**Final Project**

Predict Customer Life-time Value for an Auto Insurance Company

**Description**: The dataset is of auto insurance company which provides insurance to customers. The dataset consists of 24 columns with various variables. The variable which we identified as dependent variable is “Customer Lifetime Value” and some variables which we identified as independent variable are: Education, Salary, Gender, Employment status and so on.

**Problem**: Predict the conditions effecting “Customer Lifetime Value”.

**Solution**: The solution is implemented in R statistical tool and the steps used to derive the solution is as follows:

**Step 1: Reading the data from the file**

> data <- read.csv("data.csv")

**Step 2: Displaying and understanding the data**

> head(data)

Customer State Customer.Lifetime.Value Response Coverage Education

1 BU79786 Washington 2763.519 No Basic Bachelor

2 QZ44356 Arizona 6979.536 No Extended Bachelor

3 AI49188 Nevada 12887.432 No Premium Bachelor

4 WW63253 California 7645.862 No Basic Bachelor

5 HB64268 Washington 2813.693 No Basic Bachelor

6 OC83172 Oregon 8256.298 Yes Basic Bachelor

Effective.To.Date EmploymentStatus Gender Income Location.Code Marital.Status

1 2/24/11 Employed F 56274 Suburban Married

2 1/31/11 Unemployed F 0 Suburban Single

3 2/19/11 Employed F 48767 Suburban Married

4 1/20/11 Unemployed M 0 Suburban Married

5 2/3/11 Employed M 43836 Rural Single

6 1/25/11 Employed F 62902 Rural Married

Monthly.Premium.Auto Months.Since.Last.Claim Months.Since.Policy.Inception

1 69 32 5

2 94 13 42

3 108 18 38

4 106 18 65

5 73 12 44

6 69 14 94

Number.of.Open.Complaints Number.of.Policies Policy.Type Policy

1 0 1 Corporate Auto Corporate L3

2 0 8 Personal Auto Personal L3

3 0 2 Personal Auto Personal L3

4 0 7 Corporate Auto Corporate L2

5 0 1 Personal Auto Personal L1

6 0 2 Personal Auto Personal L3

Renew.Offer.Type Sales.Channel Total.Claim.Amount Vehicle.Class Vehicle.Size

1 Offer1 Agent 384.8111 Two-Door Car Medsize

2 Offer3 Agent 1131.4649 Four-Door Car Medsize

3 Offer1 Agent 566.4722 Two-Door Car Medsize

4 Offer1 Call Center 529.8813 SUV Medsize

5 Offer1 Agent 138.1309 Four-Door Car Medsize

6 Offer2 Web 159.3830 Two-Door Car Medsize

**Step 3: Checking the data types:**

> str(data)

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

$ Customer : chr "BU79786" "QZ44356" "AI49188" "WW63253" ...

$ State : chr "Washington" "Arizona" "Nevada" "California" ...

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

$ Response : chr "No" "No" "No" "No" ...

$ Coverage : chr "Basic" "Extended" "Premium" "Basic" ...

$ Education : chr "Bachelor" "Bachelor" "Bachelor" "Bachelor" ...

$ Effective.To.Date : chr "2/24/11" "1/31/11" "2/19/11" "1/20/11" ...

$ EmploymentStatus : chr "Employed" "Unemployed" "Employed" "Unemployed" ...

$ Gender : chr "F" "F" "F" "M" ...

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

$ Location.Code : chr "Suburban" "Suburban" "Suburban" "Suburban" ...

$ Marital.Status : chr "Married" "Single" "Married" "Married" ...

$ 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 : chr "Corporate Auto" "Personal Auto" "Personal Auto" "Corporate Auto" ...

$ Policy : chr "Corporate L3" "Personal L3" "Personal L3" "Corporate L2" ...

$ Renew.Offer.Type : chr "Offer1" "Offer3" "Offer1" "Offer1" ...

$ Sales.Channel : chr "Agent" "Agent" "Agent" "Call Center" ...

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

$ Vehicle.Class : chr "Two-Door Car" "Four-Door Car" "Two-Door Car" "SUV" ...

$ Vehicle.Size : chr "Medsize" "Medsize" "Medsize" "Medsize" ...

**Step 4: Generating the summary of data**

> summary(data)

Customer State Customer.Lifetime.Value Response

Length:9134 Length:9134 Min. : 1898 Length:9134

Class :character Class :character 1st Qu.: 3994 Class :character

Mode :character Mode :character Median : 5780 Mode :character

Mean : 8005

3rd Qu.: 8962

Max. :83325

Coverage Education Effective.To.Date EmploymentStatus

Length:9134 Length:9134 Length:9134 Length:9134

Class :character Class :character Class :character Class :character

Mode :character Mode :character Mode :character Mode :character

Gender Income Location.Code Marital.Status

Length:9134 Min. : 0 Length:9134 Length:9134

Class :character 1st Qu.: 0 Class :character Class :character

Mode :character Median :33890 Mode :character Mode :character

Mean :37657

3rd Qu.:62320

Max. :99981

Monthly.Premium.Auto Months.Since.Last.Claim Months.Since.Policy.Inception

Min. : 61.00 Min. : 0.0 Min. : 0.00

1st Qu.: 68.00 1st Qu.: 6.0 1st Qu.:24.00

Median : 83.00 Median :14.0 Median :48.00

Mean : 93.22 Mean :15.1 Mean :48.06

3rd Qu.:109.00 3rd Qu.:23.0 3rd Qu.:71.00

Max. :298.00 Max. :35.0 Max. :99.00

Number.of.Open.Complaints Number.of.Policies Policy.Type Policy

Min. :0.0000 Min. :1.000 Length:9134 Length:9134

1st Qu.:0.0000 1st Qu.:1.000 Class :character Class :character

Median :0.0000 Median :2.000 Mode :character Mode :character

Mean :0.3844 Mean :2.966

3rd Qu.:0.0000 3rd Qu.:4.000

Max. :5.0000 Max. :9.000

Renew.Offer.Type Sales.Channel Total.Claim.Amount Vehicle.Class

Length:9134 Length:9134 Min. : 0.099 Length:9134

Class :character Class :character 1st Qu.: 272.258 Class :character

Mode :character Mode :character Median : 383.945 Mode :character

Mean : 434.089

3rd Qu.: 547.515

Max. :2893.240

Vehicle.Size

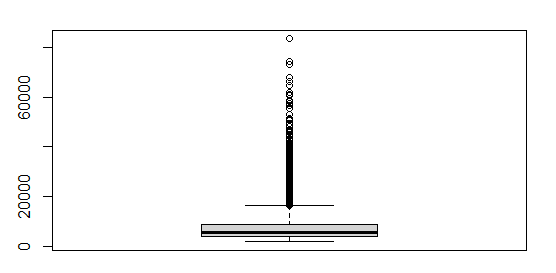
Length:9134

Class :character

Mode :character

**Step 5: Plotting the data using Boxplot to see the outliers.**

> boxplot(data$Customer.Lifetime.Value)



**Step 6 : Analyzing the Boxplot figure generated from Step 5, we can see there are outliers on the data. We need to remove the outliers.**

Checking the Quartiles

> quantile(data$Customer.Lifetime.Value, c(0,0.05,0.1,0.25,0.5,0.75,0.90,0.95,0.99,0.995,1))

0% 5% 10% 25% 50% 75% 90% 95%

1898.008 2475.109 2661.758 3994.252 5780.182 8962.167 15433.385 22064.361

99% 99.5% 100%

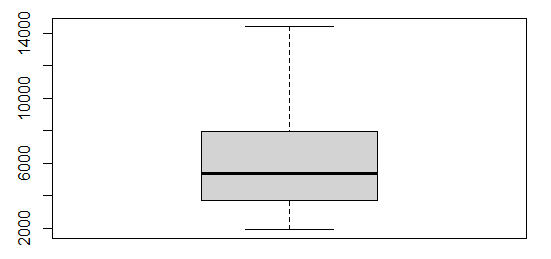
35971.105 41787.903 83325.381

**Step 7 : On the basis of Step 5 and Step 6, we remove the outliers using below command**

> data2 <- data[data$Customer.Lifetime.Value <14400, ]

> data <- data2

**Step 8: Plotting the data using Boxplot to see the outliers.**



**Step 9: As the outliers has been removed, we now clean the data. We check for the missing values on the data using below command:**

> sapply(data, function(x) sum(is.na(x)))

Customer State

0 0

Customer.Lifetime.Value Response

0 0

Coverage Education

0 0

Effective.To.Date EmploymentStatus

0 0

Gender Income

0 0

Location.Code Marital.Status

0 0

Monthly.Premium.Auto Months.Since.Last.Claim

0 0

Months.Since.Policy.Inception Number.of.Open.Complaints

0 0

Number.of.Policies Policy.Type

0 0

Policy Renew.Offer.Type

0 0

Sales.Channel Total.Claim.Amount

0 0

Vehicle.Class Vehicle.Size

0 0

**Step 10 : As seen on output of Step 9, there are no missing values, however, we run the below command to remove the missing values.**

> data <- na.omit(data)

**Step 11: After studying the data, we find the dependent variable is “Customer Lifetime Value”. The depended variable is continuous and as such we use linear regression model to predict the values. Also noting that there are multiple independent variables in the data, we use specifically multiple linear regression model.**

**Before fitting the data into linear model, we split the data into 70% training data (Development Sample) and 30% testing data ( Validation Sample) using below command:**

> dt = sort(sample(nrow(data), nrow(data)\*.7))

> development\_sample <- data[dt,]

> validation\_sample <- data[-dt,]

We then check if the split data adds up to the original data :

> #counting the total number of rows in data

> nrow(data)

[1] 8089

> #counting the total number of rows in development sample

> nrow(development\_sample)

[1] 5662

> #counting the total number of rows in validation sample

> nrow(validation\_sample)

[1] 2427

> #checking if the total number of rows of development sample and validation sample

> all.equal(nrow(data),(nrow(development\_sample) + nrow(validation\_sample)))

[1] TRUE

**Step 12: We now run multiple regression model into the data**

> fit = lm(Customer.Lifetime.Value ~ State + Response + Coverage +

+ Education + Effective.To.Date + 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 + Policy.Type +

+ Policy + Renew.Offer.Type + Sales.Channel + Total.Claim.Amount + Vehicle.Class +

+ Vehicle.Size, data = development\_sample)

**Step 13: Checking the summary and significant codes in the model:**

> summary(fit)

Call:

lm(formula = Customer.Lifetime.Value ~ State + Response + Coverage +

Education + Effective.To.Date + 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 + Policy.Type +

Policy + Renew.Offer.Type + Sales.Channel + Total.Claim.Amount +

Vehicle.Class + Vehicle.Size, data = development\_sample)

Residuals:

Min 1Q Median 3Q Max

-3843.2 -1680.8 -778.8 867.2 9961.4

Coefficients: (2 not defined because of singularities)

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

(Intercept) 2.076e+03 5.211e+02 3.983 6.90e-05 \*\*\*

StateCalifornia -2.082e+01 9.194e+01 -0.226 0.82084

StateNevada -2.133e+02 1.283e+02 -1.662 0.09650 .

StateOregon -3.062e+01 9.513e+01 -0.322 0.74760

StateWashington -3.181e+02 1.309e+02 -2.429 0.01515 \*

ResponseYes -3.233e+01 1.022e+02 -0.316 0.75171

CoverageExtended 1.700e+02 1.275e+02 1.333 0.18258

CoveragePremium 6.322e+02 2.706e+02 2.337 0.01950 \*

EducationCollege 3.300e+02 8.346e+01 3.954 7.78e-05 \*\*\*

EducationDoctor 5.372e+01 1.834e+02 0.293 0.76962

EducationHigh School or Below 1.736e+02 8.497e+01 2.043 0.04111 \*

EducationMaster 6.357e+01 1.304e+02 0.487 0.62594

Effective.To.Date1/10/11 -2.356e+02 3.293e+02 -0.715 0.47441

Effective.To.Date1/11/11 -4.632e+02 3.549e+02 -1.305 0.19191

Effective.To.Date1/12/11 -6.115e+01 3.740e+02 -0.164 0.87011

Effective.To.Date1/13/11 5.546e+00 3.615e+02 0.015 0.98776

Effective.To.Date1/14/11 1.450e+02 3.550e+02 0.408 0.68305

Effective.To.Date1/15/11 -4.712e+02 3.569e+02 -1.320 0.18673

Effective.To.Date1/16/11 1.051e+02 3.620e+02 0.290 0.77148

Effective.To.Date1/17/11 -1.492e+02 3.411e+02 -0.437 0.66180

Effective.To.Date1/18/11 -4.925e+02 3.463e+02 -1.422 0.15505

Effective.To.Date1/19/11 -3.508e+02 3.397e+02 -1.033 0.30183

Effective.To.Date1/2/11 3.500e+01 3.566e+02 0.098 0.92183

Effective.To.Date1/20/11 -2.818e+02 3.484e+02 -0.809 0.41865

Effective.To.Date1/21/11 9.736e+00 3.414e+02 0.029 0.97725

Effective.To.Date1/22/11 4.522e+02 3.650e+02 1.239 0.21546

Effective.To.Date1/23/11 -7.085e+01 3.434e+02 -0.206 0.83655

Effective.To.Date1/24/11 -5.297e+01 3.541e+02 -0.150 0.88109

Effective.To.Date1/25/11 -4.197e+02 3.504e+02 -1.198 0.23112

Effective.To.Date1/26/11 -1.050e+02 3.383e+02 -0.310 0.75630

Effective.To.Date1/27/11 4.485e+02 3.346e+02 1.340 0.18016

Effective.To.Date1/28/11 -1.201e+02 3.440e+02 -0.349 0.72698

Effective.To.Date1/29/11 -8.546e+02 3.428e+02 -2.493 0.01270 \*

Effective.To.Date1/3/11 1.592e+02 3.363e+02 0.473 0.63607

Effective.To.Date1/30/11 -5.547e+02 3.513e+02 -1.579 0.11440

Effective.To.Date1/31/11 -1.758e+02 3.515e+02 -0.500 0.61706

Effective.To.Date1/4/11 -4.220e+02 3.831e+02 -1.102 0.27068

Effective.To.Date1/5/11 -5.594e+02 3.466e+02 -1.614 0.10656

Effective.To.Date1/6/11 -3.489e+02 3.666e+02 -0.951 0.34139

Effective.To.Date1/7/11 1.666e+02 3.464e+02 0.481 0.63054

Effective.To.Date1/8/11 -1.538e+02 3.529e+02 -0.436 0.66296

Effective.To.Date1/9/11 3.290e+02 3.627e+02 0.907 0.36433

Effective.To.Date2/1/11 2.246e+02 3.522e+02 0.638 0.52369

Effective.To.Date2/10/11 -4.099e+02 3.480e+02 -1.178 0.23893

Effective.To.Date2/11/11 -7.917e+02 3.646e+02 -2.171 0.02995 \*

Effective.To.Date2/12/11 -7.006e+02 3.535e+02 -1.982 0.04751 \*

Effective.To.Date2/13/11 -1.547e+02 3.566e+02 -0.434 0.66455

Effective.To.Date2/14/11 -2.254e+02 3.406e+02 -0.662 0.50820

Effective.To.Date2/15/11 -3.582e+02 3.751e+02 -0.955 0.33972

Effective.To.Date2/16/11 -3.782e+02 3.630e+02 -1.042 0.29742

Effective.To.Date2/17/11 -4.763e+02 3.721e+02 -1.280 0.20058

Effective.To.Date2/18/11 -2.962e+02 3.572e+02 -0.829 0.40700

Effective.To.Date2/19/11 1.466e+02 3.475e+02 0.422 0.67322

Effective.To.Date2/2/11 -7.670e+01 3.472e+02 -0.221 0.82515

Effective.To.Date2/20/11 -4.043e+02 3.741e+02 -1.081 0.27988

Effective.To.Date2/21/11 -3.388e+02 3.610e+02 -0.939 0.34799

Effective.To.Date2/22/11 -6.007e+01 3.447e+02 -0.174 0.86167

Effective.To.Date2/23/11 -3.549e+02 3.537e+02 -1.003 0.31574

Effective.To.Date2/24/11 -3.610e+02 3.696e+02 -0.977 0.32877

Effective.To.Date2/25/11 -6.962e+02 3.526e+02 -1.975 0.04835 \*

Effective.To.Date2/26/11 -1.985e+02 3.498e+02 -0.568 0.57034

Effective.To.Date2/27/11 2.895e+02 3.512e+02 0.824 0.40986

Effective.To.Date2/28/11 -3.483e+02 3.499e+02 -0.995 0.31954

Effective.To.Date2/3/11 1.067e+02 3.408e+02 0.313 0.75421

Effective.To.Date2/4/11 -6.318e+02 3.509e+02 -1.801 0.07183 .

Effective.To.Date2/5/11 -1.833e+02 3.463e+02 -0.529 0.59671

Effective.To.Date2/6/11 1.324e+02 3.621e+02 0.366 0.71462

Effective.To.Date2/7/11 -2.635e+02 3.456e+02 -0.762 0.44585

Effective.To.Date2/8/11 -1.599e+02 3.609e+02 -0.443 0.65785

Effective.To.Date2/9/11 -2.734e+02 3.640e+02 -0.751 0.45261

EmploymentStatusEmployed 4.661e+01 1.725e+02 0.270 0.78706

EmploymentStatusMedical Leave -3.318e+02 2.161e+02 -1.535 0.12473

EmploymentStatusRetired -9.270e+01 2.494e+02 -0.372 0.71015

EmploymentStatusUnemployed -4.195e+02 1.744e+02 -2.406 0.01617 \*

GenderM 1.297e+01 6.505e+01 0.199 0.84193

Income 2.765e-03 1.877e-03 1.473 0.14078

Location.CodeSuburban -1.152e+02 1.289e+02 -0.893 0.37179

Location.CodeUrban -5.204e+01 1.176e+02 -0.443 0.65809

Marital.StatusMarried 2.121e+01 9.677e+01 0.219 0.82654

Marital.StatusSingle -1.630e+02 1.123e+02 -1.451 0.14672

Monthly.Premium.Auto 3.619e+01 5.320e+00 6.802 1.14e-11 \*\*\*

Months.Since.Last.Claim -4.483e+00 3.198e+00 -1.402 0.16109

Months.Since.Policy.Inception -1.643e+00 1.162e+00 -1.414 0.15747

Number.of.Open.Complaints -1.306e+02 3.471e+01 -3.763 0.00017 \*\*\*

Number.of.Policies 3.348e+02 1.318e+01 25.402 < 2e-16 \*\*\*

Policy.TypePersonal Auto -1.330e+02 1.697e+02 -0.784 0.43304

Policy.TypeSpecial Auto 3.719e+02 2.968e+02 1.253 0.21032

PolicyCorporate L2 -2.743e+02 2.045e+02 -1.341 0.17987

PolicyCorporate L3 -1.629e+02 1.883e+02 -0.865 0.38703

PolicyPersonal L1 -2.740e+00 1.005e+02 -0.027 0.97825

PolicyPersonal L2 1.274e+02 8.503e+01 1.498 0.13425

PolicyPersonal L3 NA NA NA NA

PolicySpecial L1 8.697e+01 4.576e+02 0.190 0.84925

PolicySpecial L2 -2.915e+02 3.538e+02 -0.824 0.41006

PolicySpecial L3 NA NA NA NA

Renew.Offer.TypeOffer2 -7.244e+02 8.117e+01 -8.925 < 2e-16 \*\*\*

Renew.Offer.TypeOffer3 -4.817e+02 9.807e+01 -4.912 9.29e-07 \*\*\*

Renew.Offer.TypeOffer4 -6.850e+02 1.132e+02 -6.052 1.52e-09 \*\*\*

Sales.ChannelBranch -1.391e+01 8.073e+01 -0.172 0.86323

Sales.ChannelCall Center 1.150e+02 9.071e+01 1.268 0.20481

Sales.ChannelWeb -4.868e+01 1.005e+02 -0.485 0.62803

Total.Claim.Amount 2.343e-01 2.422e-01 0.967 0.33339

Vehicle.ClassLuxury Car -5.938e+02 7.402e+02 -0.802 0.42248

Vehicle.ClassLuxury SUV -1.042e+03 7.486e+02 -1.393 0.16381

Vehicle.ClassSports Car 3.493e+02 2.777e+02 1.258 0.20859

Vehicle.ClassSUV 6.960e+02 2.399e+02 2.901 0.00373 \*\*

Vehicle.ClassTwo-Door Car 2.677e+01 8.288e+01 0.323 0.74674

Vehicle.SizeMedsize 2.324e+02 1.071e+02 2.170 0.03005 \*

Vehicle.SizeSmall 2.102e+02 1.247e+02 1.685 0.09197 .

---

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

Residual standard error: 2397 on 5555 degrees of freedom

Multiple R-squared: 0.3252, Adjusted R-squared: 0.3123

F-statistic: 25.25 on 106 and 5555 DF, p-value: < 2.2e-16

**Step 14: We model only the significant codes derived from the Step 13.**

> model = lm(Customer.Lifetime.Value ~ I(State == "Nevada") + I(State == "Washington")+

+ I(Coverage == "Premium") +

+ I(Education == "College") + I(Education == "High School or Below")+

+ I(Effective.To.Date == "1/29/11") + I(Effective.To.Date == "2/25/11")+

+ I(Effective.To.Date == "2/11/11") + I(Effective.To.Date == "2/12/11")+

+ I(Effective.To.Date == "2/4/11")+

+ I(EmploymentStatus == "Unemployed") + Monthly.Premium.Auto +

+ Number.of.Open.Complaints + Number.of.Policies +

+ Renew.Offer.Type + I(Vehicle.Class == "Luxury SUV") +

+ I(Vehicle.Class == "SUV"), data = development\_sample)

**Step 15: Checking the summary of the model**

> summary(model)

Call:

lm(formula = Customer.Lifetime.Value ~ I(State == "Nevada") +

I(State == "Washington") + I(Coverage == "Premium") + I(Education ==

"College") + I(Education == "High School or Below") + I(Effective.To.Date ==

"1/29/11") + I(Effective.To.Date == "2/25/11") + I(Effective.To.Date ==

"2/11/11") + I(Effective.To.Date == "2/12/11") + I(Effective.To.Date ==

"2/4/11") + I(EmploymentStatus == "Unemployed") + Monthly.Premium.Auto +

Number.of.Open.Complaints + Number.of.Policies + Renew.Offer.Type +

I(Vehicle.Class == "Luxury SUV") + I(Vehicle.Class == "SUV"),

data = development\_sample)

Residuals:

Min 1Q Median 3Q Max

-3991.6 -1718.1 -815.5 880.8 9930.1

Coefficients:

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

(Intercept) 1909.835 157.891 12.096 < 2e-16 \*\*\*

I(State == "Nevada")TRUE -209.801 110.095 -1.906 0.056748 .

I(State == "Washington")TRUE -303.558 113.356 -2.678 0.007430 \*\*

I(Coverage == "Premium")TRUE 525.791 135.173 3.890 0.000101 \*\*\*

I(Education == "College")TRUE 289.268 76.644 3.774 0.000162 \*\*\*

I(Education == "High School or Below")TRUE 146.496 78.049 1.877 0.060573 .

I(Effective.To.Date == "1/29/11")TRUE -712.944 235.186 -3.031 0.002445 \*\*

I(Effective.To.Date == "2/25/11")TRUE -542.613 249.322 -2.176 0.029570 \*

I(Effective.To.Date == "2/11/11")TRUE -670.389 264.402 -2.535 0.011256 \*

I(Effective.To.Date == "2/12/11")TRUE -541.494 250.371 -2.163 0.030601 \*

I(Effective.To.Date == "2/4/11")TRUE -478.189 246.517 -1.940 0.052456 .

I(EmploymentStatus == "Unemployed")TRUE -626.030 74.452 -8.408 < 2e-16 \*\*\*

Monthly.Premium.Auto 38.407 1.675 22.933 < 2e-16 \*\*\*

Number.of.Open.Complaints -123.018 34.492 -3.567 0.000365 \*\*\*

Number.of.Policies 336.475 13.017 25.848 < 2e-16 \*\*\*

Renew.Offer.TypeOffer2 -652.910 77.098 -8.469 < 2e-16 \*\*\*

Renew.Offer.TypeOffer3 -439.397 95.809 -4.586 4.61e-06 \*\*\*

Renew.Offer.TypeOffer4 -615.797 108.777 -5.661 1.58e-08 \*\*\*

I(Vehicle.Class == "Luxury SUV")TRUE -1217.279 364.891 -3.336 0.000855 \*\*\*

I(Vehicle.Class == "SUV")TRUE 644.637 105.025 6.138 8.93e-10 \*\*\*

---

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

Residual standard error: 2402 on 5642 degrees of freedom

Multiple R-squared: 0.3115, Adjusted R-squared: 0.3092

F-statistic: 134.4 on 19 and 5642 DF, p-value: < 2.2e-16

**Step 16: Checking the Assumption Diagnostic Test of Linear Model:**

> # Normalization Test

> # Anderson - Daring Test

> resids <- model$residuals

> ad.test(resids)

Anderson-Darling normality test

data: resids

A = 267.13, p-value < 2.2e-16

> #Homoscedasticity Test

> bptest(model)

studentized Breusch-Pagan test

data: model

BP = 353.36, df = 19, p-value < 2.2e-16

> #Multicollinearity Test

> vif(model)

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

I(State == "Nevada") 1.012662 1 1.006311

I(State == "Washington") 1.013432 1 1.006694

I(Coverage == "Premium") 1.388664 1 1.178416

I(Education == "College") 1.213243 1 1.101473

I(Education == "High School or Below") 1.223794 1 1.106252

I(Effective.To.Date == "1/29/11") 1.006450 1 1.003220

I(Effective.To.Date == "2/25/11") 1.006395 1 1.003192

I(Effective.To.Date == "2/11/11") 1.002738 1 1.001368

I(Effective.To.Date == "2/12/11") 1.004380 1 1.002188

I(Effective.To.Date == "2/4/11") 1.004230 1 1.002113

I(EmploymentStatus == "Unemployed") 1.033082 1 1.016407

Monthly.Premium.Auto 2.494086 1 1.579268

Number.of.Open.Complaints 1.008869 1 1.004425

Number.of.Policies 1.023068 1 1.011468

Renew.Offer.Type 1.073277 3 1.011856

I(Vehicle.Class == "Luxury SUV") 1.527925 1 1.236093

I(Vehicle.Class == "SUV") 1.630370 1 1.276860

> dwt(model)

lag Autocorrelation D-W Statistic p-value

1 0.0002394757 1.999423 0.946

Alternative hypothesis: rho != 0

**Summary**:

Based on the above Assumption Diagnostic Test, we generate the below insights of the linear model:

1. **Normality Test**: We start with Hypothesis as below

* Null Hypothesis: The errors are normally distributed. p-value should be more than 0.05.
* Alternative Hypothesis: The errors are not normally distributed. p-value is less than 0.05.

Based on the Anderson- Daring Test, we find the p-value is very less. Since the p-value is less than 0.05, we reject the null hypothesis and accept the alternative hypothesis. Finally, we conclude that the errors are not normally distributed. We have failed the Normality Test for this linear model.

1. **Homoscedasticity Test**: The assumption means that the variance around the regression line is same for all values of the predictor variable (X). For this test, we use Breusch-Pagan Test.

* Null Hypothesis: The error variances are all equal. p-value should be greater than 0.05.
* Alternative Hypothesis: The error variances are not equal. p-value should be less than 0.05

Based on the Breusch-Pagan Test, we find the p-value is very less. Since the p-value is less than 0.05, we reject the null hypothesis which says the variance is scattered similarly. We accept the alternative hypothesis and the regression model is heteroscedasticity – the variances are differently scattered.

1. **Multicollinearity Test**: Here, we check if the independent variables have relationship between them, i.e, the correlation between them. The correlation between independent or explanatory variables are called multicollinearity. We use VIF (Variance Inflation Factor) for this test.

* The value of VIF should be lower than 1.7.

Based on the test, we find the VIF score for the variables are less than 1.7. As such, we conclude that there is no multicollinearity on the linear model i.e, no correlation between independent variables in the model.

1. **Auto – Correlation Test**: The good model should not have autocorrelation in the data. Auto- Correlation occurs when the residuals are not independent from each other. For this, we use Durbin-Watson’s d test.

* Null Hypothesis: Residuals are not linearly auto correlated. Values 1.5 < d < 2.5 shows there is no auto correlation in data.
* Alternative Hypothesis: The residuals are linearly auto correlated. Values 1.5 > d > 2.5 shows there is linear auto correlations between data.

Based on the Durbin-Watson Test, we find that the values are less than 2.5, as such there is no linear auto correlation between data.

Finally, we pass all the assumption diagnostic test except the homoscedasticity test and normality test.

**Step 17: We now use the linear model to predict the values on unseen data, i.e Validation Sample ( Testing Data).**

> validation\_sample$pred <- predict(model,validation\_sample)

The data is predicted and saved.

Finally, we calculate MAPE. MAPE stands for Mean Absolute Percentage Error. MAPE tells how different the prediction from true value values. Its values are between 0 to 1. Less MAPE value corresponds to better model.

> (sum((abs(validation\_sample$Customer.Lifetime.Value-validation\_sample$pred))/validation\_sample$Customer.Lifetime.Value))/nrow(validation\_sample)

[1] 0.3528699

We received 0.352 as MAPE value of the model. The model is 65% accurate.

We save the Validation Data into csv file along with the predicted values.

> write.csv(validation\_sample,"validation\_data.csv")

**Overall Insights and suggestions:**

We can suggest as following:

* The number of complaints should be removed.
* More attention should be given to premium customers.
* The company should start increasing their policy advertisement to the customers as the number of policy effects the CLV.