Insurance\_Marketing- Customer Value Analysis

# Submitted towards partial fulfilment of the criteria for award of PGPDSE by GLIM

## **Members: Gopal Sharma, Rishima Makhija, Shoubhik Guha, Tanusha Singh, Vishal Walia, Yash Gulati**

## **Mentor: Ankush Bansal**

Abstract

In this project we tried to determine the customer lifetime value for an insurance company by analysing the present and past behaviour of the customers. Furthermore, we tried to optimize our results by doing outlier treatment through capping the outliers. We derived four new features and tried to generalize our model with help of EDA and statistical modelling. Also, we compared various machine learning models and chose the *Gradient Boosting* model which was giving us least RMSE.

**Keywords: Machine Learning, Statistical Modelling, Gradient Boosting, Random Forest, EDA, Outlier Treatment**



Ackowledgment

I have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and organizations. I would like to extend my sincere thanks to all of them.

I am highly indebted to Mr Ankush Bansal for his guidance and constant supervision as well as for providing necessary information regarding the project & also for his support in completing the project.

I would like to express my gratitude towards my parents and teachers of Great Learning for their kind co-operation and encouragement which help me in completion of this project.

My thanks and appreciations also go to my colleague in developing the project and people who have willingly helped me out with their abilities.

Ankush Bansal

CHAPTER 1

**Introduction**

Customer Lifetime Value, or CLV, is a key metric that represents how much cash the average customer spends during their relationship with a company – or in other words, how much they’re worth to a business. Depending on your business model, your customers might make repeat purchases or sign up to a recurring contract. Focusing only on that initial metric (the first purchase) is incredibly short-sighted, and likely to drastically slow growth. Understanding CLV can help with forecasting. You can predict (with varying degrees of accuracy) your future cash flow. It is a tangible metric that can help you make smarter business decisions.

CLV helps you numerically analyse your expenditure and project future spends and growth. It also helps you determine how much money you can spend to acquire new customers based on the revenue you generate. Additionally, you can evaluate the financial impact on your customers when you rebrand, run special discount campaigns, change your pricing and more. A huge aspect of a customer’s user experience depends on the end-to-end experience.

**Business Problem**

CLV is a concept used across many industries. Calculating Customer Lifetime Value (CLV) helps us to predict the profit that a customer is expected to bring to the company, taking into account the entire future relationship of a customer wit the company.

The movement of total CLV and that of different customer segment may provide good insights as far as profitability of the company is concerned. It helps in monitoring the portfolio in a better way. The whole definition of profitability changes and the company may take better decisions based on the CLV instead of other individual parameters to succeed their business growth and expand their business outlook for values.

Insurance industry, like other businesses, would like to get more out of their sales and marketing campaigns. That being the starting point during customer acquisition phase and how do they behave once they are part of the company is equally important. It needs to be learned whether the cost of acquiring a customer is actually profitable for the company and to forecast the future behaviour of customers.

During this project we’ll try to forecast the customer lifetime value for an automobile insurance company using regression models.

**Data Dictionary**

|  |  |  |
| --- | --- | --- |
| Feature | Explanation | Unique Values/ Data Types |
| State | US province to where the customer belongs to | Arizona, California, Nevada, Oregon, Washington |
| Response | Refers to whether customers have responded to marketing calls or not | Yes, No |
| Coverage | Nature of Insurance Coverage | Basic, Extended, Premium |
| Education | Education level of Customer | below high school, College, Graduate, Masters, Doctor |
| Effective to Date | Expiry Date of policy | Date-Time |
| Gender |  | Male, Female |
| Employment Status |  | Employed, Unemployed, Retired, Disabled, Medical leave |
| Income | Customer Annual Income in USD | Continuous |
| Location Code | Type of location where customer lives | Rural, Suburban, Urban |
| Marital Status |  | Single, Married, Divorced |
| Vehicle Size |  | Small, Medium, Large |
| Vehicle Class |  | Two-door, Four-door, SUV, Sports car, Luxury sport, Luxury car |
| Sales Channel |  | Agent, Branch, Web, Call Centre |
| Renew Offer Type | Offer given during renewal | Offer1, Offer2, Offer3, Offer4 |
| Total Claim Amount | Amount claimed till date | Continuous |
| Monthly Premium Auto | Monthly Premium for Auto Insurance | Continuous |
| Months Since Last Claim | No: of months before which the last claim was made | Continuous |
| Months Since policy Inception | No: of months before which the policy commenced | Continuous |
| Number of Open complaints | No: of unresolved complaints from the customer | Continuous |
| Number of Policies | No: of policies with the current customer | Continuous |
| Policy Type |  | Personal, Corporate, Special |
| Policy | Sub category of Policy Type | Categorical |

**Hypothesis regarding Features:**

1. *Coverage will have very significant impact on monthly premium paid by a customer.*
2. *Customers living in rural areas are less valuable than customers living in suburban or urban areas.*
3. *People with higher education level are likely to have higher Coverage and therefore, more valuable asset for the company.*
4. *Vehicle size and Vehicle class might have an impact on the amounts of claims taken by the customer.*
5. *If the sales channel is agent or a branch, a customer is likely to stay for longer period in the future with the company, thus increasing the CLV for a customer.*
6. *Total Claim Amount is likely to have negative impact on CLV as it is considered a cost for retaining a customer.*
7. *Premium will have positive correlation with CLV.*
8. *Months Since policy inception should also have a positive correlation, since longer the customer has been with a company, it is very likely the customer will continue to stay with the company.*
9. *No of open complaints can have a negative impact, as higher the complaints, more likely the customer is to switch to other company.*
10. *No: of policies will have a positive correlation with CLV, as high no: of policy means that the customer is satisfied with the services and would likely to retain with a particular company.*

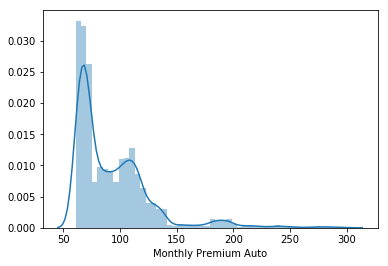
CHAPTER 2

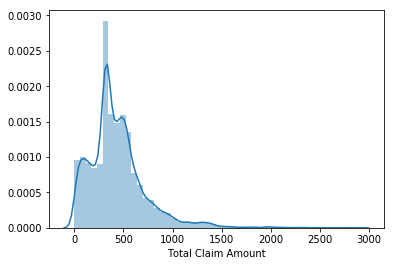
**Exploring Data Analysis (EDA)**

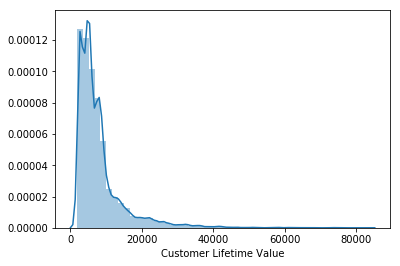
We noticed that there are no null values in our data. Also, we have 6 numerical features, 14 categorical features and 1 date-time feature. Our target variable is Customer Lifetime Value. It is a regression problem. We have 9134 instances.

**Univariate Analysis**

Upon conducting a univariate analysis, we found that *Customer Lifetime Value (CLV)*, *Monthly Premium Auto (MPA)* and *Total Claim Amount (TCA)* have outliers among them and all of them highly skewed towards right.

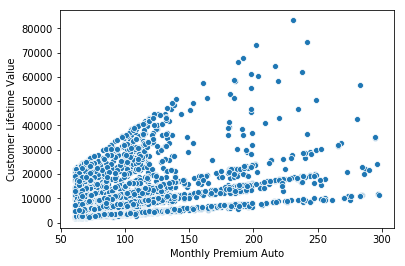






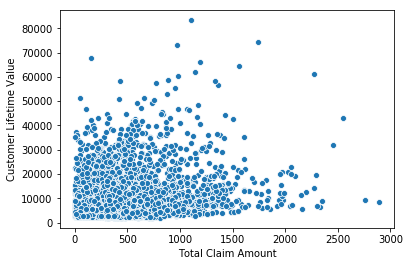
**Bivariate Analysis**

1. Customer Lifetime Value and Monthly Premium Auto



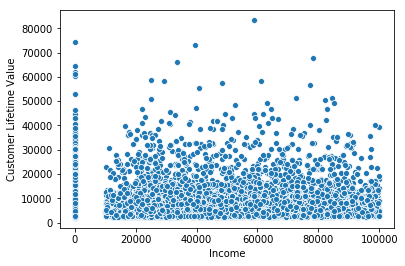
We see that there is a positive relation among them, but it also has a heteroskedastic relationship among them.

1. Customer Lifetime Value and Total Claim Amount



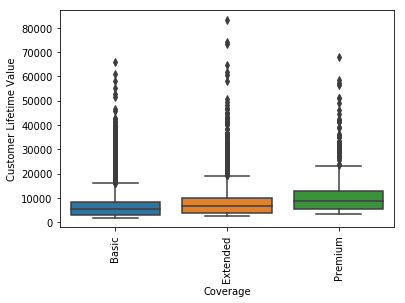
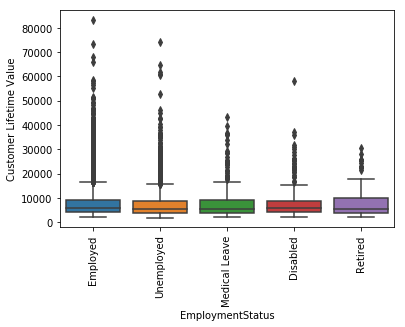
We assumed there should be a negative correlation among them, but there seems to be a positive correlation among them and we can see a heteroskedastic relation between these two variables. Also, we can detect the outliers.

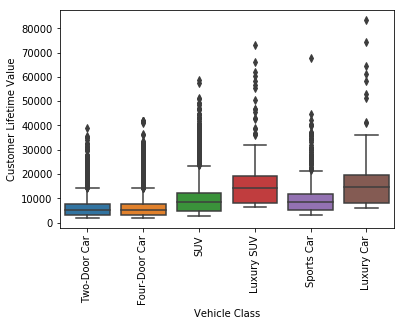
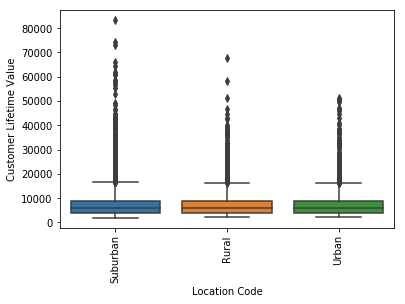
1. Customer Lifetime Value and Income



There doesn’t seem to exist any relation between *Customer Lifetime Value* and *Income*. Most of income values are zero.

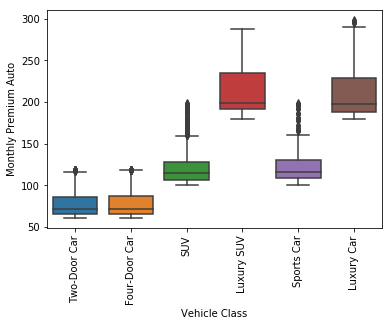
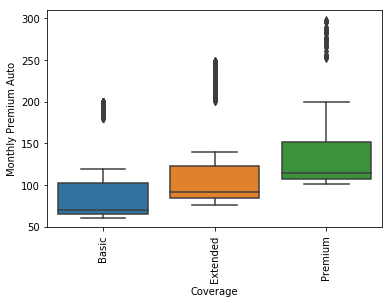
1. Customer Lifetime Value and Categorical Features

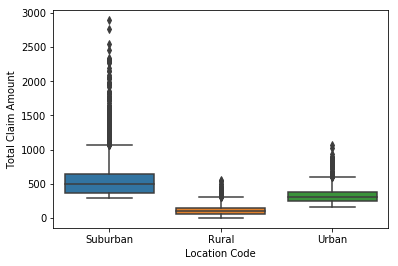
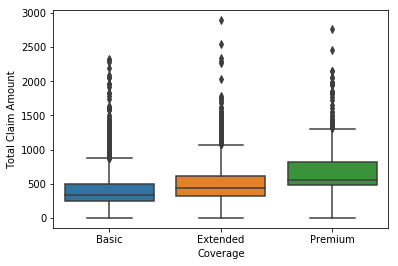
We notice that we can see clear difference with respect to *Coverage* and vehicle class in *CLV*. On the other hand, *Employment Status* and *Location Code* are not very different for *CLV*. Customers with *Premium Coverage* generally has high value for company. Similarly, customer with *Luxury Car* and *Luxury SUV* have high value for a company.

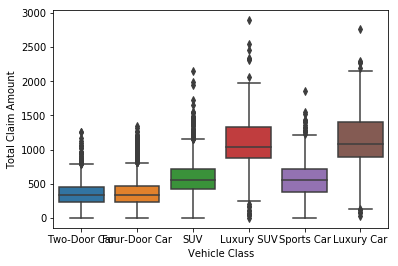
1. Monthly Premium Auto and Categorical Features



Similarly, for *MPA* clear distinction can be seen with respect to *Vehicle Class* and *Coverage.*

1. Total Claim Amount and Categorical Features





While, for *TCA* a distinction can also be seen with respect to *Location Code* apart from above mentioned categories. *Suburban* people are generally going for high amounts of claims while *Rural* people are going for a smaller number of claims.

1. ANOVA Tests

After pictorial analysis, we also conducted one-way ANOVA and pairwise Tukey test. Our null hypothesis was whether there is a variance between different categories with respect to our target variable. We observed the following-

1. Upon further investigation we found that *TCA*, *MPA* and *CLV* vary with respect to *Coverage*.

2. Upon further investigation we found that *TCA*, *MPA* and *CLV* vary with respect to *Vehicle Class* where *Two-Door* and *Four-Door* are not different, *SUV* and *Sports Car* are not different from each other and *Luxury Car* and *Luxury SUV* are not different from each other.

3. While, *Location Code* impacts only *MPA* and *TCA*.

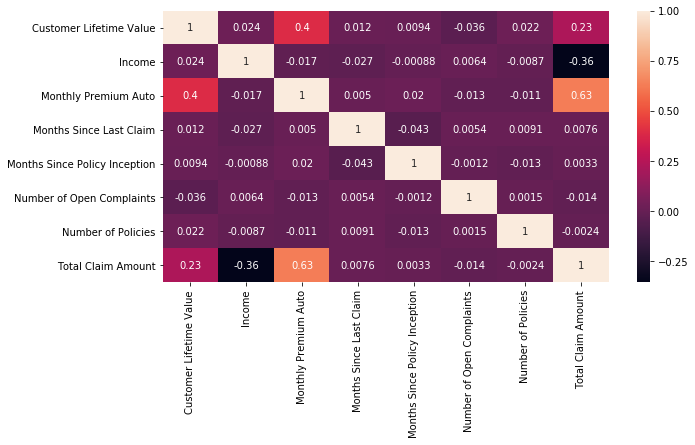
4. *Marital Status* impacts *TCA*, where single is significantly different from *Married* and *Divorced*.

5. *Employment Status* impacts *TCA* and *CLV* where, *Retired, Medical Leave* and *Disable* have no significant difference among them.

6. *Vehicle Size* impacts *TCA*, where *Small* car is different than *Medium* and *Large* cars

7. *Education* also with respect to *TCA* where *Bachelors* and *College* have no difference among them and *Masters* and *Doctor* have no difference among them.

**Multivariate Analysis**



We notice that only *MPA* and *TCA* have a strong correlation with our target. Also, *Number of Complaints* have a negative correlation as hypothesised by us, although not very strong. Similarly, *Number of Policies* have positive correlation as hypothesised by us. We also notice that *Income* and *TCA* have a negative correlation among them. While, *MPA* and *TCA* have a strong positive correlation among them.

CHAPTER 3

**Data Pre-processing and Preparation**

We started with 24 features, wherein we dropped *Customer* and *Effective to Date* because they were unique. Furthermore, we also dropped *Policy Type* as that information was being captured by *Policy*.

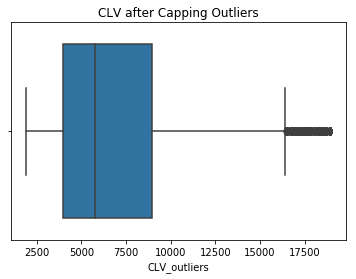
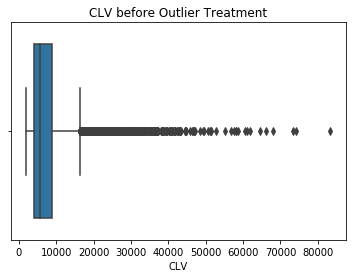
**Outlier Treatment**

As captured by our univariate and bivariate analysis, there were outliers present in our target variable *Customer Lifetime Value (CLV).* To treat that we followed two distinct approach-

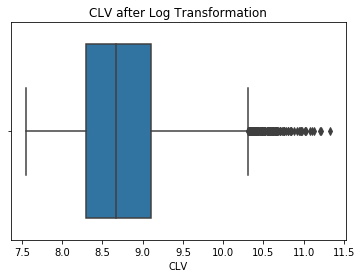
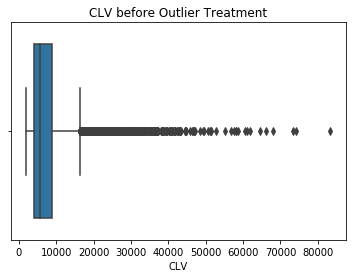
1. Capping Outliers

For this, we took the 99 per cent of data till 4 standard deviations of mean and took out the new mean and new standard deviation for that data. Further, we capped the extreme value by taking-

**new\_mean + 2\*new\_sd**

****

1. Applying log-transformation



From the above graphs it is visible that capping outliers seems to be better approach than applying log transformation.

**Feature Extraction**

We were given *Income* on an annual basis, *Premium Monthly Auto* on a monthly basis and *Total Claim Amount* since the inception of policy. To bring all variables within same time measurement we decided to extracted two new features-

1. *Annual\_Claim\_Amount* = *Total Claim Amount* / (*Months Since Policy Inception*/12)
2. *Annual\_Premium* = *Monthly Premium Auto*\*12

Apart from these we derived two more features, namely-

1. *Per\_policy\_cost* = *Total Claim Amount/Number of Policies*
2. *Claim* = It was a Boolean feature, where if a person has made a claim since policy inception, it returns yes and if not, then it returns no. It was derived by comparing *Months Since Policy Inception* and *Months Since Last Claim* made.

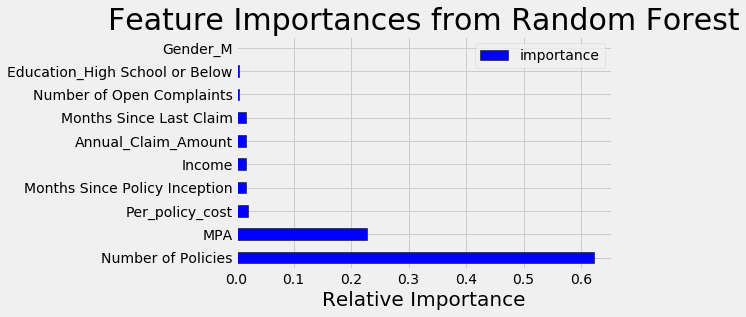
**Data Manipulation**

Based on pictorial analysis with boxplots and statistical tests, we decided to generalize our model and minimize our features-

1. We converted *Vehicle Class* from 6 categories to 3 categories, where *Two-Door car* and *Four-Door car* were under *low\_class*, *SUV* and *Sports Car* were under *med\_class*, and *Luxury SUV* and *Luxury Sports* were under *high\_class*.
2. Furthermore, in *Education*, we clubbed *College* and *Graduate* under one category as *College\_Graduate* and *Master* and *Doctor* under one category as *Post\_Graduate*.
3. We also clubbed together *Retired*, *Disabled*, and on *Medical Leave* in *EmploymentStatus* as *employes\_special*.

**Feature Selection**

We decided to select features through random forest model.



However, reducing the features did not improve the results. It turns out that the extra information in the features with low importance do actually improve performance. The model results are slightly worse with the reduced set of features and therefore, we will keep all of the features for the final model.

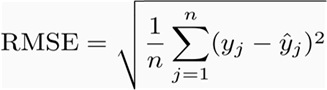
|  |  |
| --- | --- |
| RMSE with all features | 1865.9755 |
| RMSE with selected features | 1898.6018 |

The desire to reduce the number of features is because we are always looking to build the most parsimonious model: that is, the simplest model with adequate performance. A model that uses fewer features will be faster to train and generally easier to interpret. In this case, keeping all of the features is not a major concern because the training time is not significant and we can still make interpretations with many features.

CHAPTER 4

**Modelling**

Before applying any data manipulation, we decided to build a base model, to which we could compare our results. Our model evaluation metric is RMSE, it’s the square root of the average of squared differences between prediction and actual observation-



Here are the key points to consider for using RMSE:

1. The power of ‘square root’ empowers this metric to show large number deviations.
2. The ‘squared’ nature of this metric helps to deliver more robust results which prevents cancelling the positive and negative error values. In other words, this metric aptly displays the plausible magnitude of error term.
3. It avoids the use of absolute error values which is highly undesirable in mathematical calculations.
4. When we have more samples, reconstructing the error distribution using RMSE is considered to be more reliable.
5. As compared to mean absolute error, RMSE gives higher weightage and punishes large errors.

In the base model we simply took the mean of *CLV* as predicted value and determined the

RMSE through it. We achieved an RMSE of 7087.2480.

**Linear Regression Model**

Since *CLV* is continuous in nature, we’ll apply the regression model. When we do not know

the parameters of a hypothetical linear model in advance, linear regression is the method we

use to estimate those parameters. Linear regression scans the data and computes parameters

for the linear model that “best” fits the data. The method chooses an optimal model through

the least squares’ criterion, which minimizes the squared errors between predicted and actual

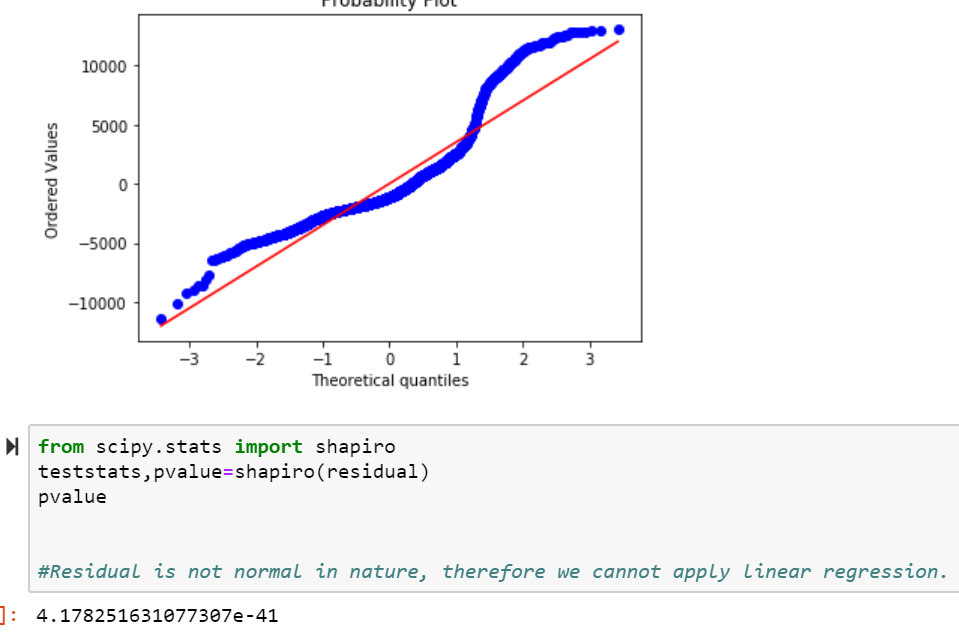
values.

Linear regression is a powerful and widely used method and relatively easy to implement.

However, the method has a number of properties that limit its application, require the analyst to prepare the data in certain ways or, in the worst case, lead to spurious results.

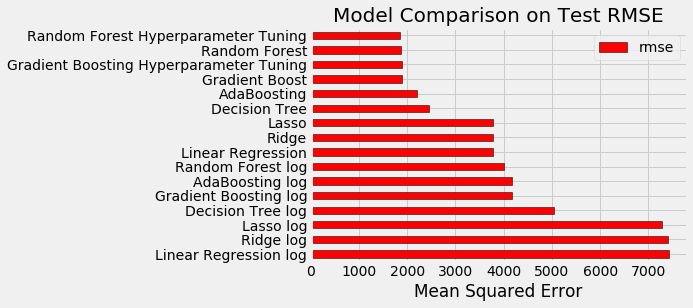
We choose to apply linear regression model because it has a very high explaining power however, it is a parametric algorithm which works on number of assumptions. If our dataset is able to meet all the assumptions, then linear regression is the best model. After fitting linear regression on the above explained data, we found the RMSE of 3783.

However, our dataset was not fit for linear regression model since it was unable to meet the assumptions of normality of residuals.



We can move forward with this and try applying non-parametric algorithms on our dataset and see which algorithm after hyperparameter tuning helps us achieve the best results.

**Comparing Different Models**



From the above figure it is clear that capping outlier results in better model than by using log transformation. Also, we can say that hyper parameter tuning has resulted in decrease in RMSE.

However, although the above figure suggests that we should go with random forest as our final model, but after comparing the training and testing score, it seems that random forest is over fitting on training data and therefore, it is best for us to move forward with gradient boost as our final model.

|  |  |  |  |
| --- | --- | --- | --- |
|  | RMSE | Training Score | Testing Score |
| Random Forest | 1812 | 0.9658 | 0.8533 |
| Gradient Boosting | 1854 | 0.8834 | 0.8464 |

The logic behind gradient boosting is simple. A basic assumption of linear regression is that sum of its residuals is 0, i.e. the residuals should be spread randomly around zero. Now think of these residuals as mistakes committed by our predictor model. We might arguethat, if we are able to see some pattern of residuals around 0, we can leverage that pattern to fit a model. So, the intuition behind gradient boosting algorithm is to repetitively leverage the patterns in residuals and strengthen a model with weak predictions and make it better. Once we reach a stage that residuals do not have any pattern that could be modelled, we can stop modelling residuals (otherwise it might lead to overfitting). Algorithmically, we are minimizing our loss function, such that test loss reach its minima.

CHAPTER 5

We have decided to go with gradient boosting as our final model, even though it was not giving us the least RMSE among all the models, however model with least RMSE was overfitting on training data. After implementing this model, a business can be benefitted in number of ways including-

1. **Customer Lifetime Value by Segment:**This takes into consideration the customer variance. This means that different customers have different needs and spending tendencies. While averages are an easy way to get an overall picture, creating segments help analyse targeted customers and their spends. This will help gain a better understanding of specific customers. These segments can be based on location, modes of distributions, select marketing campaigns and more.
2. CLV can helps us in planning better business strategies. For example, higher commission and brokerage (initial and renewal) can be offered to those customers who have high CLV even if the initial premium paid by them may not be very high.
3. It can also help in making best use of marketing budgets by directing the advertising campaigns, etc. towards the most profitable customers in terms of CLV.

However, due to lack of resources and time there were certain limitations in our model including-

1. Due to lack of proper source of data, we had to drop *Effective to Date*, which could have been used as an important feature in deriving retention period of a customer.
2. We could have also applied more advanced algorithms from deep learning, which would have resulted in even better understanding of the interactions between various features.
3. Also, we could have done some further investigation as to figure out the reason for overfitting of random forest and try to work on it and choose the final model with least RMSE.