2019

Project By:

Ritesh Natekar

8/23/2019



**Customer Lifetime Value**

**for**

**Auto Insurance Company**

-PROJECT REPORT-

**Table of Contents**

1. Abtract 3

2. Business Strategy 4

3. Approach 5

4. Steps to prepare fit regression model 6

5. Project Results and it's Interpretion 7

6. Linear Regression model assumptions 8

6.1 Auto-correlation

6.2 Multicollinearity

6.2 Constant Error Variance

6.3 Normality 9

7. Business Decisions after model analysis 10

Abstract

What is CLV?

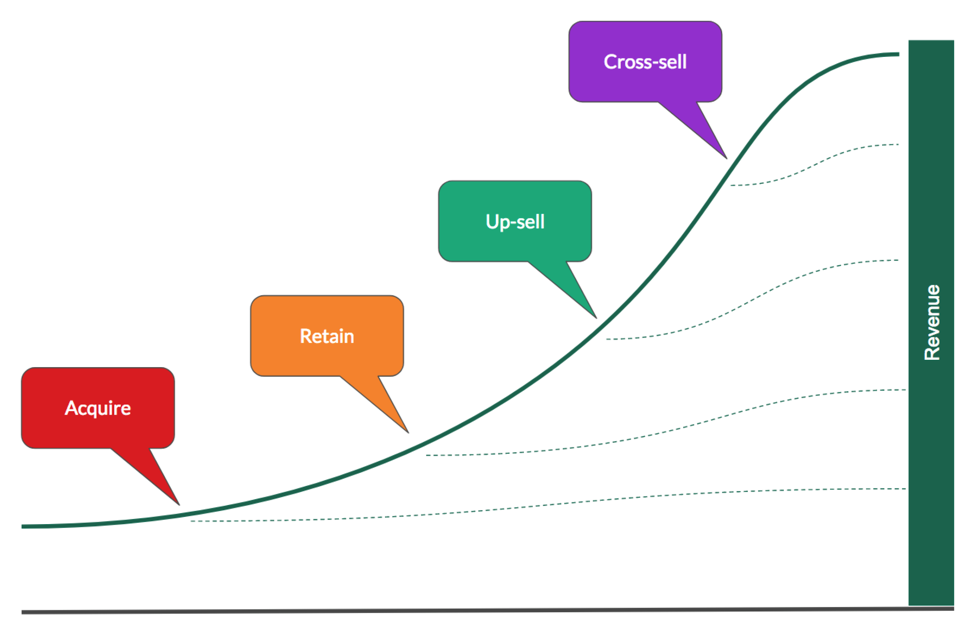
The Customer Lifetime Value is the net present value of a customer. It considers the difference between the total amount of revenues from a customer and the companies` expenses for this customer during the whole duration of relationship.

For example, if a new customer costs $50 to acquire (COCA, or cost of customer acquisition), and their lifetime value is $60, then the customer is judged to be profitable, and acquisition of additional similar customers is acceptable. Additionally, CLV is used to calculate customer equity.

Why is it important?

Customer lifetime value is important because, the higher the number, the greater the profits. You'll always have to spend money to acquire new customers and to retain existing ones, but the former costs five times as much. When you know your customer lifetime value, you can improve it.

Business Strategy



Approach

* Getting Dataset

Dataset has been collected in a “.csv” format. Examined the dataset with all variables included. Understand the importance of each variable in the dataset. Checking with data type of variables and how much relevant information they are providing for model building.

* Choosing Regression Model

In the given dataset, having variables with continuous type of data. Dependent variable i.e. Customer Lifetime Value is of type continuous and other independent variable which correlate the dependent variable more accurately are of type continuous. We have explanatory variable as quantitative and Response variable as quantitative, so **Linear Regression** model is best fit for predicting the response variable i.e. Customer Lifetime Value.

* Cleaning dataset

Once regression model is confirmed, to start the working on dataset, it is very important to clean it. Finding missing values/null values and replacing it with relevant values in order to avoid discrepancy in data. Treating outliers is very important as it affect statistical analysis.

* Development Model and Testing Model

Divide dataset into two different model:

1. Development Model
2. Testing Model

Development Model, It is used to develop regression model. Considering different explanatory variables to predict response variable and examined P-value in order to decide whether to reject null hypothesis or not.

Testing Model, It is used to conduct testing on developed fit model.

* Developing Regression Model

Linear Regression model, considering all relevant explanatory variables to predict response variable. Check the P-value, if variable having least p-values then it more accurate in predicting response variable. Check R-squared, as it ranges from 0 to 1.

* Regression Model Assumption.

Checking developed model for different assumption like autocorrelation, multicollinearity, Constant error variance and normality of dataset.

* Testing Model

Once model developed fully then perform testing on it. And validate it.

* Inferring data insights and business decisions.

Now model is ready to analyses the thing and make business decisions.

Steps to prepare fit Regression Model

* Loading Dataset into R environment.

Dataset with: Observations- 9134

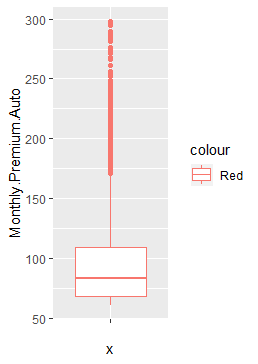
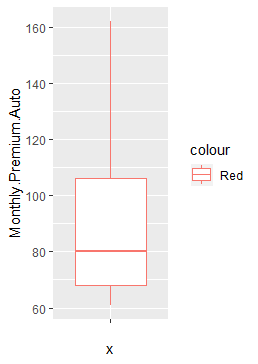
Variables – 24.

* Removing the explanatory variables which doesn’t make any sense in predicting response variable.

Here Customer ID doesn’t make any sense in predicting values, so remove it.

* Detecting and treating outliers.

Below shown outliers in variable field Monthly Premium Auto.



* Splitting dataset into two model:

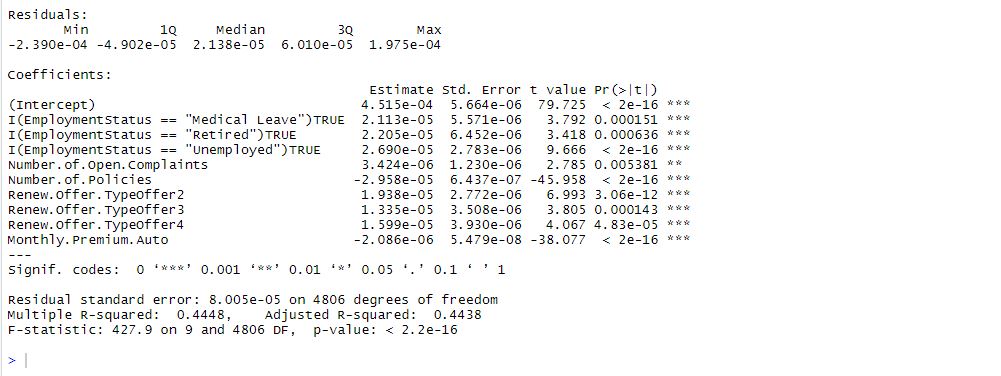
1. Development Model
2. Testing Model

* Preparing Final Regression Model

Considering variables like Employment Status, Number of open complaints, Number of policies, Renew Offer type and Monthly Premium Auto which are contributing best to predict response variable i.e. Customer Lifetime Value.

Result and its Interpretation

Result for the best fit regression model:



Customers getting new offers on policies tends to score high lifetime value and long term relationship with company. On every single offer type lifetime value of customer is increasing as there is positive relationship.

As the monthly premium is less, longer the relationship of customer with company can be observed and potential increase in lifetime value. As monthly premium is greater, there is less possibility to end up with short term business benefits for company.

With the less number of policies customer are tend to continue their relationship with company and results into high customer lifetime value. With less number of policies customer are more likely to pay premium regularly.

Customers with more number of complaints are more interacted with company and that results into making them more satisfied customer for long term relationship.

Employment status contributing positively in lifetime value. Every couple of years, getting increased in lifetime value.

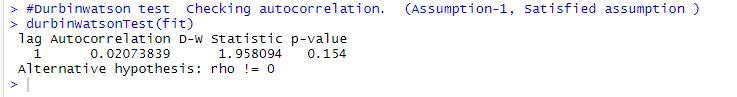
Linear Regression Model Assumptions

* Checking for autocorrelation.

Autocorrelation, also known as serial correlation, is the [correlation](https://en.wikipedia.org/wiki/Correlation) of a [signal](https://en.wikipedia.org/wiki/Signal_(information_theory)) with a delayed copy of itself as a function of delay. Informally, it is the similarity between observations as a function of the time lag between them. The analysis of autocorrelation is a mathematical tool for finding repeating patterns, such as the presence of a [periodic signal](https://en.wikipedia.org/wiki/Periodic_signal) obscured by [noise](https://en.wikipedia.org/wiki/Noise_(signal_processing)), or identifying the [missing fundamental frequency](https://en.wikipedia.org/wiki/Missing_fundamental_frequency) in a signal implied by its [harmonic](https://en.wikipedia.org/wiki/Harmonic) frequencies. It is often used in [signal processing](https://en.wikipedia.org/wiki/Signal_processing) for analyzing functions or series of values, such as [time domain](https://en.wikipedia.org/wiki/Time_domain) signals.

Run Durbin Watson Test as below:

DurbinWatsonTest(fit)



If et is the [residual](https://en.wikipedia.org/wiki/Errors_and_residuals_in_statistics) given by {\displaystyle e\_{t}=\rho e\_{t-1}+\nu \_{t},} Durbin -Watson statistic states that null hypothesis: {\displaystyle \rho =0} , alternative hypothesis {\displaystyle \rho \neq 0}, then the [test statistic](https://en.wikipedia.org/wiki/Test_statistic) is

{\displaystyle d={\sum \_{t=2}^{T}(e\_{t}-e\_{t-1})^{2} \over {\sum \_{t=1}^{T}e\_{t}^{2}}},}

where T is the number of observations. If one has a lengthy sample, then this can be linearly mapped to the Pearson correlation of the time-series data with its lags.[[2]](https://en.wikipedia.org/wiki/Durbin%E2%80%93Watson_statistic#cite_note-2) Since d is approximately equal to 2(1 − {\displaystyle {\hat {\rho }}}), where {\displaystyle {\hat {\rho }}} is the sample autocorrelation of the residuals,[[3]](https://en.wikipedia.org/wiki/Durbin%E2%80%93Watson_statistic#cite_note-Gujarati_2003-3) d = 2 indicates no autocorrelation. The value of d always lies between 0 and 4. If the Durbin–Watson statistic is substantially less than 2, there is evidence of positive serial correlation. As a rough rule of thumb, if Durbin–Watson is less than 1.0, there may be cause for alarm. Small values of d indicate successive error terms are positively correlated. If d > 2, successive error terms are negatively correlated. In regressions, this can imply an underestimation of the level of [statistical significance](https://en.wikipedia.org/wiki/Statistical_significance).

To test for positive autocorrelation at significance α, the test statistic d is compared to lower and upper critical values (dL,α and dU,α):

* If d < dL,α, there is statistical evidence that the error terms are positively autocorrelated.
* If d > dU,α, there is no statistical evidence that the error terms are positively autocorrelated.
* If dL,α < d < dU,α, the test is inconclusive.

Positive serial correlation is serial correlation in which a positive error for one observation increases the chances of a positive error for another observation.

To test for negative autocorrelation at significance α, the test statistic (4 − d) is compared to lower and upper critical values (dL,α and dU,α):

* If (4 − d) < dL,α, there is statistical evidence that the error terms are negatively autocorrelated.
* If (4 − d) > dU,α, there is no statistical evidence that the error terms are negatively autocorrelated.
* If dL,α < (4 − d) < dU,α, the test is inconclusive.

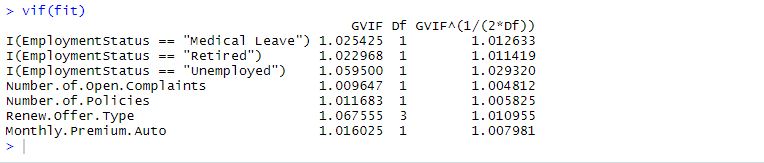
Negative serial correlation implies that a positive error for one observation increases the chance of a negative error for another observation and a negative error for one observation increases the chances of a positive error for another.

* Checking Multicollinearity

In regression, "multicollinearity" refers to predictors that are correlated with other predictors.  Multicollinearity occurs when your model includes multiple factors that are correlated not just to your response variable, but also to each other. In other words, it results when you have factors that are a bit redundant.

Run VIF test as below:

VIF(fit)



If multicollinearity is a problem in your model -- if the VIF for a factor is near or above 5 -- the solution may be relatively simple. Try one of these:

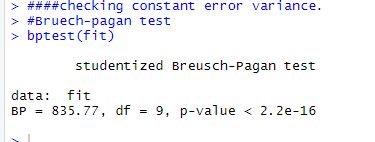
* Remove highly correlated predictors from the model.  If you have two or more factors with a high VIF, remove one from the model. Because they supply redundant information, removing one of the correlated factors usually doesn't drastically reduce the R-squared.
* Constant Error Variance

heteroskedasticity occurs when the variance for all observations in a data set are not the same. Conversely, when the variance for all observations are equal, we call that homoskedasticity. Why should we care about heteroskedasticity? Because it is a violation of the ordinary least square assumption that var(yi)=var(ei)=σ2var(yi)=var(ei)=σ2. In the presence of heteroskedasticity, there are two main consequences on the least squares estimators:

1. The least squares estimator is still a linear and unbiased estimator, but it is no longer best. That is, there is another estimator with a smaller variance.
2. The standard errors computed for the least squares estimators are incorrect. This can affect confidence intervals and hypothesis testing that use those standard errors, which could lead to misleading conclusions.

Run Bruech-pagan test as below:

Bptest(fit)



mathematical way of detecting heteroskedasticity is what is known as the **Breusch-Pagan test**. It involves using a variance function and using a χ2χ2-test to test the null hypothesis that heteroskedasticity is not present (i.e. homoskedastic) against the alternative hypothesis that heteroskedasticity is present.

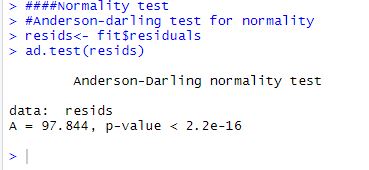
* Checking for Normality.

Normality tests are a form of [hypothesis test](http://www.itl.nist.gov/div898/handbook/prc/section1/prc13.htm), which is used to make an inference about the population from which we have collected a sample of data.

* H0: No observable difference between data and normal distribution
* Ha: Clear observable difference between data and normal distribution

Run Anderson-darling test for normality as below:

Ad.test(resids)



QQ Plot for Residual Analysis:

qqnorm is a generic function the default method of which produces a normal QQ plot of the values in y. qqline adds a line to a “theoretical”, by default normal, quantile-quantile plot which passes through the probs quantiles, by default the first and third quartiles.

qqplot produces a QQ plot of two datasets.

Graphical parameters may be given as arguments to qqnorm, qqplot and qqline.

### Usage

qqnorm(y, ...)

## Default S3 method:

qqnorm(y, ylim, main = "Normal Q-Q Plot",

xlab = "Theoretical Quantiles", ylab = "Sample Quantiles",

plot.it = TRUE, datax = FALSE, ...)

qqline(y, datax = FALSE, distribution = qnorm,

probs = c(0.25, 0.75), qtype = 7, ...)

qqplot(x, y, plot.it = TRUE, xlab = deparse(substitute(x)),

ylab = deparse(substitute(y)), ...)

### Arguments

|  |  |
| --- | --- |
| x | The first sample for qqplot. |
| y | The second or only data sample. |
| xlab, ylab, main | plot labels. The xlab and ylab refer to the y and x axes respectively if datax = TRUE. |
| plot.it | logical. Should the result be plotted? |
| datax | logical. Should data values be on the x-axis? |
| distribution | quantile function for reference theoretical distribution. |
| probs | numeric vector of length two, representing probabilities. Corresponding quantile pairs define the line drawn. |
| qtype | the type of quantile computation used in [quantile](https://stat.ethz.ch/R-manual/R-devel/library/stats/html/quantile.html). |
| ylim, ... | graphical parameters. |

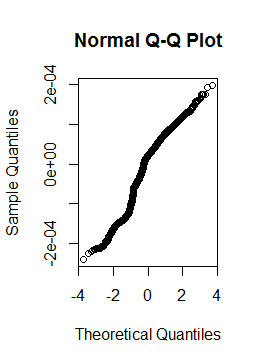
### Value

For qqnorm and qqplot, a list with components

|  |  |
| --- | --- |
| x | The x coordinates of the points that were/would be plotted |
| y | The original y vector, i.e., the corresponding y coordinates |

Run as below:

qqnorm(resids)



Business Decisions after model analysis

* Customers with moderate valued can be converted into highly valued customer by providing good offers on policies. It has been observed that customers with renew offer type more liked high value customer. Making new offers on policies will help to make customer high value.
* Association of customer with company for longer period of time will automatically make the lifetime value high. Keeping monthly premium amount small but for longer period of time will help company to have long time association with customers. And customers also will be comfortable to pay small amount of money on every month regularly.
* Customers with less number of policies tends to be more interested to continue so rather than going for number of policies with single customers. It will help to increase customer base also. Different customers with high lifetime value will definitely results into good business decision.
* Improving customer service is important, the more customer satisfied, the more business can be improved. Number of open complaints shows that company having to way communication with customers which will allow to customer to know the policies of company more. Company also can be aware with customer requirement and improve itself as per the market. To retain the customer by satisfying their need through their complaints.