Number.of.Policies + Renew.Offer.Type+ Sales.Channel+ Total.Claim.Amount+ Vehicle.Class+ Vehicle.Size, data=train4)

summary(L1)

**Outcome & interpretation:** In the previous model we had seen that not all the variables were significant. First, we have to remove two insignificant variables i.e. policy and policy type.

Call:

lm(formula = clv ~ Response + Coverage + Education + EmploymentStatus +

Gender + Income +Location.Code + Marital.Status +Monthly.Premium.Auto +

Months.Since.Last.Claim + Months.Since.Policy.Inception + Number.of.Open.Complaints + Number.of.Policies + Renew.Offer.Type + Sales.Channel + Total.Claim.Amount + Vehicle.Class + Vehicle.Size, data = train4)

Residuals:

Min 1Q Median 3Q Max

-954.28 -301.92 1.57 248.97 1024.13

Coefficients:

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

(Intercept) -1.358e+03 9.467e+01 -14.350 < 2e-16 \*\*\*

ResponseYes -3.352e+01 1.865e+01 -1.798 0.072275 .

CoverageExtended -1.652e+01 2.854e+01 -0.579 0.562682

CoveragePremium 1.101e+02 5.840e+01 1.886 0.059423 .

EducationCollege 3.426e-01 1.609e+01 0.021 0.983017

EducationDoctor 4.708e+01 3.205e+01 1.469 0.141935

EducationHigh School or Below 5.479e+01 1.658e+01 3.304 0.000963 \*\*\*

EducationMaster 6.857e+01 2.356e+01 2.910 0.003637 \*\*

EmploymentStatusEmployed 1.396e+02 2.885e+01 4.837 1.37e-06 \*\*\*

EmploymentStatusMedical Leave -7.051e+01 3.567e+01 -1.977 0.048151 \*

EmploymentStatusRetired -1.658e+02 4.190e+01 -3.957 7.75e-05 \*\*\*

GenderM -1.926e+01 1.251e+01 -1.539 0.123852

Income 2.074e-03 3.141e-04 6.602 4.69e-11 \*\*\*

Location.CodeSuburban -4.278e+01 2.883e+01 -1.484 0.137915

Location.CodeUrban -3.254e+01 2.292e+01 -1.420 0.155728

Marital.StatusMarried 4.204e+01 1.722e+01 2.441 0.014714 \*

Marital.StatusSingle -9.267e+01 2.141e+01 -4.328 1.55e-05 \*\*\*

Monthly.Premium.Auto 5.316e+01 1.267e+00 41.962 < 2e-16 \*\*\*

Months.Since.Last.Claim -1.533e+00 6.311e-01 -2.429 0.015209 \*

Months.Since.Policy.Inception 1.008e-01 2.245e-01 0.449 0.653534

Number.of.Open.Complaints1 -3.595e+01 2.000e+01 -1.797 0.072381 .

Number.of.Open.Complaints2 -6.796e+01 3.317e+01 -2.049 0.040522 \*

Number.of.Open.Complaints3 -1.369e+02 3.445e+01 -3.975 7.18e-05 \*\*\*

Number.of.Open.Complaints4 -3.438e+02 4.545e+01 -7.564 5.01e-14 \*\*\*

Number.of.Open.Complaints5 -2.774e+02 8.649e+01 -3.207 0.001352 \*\*

Number.of.Policies2 5.543e+03 2.354e+01 235.462 < 2e-16 \*\*\*

Number.of.Policies3 3.233e+03 1.878e+01 172.153 < 2e-16 \*\*\*

Number.of.Policies4 3.211e+03 3.004e+01 106.865 < 2e-16 \*\*\*

Number.of.Policies5 3.138e+03 3.034e+01 103.429 < 2e-16 \*\*\*

Number.of.Policies6 3.168e+03 3.002e+01 105.505 < 2e-16 \*\*\*

Number.of.Policies7 3.262e+03 2.980e+01 109.458 < 2e-16 \*\*\*

Number.of.Policies8 3.242e+03 2.940e+01 110.274 < 2e-16 \*\*\*

Number.of.Policies9 3.256e+03 2.815e+01 115.652 < 2e-16 \*\*\*

Renew.Offer.TypeOffer2 1.045e+01 1.596e+01 0.655 0.512725

Renew.Offer.TypeOffer3 -8.517e-01 1.958e+01 -0.043 0.965307

Renew.Offer.TypeOffer4 -1.065e+01 2.186e+01 -0.487 0.626041

Sales.ChannelBranch 3.571e+00 1.560e+01 0.229 0.818986

Sales.ChannelCall Center -1.462e+01 1.764e+01 -0.829 0.407079

Sales.ChannelWeb 1.156e+01 1.961e+01 0.590 0.555522

Total.Claim.Amount -1.962e-02 7.080e-02 -0.277 0.781738

Vehicle.ClassSports Car -6.476e+01 6.331e+01 -1.023 0.306456

Vehicle.ClassSUV -1.040e+01 5.475e+01 -0.190 0.849353

Vehicle.ClassTwo-Door Car -5.680e-01 1.515e+01 -0.037 0.970096

Vehicle.SizeMedsize 1.065e+01 2.084e+01 0.511 0.609318

Vehicle.SizeSmall 1.396e+01 2.449e+01 0.570 0.568739

---

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

Residual standard error: 362.9 on 3413 degrees of freedom

Multiple R-squared: 0.9667, Adjusted R-squared: 0.9662

F-statistic: 2249 on 44 and 3413 DF, p-value: < 2.2e-16

Here maximum variables have a p-value which is greater than significant level i.e. 95% or 99%. So, we have to accept the null hypothesis for those variables and remove them one by one as they have no impact on clv. Here also not all the variables are significant. It means they have no impact on clv. So again, we need to remove them.

**3rd model:**

Code: L2 <- lm(clv~Response+ Coverage+ Education+ EmploymentStatus+Gender+Income+Location.Code+Marital.Status+Monthly.Premium.Auto+Months.Since.Last.Claim+Months.Since.Policy.Inception+Number.of.Open.Complaints+Number.of.Policies +Renew.Offer.Type+Sales.Channel+Total.Claim.Amount, data=train4)

summary(L2)

**Outcome & interpretation:**

Call:

lm(formula = clv ~ Response + Coverage + Education + EmploymentStatus +

Gender + Income + Location.Code + Marital.Status + Monthly.Premium.Auto + Months.Since.Last.Claim + Months.Since.Policy.Inception + Number.of.Open.Complaints + Number.of.Policies + Renew.Offer.Type + Sales.Channel + Total.Claim.Amount, data = train4)

Residuals:

Min 1Q Median 3Q Max

-954.02 -300.81 0.13 249.89 1032.28

Coefficients:

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

(Intercept) -1.319e+03 5.258e+01 -25.091 < 2e-16 \*\*\*

ResponseYes -3.455e+01 1.861e+01 -1.856 0.06350 .

CoverageExtended -7.186e+00 1.592e+01 -0.451 0.65173

CoveragePremium 1.297e+02 3.115e+01 4.163 3.22e-05 \*\*\*

EducationCollege -4.463e-02 1.605e+01 -0.003 0.99778

EducationDoctor 4.770e+01 3.202e+01 1.490 0.13633

EducationHigh School or Below 5.439e+01 1.656e+01 3.285 0.00103 \*\*

EducationMaster 6.878e+01 2.349e+01 2.928 0.00343 \*\*

EmploymentStatusEmployed 1.415e+02 2.881e+01 4.912 9.46e-07 \*\*\*

EmploymentStatusMedical Leave -6.807e+01 3.561e+01 -1.912 0.05599 .

EmploymentStatusRetired -1.694e+02 4.177e+01 -4.055 5.12e-05 \*\*\*

GenderM -2.014e+01 1.248e+01 -1.614 0.10666

Income 2.083e-03 3.133e-04 6.648 3.44e-11 \*\*\*

Location.CodeSuburban -4.274e+01 2.871e+01 -1.489 0.13665

Location.CodeUrban -3.259e+01 2.289e+01 -1.424 0.15461

Marital.StatusMarried 4.123e+01 1.719e+01 2.399 0.01651 \*

Marital.StatusSingle -9.253e+01 2.139e+01 -4.327 1.56e-05 \*\*\*

Monthly.Premium.Auto 5.268e+01 4.827e-01 109.130 < 2e-16 \*\*\*

Months.Since.Last.Claim -1.527e+00 6.306e-01 -2.421 0.01553 \*

Months.Since.Policy.Inception 1.023e-01 2.242e-01 0.456 0.64814

Number.of.Open.Complaints1 -3.541e+01 1.997e+01 -1.773 0.07631 .

Number.of.Open.Complaints2 -6.956e+01 3.313e+01 -2.100 0.03584 \*

Number.of.Open.Complaints3 -1.363e+02 3.438e+01 -3.964 7.53e-05 \*\*\*

Number.of.Open.Complaints4 -3.451e+02 4.539e+01 -7.602 3.73e-14 \*\*\*

Number.of.Open.Complaints5 -2.760e+02 8.643e+01 -3.193 0.00142 \*\*

Number.of.Policies2 5.543e+03 2.349e+01 235.965 < 2e-16 \*\*\*

Number.of.Policies3 3.234e+03 1.874e+01 172.618 < 2e-16 \*\*\*

Number.of.Policies4 3.209e+03 3.002e+01 106.911 < 2e-16 \*\*\*

Number.of.Policies5 3.139e+03 3.029e+01 103.638 < 2e-16 \*\*\*

Number.of.Policies6 3.169e+03 2.994e+01 105.849 < 2e-16 \*\*\*

Number.of.Policies7 3.263e+03 2.976e+01 109.661 < 2e-16 \*\*\*

Number.of.Policies8 3.243e+03 2.936e+01 110.455 < 2e-16 \*\*\*

Number.of.Policies9 3.255e+03 2.812e+01 115.762 < 2e-16 \*\*\*

Renew.Offer.TypeOffer2 1.063e+01 1.594e+01 0.667 0.50500

Renew.Offer.TypeOffer3 -9.659e-01 1.957e+01 -0.049 0.96064

Renew.Offer.TypeOffer4 -1.040e+01 2.182e+01 -0.477 0.63370

Sales.ChannelBranch 3.959e+00 1.558e+01 0.254 0.79944

Sales.ChannelCall Center -1.444e+01 1.763e+01 -0.819 0.41264

Sales.ChannelWeb 1.141e+01 1.960e+01 0.582 0.56052

Total.Claim.Amount -1.500e-02 7.053e-02 -0.213 0.83156

---

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

Residual standard error: 362.8 on 3418 degrees of freedom

Multiple R-squared: 0.9666, Adjusted R-squared: 0.9663

F-statistic: 2540 on 39 and 3418 DF, p-value: < 2.2e-16

Here we can see that variables like total claim, vehicle class, vehicle size, renew offer,has a very high p-value, it means we have to accept the null hypothesis. As these variables has no effect on clv we have to remove them. In this step we remove vehicle class and vehicle size.

**4th model:**

Code:L3<-lm(clv~Response+ Coverage+ Education+ EmploymentStatus+ Gender+ Income+Location.Code+Marital.Status+Monthly.Premium.Auto+Months.Since.Last.Claim+Months.Since.Policy.Inception+Number.of.Open.Complaints+Number.of.Policies+Renew.Offer.Type, data=train4)

summary(L3)

**Outcome & interpretation**:

Call:

lm(formula = clv ~ Response + Coverage + Education + EmploymentStatus +

Gender + Income + Location.Code + Marital.Status + Monthly.Premium.Auto +

Months.Since.Last.Claim + Months.Since.Policy.Inception + Number.of.Open.Complaints + Number.of.Policies + Renew.Offer.Type, data = train4)

Residuals:

Min 1Q Median 3Q Max

-950.2 -301.4 -2.8 251.2 1024.6

Coefficients:

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

(Intercept) -1.316e+03 4.944e+01 -26.626 < 2e-16 \*\*\*

ResponseYes -3.493e+01 1.843e+01 -1.896 0.05810 .

CoverageExtended -7.572e+00 1.590e+01 -0.476 0.63386

CoveragePremium 1.294e+02 3.112e+01 4.157 3.30e-05 \*\*\*

EducationCollege 7.655e-01 1.601e+01 0.048 0.96188

EducationDoctor 4.712e+01 3.196e+01 1.474 0.14047

EducationHigh School or Below 5.456e+01 1.654e+01 3.299 0.00098 \*\*\*

EducationMaster 6.896e+01 2.346e+01 2.939 0.00331 \*\*

EmploymentStatusEmployed 1.425e+02 2.876e+01 4.956 7.54e-07 \*\*\*

EmploymentStatusMedical Leave -6.672e+01 3.555e+01 -1.877 0.06063 .

EmploymentStatusRetired -1.662e+02 4.165e+01 -3.990 6.75e-05 \*\*\*

GenderM -2.046e+01 1.246e+01 -1.642 0.10076

Income 2.082e-03 3.127e-04 6.657 3.24e-11 \*\*\*

Location.CodeSuburban -4.804e+01 1.623e+01 -2.960 0.00310 \*\*

Location.CodeUrban -3.534e+01 1.821e+01 -1.941 0.05240 .

Marital.StatusMarried 4.126e+01 1.716e+01 2.404 0.01628 \*

Marital.StatusSingle -9.245e+01 2.137e+01 -4.326 1.56e-05 \*\*\*

Monthly.Premium.Auto 5.262e+01 3.897e-01 135.029 < 2e-16 \*\*\*

Months.Since.Last.Claim -1.527e+00 6.302e-01 -2.424 0.01542 \*

Months.Since.Policy.Inception 1.049e-01 2.241e-01 0.468 0.63985

Number.of.Open.Complaints1 -3.500e+01 1.993e+01 -1.756 0.07919 .

Number.of.Open.Complaints2 -7.015e+01 3.310e+01 -2.119 0.03415 \*

Number.of.Open.Complaints3 -1.360e+02 3.437e+01 -3.957 7.75e-05 \*\*\*

Number.of.Open.Complaints4 -3.453e+02 4.537e+01 -7.610 3.52e-14 \*\*\*

Number.of.Open.Complaints5 -2.760e+02 8.636e+01 -3.196 0.00141 \*\*

Number.of.Policies2 5.543e+03 2.345e+01 236.400 < 2e-16 \*\*\*

Number.of.Policies3 3.234e+03 1.870e+01 172.904 < 2e-16 \*\*\*

Number.of.Policies4 3.209e+03 2.998e+01 107.053 < 2e-16 \*\*\*

Number.of.Policies5 3.137e+03 3.024e+01 103.762 < 2e-16 \*\*\*

Number.of.Policies6 3.167e+03 2.990e+01 105.932 < 2e-16 \*\*\*

Number.of.Policies7 3.263e+03 2.970e+01 109.844 < 2e-16 \*\*\*

Number.of.Policies8 3.241e+03 2.928e+01 110.711 < 2e-16 \*\*\*

Number.of.Policies9 3.255e+03 2.807e+01 115.971 < 2e-16 \*\*\*

Renew.Offer.TypeOffer2 1.011e+01 1.560e+01 0.648 0.51710

Renew.Offer.TypeOffer3 -9.355e-01 1.943e+01 -0.048 0.96160

Renew.Offer.TypeOffer4 -1.142e+01 2.163e+01 -0.528 0.59756

---

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

Residual standard error: 362.6 on 3422 degrees of freedom

Multiple R-squared: 0.9666, Adjusted R-squared: 0.9663

F-statistic: 2832 on 35 and 3422 DF, p-value: < 2.2e-16

Here we remove those insignificant variables who have a high p-value. In this step we remove two variable sales channel and total claim amount. But some insignificant variables still exist in the model. So, we need to remove them.

**5th model:**

Code:L4<-lm(clv~Response+Coverage+Education+EmploymentStatus+Gender+Income+ Location.Code+Marital.Status+Monthly.Premium.Auto+Months.Since.Last.Claim+ Number.of.Open.Complaints+Number.of.Policies, data=train4)

summary(L4)

**Outcome & interpretation**:

Call:

lm(formula = clv ~ Response + Coverage + Education + EmploymentStatus +Gender + Income + Location.Code + Marital.Status + Monthly.Premium.Auto +Months.Since.Last.Claim + Number.of.Open.Complaints + Number.of.Policies, data = train4)

Residuals:

Min 1Q Median 3Q Max

-952.41 -301.73 -4.33 248.98 1023.36

Coefficients:

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

(Intercept) -1.312e+03 4.711e+01 -27.842 < 2e-16 \*\*\*

ResponseYes -2.952e+01 1.769e+01 -1.669 0.095197 .

CoverageExtended -7.408e+00 1.588e+01 -0.467 0.640810

CoveragePremium 1.294e+02 3.111e+01 4.160 3.26e-05 \*\*\*

EducationCollege 4.435e-01 1.599e+01 0.028 0.977871

EducationDoctor 4.771e+01 3.187e+01 1.497 0.134577

EducationHigh School or Below 5.434e+01 1.649e+01 3.296 0.000991 \*\*\*

EducationMaster 6.826e+01 2.340e+01 2.917 0.003552 \*\*

EmploymentStatusEmployed 1.426e+02 2.865e+01 4.977 6.79e-07 \*\*\*

EmploymentStatusMedical Leave -6.628e+01 3.552e+01 -1.866 0.062076 .

EmploymentStatusRetired -1.682e+02 4.158e+01 -4.046 5.33e-05 \*\*\*

GenderM -2.023e+01 1.244e+01 -1.626 0.103974

Income 2.107e-03 3.101e-04 6.795 1.27e-11 \*\*\*

Location.CodeSuburban -4.777e+01 1.621e+01 -2.946 0.003238 \*\*

Location.CodeUrban -3.455e+01 1.818e+01 -1.901 0.057445 .

Marital.StatusMarried 4.209e+01 1.700e+01 2.477 0.013314 \*

Marital.StatusSingle -9.225e+01 2.135e+01 -4.321 1.59e-05 \*\*\*

Monthly.Premium.Auto 5.261e+01 3.873e-01 135.858 < 2e-16 \*\*\*

Months.Since.Last.Claim -1.517e+00 6.267e-01 -2.421 0.015536 \*

Number.of.Open.Complaints1 -3.560e+01 1.981e+01 -1.797 0.072396 .

Number.of.Open.Complaints2 -7.055e+01 3.298e+01 -2.139 0.032510 \*

Number.of.Open.Complaints3 -1.361e+02 3.433e+01 -3.964 7.52e-05 \*\*\*

Number.of.Open.Complaints4 -3.444e+02 4.534e+01 -7.597 3.87e-14 \*\*\*

Number.of.Open.Complaints5 -2.773e+02 8.625e+01 -3.215 0.001316 \*\*

Number.of.Policies2 5.542e+03 2.335e+01 237.368 < 2e-16 \*\*\*

Number.of.Policies3 3.233e+03 1.860e+01 173.779 < 2e-16 \*\*\*

Number.of.Policies4 3.210e+03 2.992e+01 107.256 < 2e-16 \*\*\*

Number.of.Policies5 3.137e+03 3.016e+01 104.015 < 2e-16 \*\*\*

Number.of.Policies6 3.167e+03 2.977e+01 106.411 < 2e-16 \*\*\*

Number.of.Policies7 3.262e+03 2.963e+01 110.118 < 2e-16 \*\*\*

Number.of.Policies8 3.242e+03 2.921e+01 110.991 < 2e-16 \*\*\*

Number.of.Policies9 3.255e+03 2.789e+01 116.690 < 2e-16 \*\*\*

---

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

Residual standard error: 362.5 on 3426 degrees of freedom

Multiple R-squared: 0.9666, Adjusted R-squared: 0.9663

F-statistic: 3199 on 31 and 3426 DF, p-value: < 2.2e-16

Here education, employment status, renew offer type and coverage are the categorical variables, number of open complaints and number of policies. There are several categories under them. Some of them are significant and some are not. Under education “college”, “doctor” and have a p value 0.97 and 0.13 and under employment status the category “medical leave” has a p-value 0.657418. So, all these three variables have a higher p value greater than the level of significant. The p value mainly shows the probability of rejecting the null hypothesis. The more the p value become closer to 0 the more will be the chance of rejecting null hypothesis i.e. there are no effects on clv. Here higher p values simply indicate that we have to accept null hypothesis and reject the alternate one. It means these three variables has no effect on clv. So, again we need to remove them.

**6th Model:**

Code: L5 <- lm(clv~Response+I(Coverage == "Premium")+I(Education == "High School or Below")+ I(Education == "Master")+EmploymentStatus+Gender+Income+Marital.Status+ Monthly.Premium.Auto+ Number.of.Open.Complaints+Number.of.Policies, data=train4)

summary(L5)

**Outcome & interpretation:**

|  |
| --- |
| Call:  lm(formula = clv ~ Response + I(Coverage == "Premium") + I(Education ==  "High School or Below") + I(Education == "Master") + EmploymentStatus +  Gender + Income + Marital.Status + Monthly.Premium.Auto +  Number.of.Open.Complaints + Number.of.Policies, data = train4)  Residuals:  Min 1Q Median 3Q Max  -947.25 -304.51 -6.16 253.90 1033.13  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -1.366e+03 4.258e+01 -32.072 < 2e-16 \*\*\*  ResponseYes -3.222e+01 1.760e+01 -1.830 0.067275 .  I(Coverage == "Premium")TRUE 1.361e+02 2.964e+01 4.591 4.58e-06 \*\*\*  I(Education == "High School or Below")TRUE 4.793e+01 1.438e+01 3.334 0.000865 \*\*\*  I(Education == "Master")TRUE 6.759e+01 2.197e+01 3.077 0.002107 \*\*  EmploymentStatusEmployed 1.470e+02 2.854e+01 5.151 2.73e-07 \*\*\*  EmploymentStatusMedical Leave -6.980e+01 3.552e+01 -1.965 0.049477 \*  EmploymentStatusRetired -1.733e+02 4.157e+01 -4.169 3.13e-05 \*\*\*  GenderM -2.281e+01 1.242e+01 -1.837 0.066223 .  Income 2.295e-03 3.025e-04 7.588 4.17e-14 \*\*\*  Marital.StatusMarried 4.206e+01 1.699e+01 2.475 0.013373 \*  Marital.StatusSingle -9.281e+01 2.137e+01 -4.343 1.45e-05 \*\*\*  Monthly.Premium.Auto 5.248e+01 3.518e-01 149.176 < 2e-16 \*\*\*  Number.of.Open.Complaints1 -3.582e+01 1.981e+01 -1.808 0.070679 .  Number.of.Open.Complaints2 -7.285e+01 3.300e+01 -2.208 0.027324 \*  Number.of.Open.Complaints3 -1.367e+02 3.434e+01 -3.981 7.00e-05 \*\*\*  Number.of.Open.Complaints4 -3.482e+02 4.537e+01 -7.675 2.14e-14 \*\*\*  Number.of.Open.Complaints5 -2.819e+02 8.631e+01 -3.266 0.001101 \*\*  Number.of.Policies2 5.541e+03 2.324e+01 238.397 < 2e-16 \*\*\*  Number.of.Policies3 3.230e+03 1.861e+01 173.547 < 2e-16 \*\*\*  Number.of.Policies4 3.209e+03 2.994e+01 107.182 < 2e-16 \*\*\*  Number.of.Policies5 3.139e+03 3.016e+01 104.077 < 2e-16 \*\*\*  Number.of.Policies6 3.166e+03 2.979e+01 106.252 < 2e-16 \*\*\*  Number.of.Policies7 3.259e+03 2.959e+01 110.165 < 2e-16 \*\*\*  Number.of.Policies8 3.238e+03 2.923e+01 110.779 < 2e-16 \*\*\*  Number.of.Policies9 3.253e+03 2.788e+01 116.683 < 2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 363.1 on 3432 degrees of freedom  Multiple R-squared: 0.9664, Adjusted R-squared: 0.9662  F-statistic: 3954 on 25 and 3432 DF, p-value: < 2.2e-16 |

Among all the insignificant variables we drop months since last claim, location code. Among all the insignificant categorical variable we drop categories like extended premium under coverage and within education college and doctor. We keep all the significant variable and ran a regression model.

**Final model:**

Code: L6 <- lm(clv~Response+ I(Coverage == "Premium")+I(Education == "High School or Below")+ I(Education == "Master")+I(EmploymentStatus == "Employed")+ I(EmploymentStatus == "Retired")+Gender+Income+Marital.Status+ Monthly.Premium.Auto+ Number.of.Open.Complaints+ Number.of.Policies, data=train4)

summary(L6)

**Outcome & interpretation**:

Call:

lm(formula = clv ~ Response + I(Coverage == "Premium") + I(Education == "High School or Below") + I(Education == "Master") + I(EmploymentStatus == "Employed") + I(EmploymentStatus == "Retired") + Gender + Income + Marital.Status + Monthly.Premium.Auto + Number.of.Open.Complaints + Number.of.Policies, data = train4)

Residuals:

Min 1Q Median 3Q Max

-970.99 -303.05 -7.19 254.99 996.09

Coefficients:

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

(Intercept) -1.401e+03 3.845e+01 -36.445 < 2e-16 \*\*\*

ResponseYes -3.227e+01 1.761e+01 -1.833 0.066925 .

I(Coverage == "Premium")TRUE 1.363e+02 2.966e+01 4.595 4.49e-06 \*\*\*

I(Education == "High School or Below")TRUE 4.794e+01 1.438e+01 3.333 0.000869 \*\*\*

I(Education == "Master")TRUE 6.887e+01 2.197e+01 3.135 0.001733 \*\*

I(EmploymentStatus == "Employed")TRUE 1.829e+02 2.194e+01 8.335 < 2e-16 \*\*\*

I(EmploymentStatus == "Retired")TRUE -1.373e+02 3.733e+01 -3.678 0.000239 \*\*\*

GenderM -2.382e+01 1.241e+01 -1.919 0.055027 .

Income 2.296e-03 3.026e-04 7.589 4.13e-14 \*\*\*

Marital.StatusMarried 4.076e+01 1.699e+01 2.399 0.016472 \*

Marital.StatusSingle -9.510e+01 2.135e+01 -4.455 8.68e-06 \*\*\*

Monthly.Premium.Auto 5.250e+01 3.519e-01 149.212 < 2e-16 \*\*\*

Number.of.Open.Complaints1 -3.725e+01 1.981e+01 -1.880 0.060138 .

Number.of.Open.Complaints2 -7.180e+01 3.301e+01 -2.175 0.029687 \*

Number.of.Open.Complaints3 -1.373e+02 3.435e+01 -3.998 6.51e-05 \*\*\*

Number.of.Open.Complaints4 -3.502e+02 4.538e+01 -7.717 1.55e-14 \*\*\*

Number.of.Open.Complaints5 -2.793e+02 8.634e+01 -3.235 0.001227 \*\*

Number.of.Policies2 5.544e+03 2.319e+01 239.048 < 2e-16 \*\*\*

Number.of.Policies3 3.231e+03 1.860e+01 173.707 < 2e-16 \*\*\*

Number.of.Policies4 3.208e+03 2.994e+01 107.120 < 2e-16 \*\*\*

Number.of.Policies5 3.140e+03 3.018e+01 104.053 < 2e-16 \*\*\*

Number.of.Policies6 3.167e+03 2.980e+01 106.266 < 2e-16 \*\*\*

Number.of.Policies7 3.258e+03 2.959e+01 110.101 < 2e-16 \*\*\*

Number.of.Policies8 3.237e+03 2.924e+01 110.716 < 2e-16 \*\*\*

Number.of.Policies9 3.252e+03 2.789e+01 116.622 < 2e-16 \*\*\*

---

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

Residual standard error: 363.2 on 3433 degrees of freedom

Multiple R-squared: 0.9664, Adjusted R-squared: 0.9662

F-statistic: 4115 on 24 and 3433 DF, p-value: < 2.2e-16

Finally, we reach the final model where p-values of all the variables are close to zero. This means we can reject null hypothesis and all the variables are significant. These variables have a high impact on clv.

We run this model on the test data to predict the clv values and also to see the accuracy of

the data we run some tests on this fitted model.

**Fitted model:**

**Code:** f1 <- lm(clv~I(Education == "High School or Below")+ I(EmploymentStatus == "Employed")+ I(EmploymentStatus == "Retired")+ Gender+ Income+ I(Marital.Status == "Single")+Monthly.Premium.Auto+ Number.of.Open.Complaints+ Number.of.Policies, data=test4)

summary(f1)

**Outcome & interpretation:**

Call:

lm(formula = clv ~ I(Education == "High School or Below") + I(EmploymentStatus == "Employed") + I(EmploymentStatus == "Retired") + Gender + Income + I(Marital.Status == "Single") + Monthly.Premium.Auto + Number.of.Open.Complaints + Number.of.Policies, data = test4)

Residuals:

Min 1Q Median 3Q Max

-872.3 -284.0 -43.7 261.1 1201.0

Coefficients:

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

(Intercept) -1.365e+03 5.527e+01 -24.697 < 2e-16 \*\*\*

I(Education == "High School or Below")TRUE 4.288e+01 2.154e+01 1.991 0.04667 \*

I(EmploymentStatus == "Employed")TRUE 1.590e+02 3.314e+01 4.798 1.76e-06 \*\*\*

I(EmploymentStatus == "Retired")TRUE -2.619e+02 5.232e+01 -5.005 6.27e-07 \*\*\*

GenderM -5.790e+01 1.910e+01 -3.031 0.00248 \*\*

Income 1.897e-03 4.780e-04 3.968 7.60e-05 \*\*\*

I(Marital.Status == "Single")TRUE -7.253e+01 2.537e+01 -2.859 0.00431 \*\*

Monthly.Premium.Auto 5.286e+01 5.088e-01 103.883 < 2e-16 \*\*\*

Number.of.Open.Complaints1 -7.403e+01 3.027e+01 -2.445 0.01459 \*

Number.of.Open.Complaints2 -1.265e+02 4.895e+01 -2.584 0.00987 \*\*

Number.of.Open.Complaints3 -9.420e+01 5.390e+01 -1.748 0.08071 .

Number.of.Open.Complaints4 -3.808e+02 6.769e+01 -5.625 2.22e-08 \*\*\*

Number.of.Open.Complaints5 -3.780e+02 9.614e+01 -3.932 8.82e-05 \*\*\*

Number.of.Policies2 5.608e+03 3.745e+01 149.735 < 2e-16 \*\*\*

Number.of.Policies3 3.237e+03 2.827e+01 114.483 < 2e-16 \*\*\*

Number.of.Policies4 3.204e+03 4.644e+01 68.998 < 2e-16 \*\*\*

Number.of.Policies5 3.148e+03 4.589e+01 68.613 < 2e-16 \*\*\*

Number.of.Policies6 3.260e+03 4.425e+01 73.685 < 2e-16 \*\*\*

Number.of.Policies7 3.345e+03 4.349e+01 76.906 < 2e-16 \*\*\*

Number.of.Policies8 3.258e+03 4.436e+01 73.449 < 2e-16 \*\*\*

Number.of.Policies9 3.255e+03 4.494e+01 72.424 < 2e-16 \*\*\*

---

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

Residual standard error: 364.4 on 1462 degrees of freedom

Multiple R-squared: 0.9643, Adjusted R-squared: 0.9638

F-statistic: 1974 on 20 and 1462 DF, p-value: < 2.2e-16

At the beginning we split our data into two parts. We build a model with 70% of the dataset and want to run the model on the rest of the data. When we run the model on test data coverage and response marital status married become insignificant. So, we remove them and finally reach at f1 model where all the variables are significant. From this we predict the clv values and on the basis of the prediction the firm will target the customers so that their revenue will be maximized. Here all the variables are significant.

**R-squared and adjusted r-squared:** Here R-squared is 0.9643 and adjusted R-squared

is 0.9638. It indicates that all the independent variables can explain 96% of the dependent variable. When we drop variables or include some new variables r-squared increases, this mislead us. So, to remove the problem we use adjusted r-squared. Here adjusted r-squared value is close to r-squared. So we can conclude that the model is indeed a good model and the variables that we got as significant are indeed affecting the clv in real terms.

**Tests:**

1. **Multicollinearity test:** Code: vif(f1)

**Outcome & interpretation:** When the explanatory variables have a relationship between each other then multicollinearity exists. The variance inflation factor (VIF) identifies correlation between independent variables and the strength of that correlation. A key goal of regression analysis is to isolate the relationship between each independent variable and the dependent variable. The interpretation of a regression coefficient is that it represents the mean change in the dependent variable for each one-unit change in an independent variable when you hold all of the other independent variable’s constant.

The idea is that we can change the value of one independent variable and not the others. However, when independent variables are correlated, it indicates that changes in one variable are associated with shifts in another variable. The stronger the correlation, the more difficult it is to change one variable without changing another. It becomes difficult for the model to estimate the relationship between each independent variable and the dependent variable independently because the independent variables tend to change in unison.

vif(f1)

GVIF Df GVIF^(1/(2\*Df))

I(Education == "High School or Below") 1.014289 1 1.007119

I(EmploymentStatus == "Employed") 1.795601 1 1.340001

I(EmploymentStatus == "Retired") 1.319201 1 1.148565

Gender 1.018189 1 1.009053

Income 1.498330 1 1.224063

I(Marital.Status == "Single") 1.017146 1 1.008537

Monthly.Premium.Auto 1.114181 1 1.055548

Number.of.Open.Complaints 1.083260 5 1.008030

Number.of.Policies 1.191336 8 1.011002

If VIF is less than 2 then it means no multicollinearity and if it is greater than 2 then it means multicollinearity is present. Here in our mode for all the independent variables VIF is less than 2. It means there are no multicollinearity in this model.

1. **Autocorrelation- D-W test:** Code: durbinWatsonTest(f1)

**Outcome & interpretation:** The Durbin Watson (DW) statistic is a test for [autocorrelation](https://www.investopedia.com/terms/a/autocorrelation.asp) in the residuals from a statistical [regression analysis](https://www.investopedia.com/terms/r/regression.asp). The Durbin-Watson statistic will always have a value between 0 and 4. A value of 2.0 means that there is no autocorrelation detected in the sample. Values from 0 to less than 2 indicate positive autocorrelation and values from from 2 to 4 indicate negative autocorrelation.

**H0: null hypothesis: there is no autocorrelation**

**H1: alternative hypothesis: autocorrelation is present.**

If p value is close to 0 or d-w statistic is equals to 2 then there is no autocorrelation. If it is greater than or less than 2 and p value is high as well then autocorrelation exist.

lag Autocorrelation D-W Statistic p-value

1 0.03126396 1.935986 0.214

Alternative hypothesis: rho != 0

Here d-w statistics is close to 2.00 and p-value is greater than 0.05. It means there are no autocorrelation present in the model.

1. **Heteroscehdasticity- BP test:** Code: bptest(f1)

**Outcome & interpretation:** The Breusch-Pagan-Godfrey Test is a test for heteroscedasticity of errors in regression. Heteroscedasticity means “differently scattered”; this is opposite to homoscedastic, which means “same scatter.” Homoscedasticity in regression is an important assumption; if the assumption is violated, you won’t be able to use regression analysis.

**H0: null hypothesis: the error variances are all equal.**

**H1: alternative hypothesis: the error variances are not equal. More specifically, as Y increases, the variances increase (or decrease).**

studentized Breusch-Pagan test

data: f1

BP = 463.29, df = 20, p-value < 2.2e-16

If p-value is close to zero then we have to reject null hypothesis and accept the alternate ones. Here p-value is very small which is close to 0, it means error variances are not equal. It means if we increase Y, the variance will increase. So, we reject null hypothesis and accept the alternate ones that means heteroskedasticity is present.

**4. Anderson-Darling test:** It is a statistical test of whether or not a dataset comes from a certain probability distribution, e.g., the normal distribution. The test involves calculating the Anderson-Darling statistic.  We can use the Anderson-Darling statistic to compare how well a data set fits different distributions. The two hypotheses for the Anderson-Darling test for the normal distribution are given below:

**H0: The data follows the normal distribution**

**H1: The data do not follow the normal distribution**

The null hypothesis is that the data are normally distributed; the alternative hypothesis is that the data are non-normal.

Anderson-Darling normality test

data: resids1

A = 9.8279, p-value < 2.2e-16

Here p value is very small close to zero. It means we reject the null hypothesis that the data

is normal and accept the alternate ones that it is not.

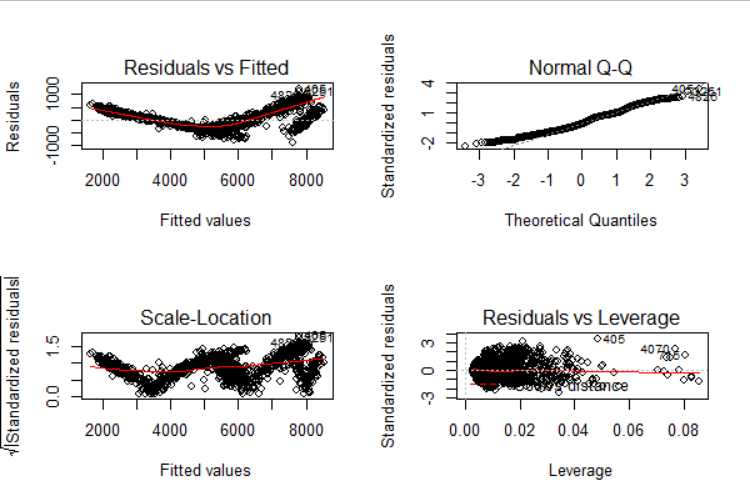
**MAPE:** Code: attach(test4)

(sum((abs(clv-prediction))/clv))/nrow(test4)

**Outcome & interpretation:**  The mean absolute percentage error (MAPE) is a statistical measure of how accurate a forecast system is. It measures this accuracy as a percentage, and can be calculated as the average absolute percent error for each time period minus actual values divided by actual values.

**Outcome:** MAPE tells how different are the predictions from actual. It ranges from 0 to 1, lesser the MAPE better the model. Here the value is 0.06. It means the model is really good and so are predictions.

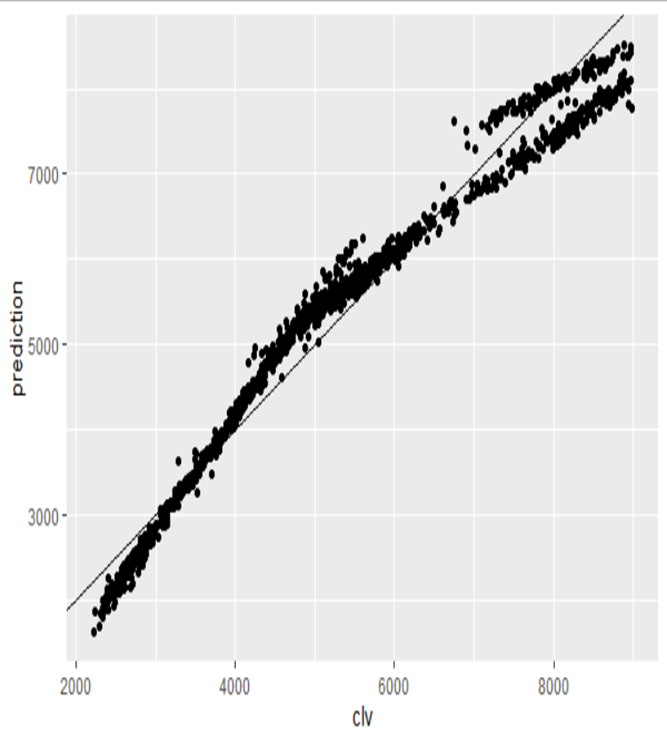
**Residual analysis:**



From the final model we get the following features:

The residual vs. fitted graph shows how likely the errors are distributed. Here we see that the data points are randomly scattered and no such pattern is observed. In 2nd diagram i.e. in Q-Q plot we see that errors are not following the corner line, which means they are not normally distributed. The 3rd diagram, scale location indicates that variance of the residuals is changing. The residual vs. leverage shows some data points have high values. Outliers might have been present in the data.

**Actual vs Predicted plot:**

****

This figure plots actual clv values on x axis and the predicted clv values on y axis. We observe that the points are close to the fitted line implying the model is a very good fit. It indicates that significant variables effect clv in real terms and the model is indeed a good model.

**Business Recommendation:**

From the analysis I would like to convey and recommend that the company should focus on increasing the values of the positive variables. Since, we want to maximize revenue, they have an increasing impact on the target variable i.e. clv.

* The agents should target mainly the customers who are employed or retired because mainly the employed and retired personnel’s need the insurance more and thus will contribute more to the company.
* The agent should target married or single as the former is more inclined towards the work and responsibility commitments and latter a more towards the fast driving.
* They should target people having either very basic or higher level of education.
* The firm should focus on customers who pays higher monthly premium as it will satisfy their object.
* The agents should target mainly the customers who are employed or retired because mainly the employed and retired personnel’s need the insurance more and thus will

contribute more to the company.

* The agent should target married or single as the former is more inclined towards the work and responsibility commitments and latter a more towards the fast driving.
* They should target people having either very basic or higher level of education.
* The firm should focus on customers who pays higher monthly premium as it will satisfy their object.