

PROJECT REPORT



# CUSTOMER LIFETIME VALUE ANALYSIS IN AUTO INSURANCE INDUSTRY

**“Submitted towards partial fulfilment of the criteria for award of PGPDSE by GLIM”**

**Submitted by**

|  |  |
| --- | --- |
| **Student Name** | **SIS ID** |
| Uday Kumar Sahukari | 5JT0U55V4I |
| Shaik Shahul Ummar Sharief | 76YG6RC47E |
| Harish Chandanam | 06QHIC3DZ1 |
| Tippani Sirisha | MYC279Y0LM |
| Ved Kumar Verma | B12FNI6CXK |

**Batch: PGPDSE\_HYDERABAD\_JUN2019**

**Mentor: Mr. Srikar Muppidi**





**Abstract:**

# Abstract

Customer lifetime value (CLTV) is one of the most important metrics to measure at any growing company. If you want your business to acquire and retain highly valuable customers, then it's essential that your team learns what customer lifetime value is. Customer lifetime value is the metric that indicates the total revenue a business can reasonably expect from a single customer account. Businesses use this metric to identify significant customer segments that are the most valuable to the company. LTV tells companies how much revenue they can expect one customer to generate over the course of the business relationship. The extracted features from the data are fed to the machine learning regression methods to build a model. Feature selection pre-processing steps are used to enhance the performance and scalability of the regression methods. As of now the results show that regression model has a good fit compared to other ensemble methods with the model being overfit.

**Keywords**:

Keywords — Auto mobile Insurance, Customer Life Time Value, Policy, Premium, Customer Behavior, Revenue.

❏ Techniques: Machine Learning – Supervised Learning Regression.

❏ Tools: Python, Tableau.

❏ Domain: Automobile Insurance.

# Certification of Completion

I hereby certify that the project titled “Customer Value Analysis in Auto Insurance” was undertaken and completed under my guidance and supervision by Harish Chandanam, Uday Kumar Sahukari, Shaik Shahul Ummar Sharief, Tippani Sirisha, Ved Kumar Verma students of the June 2019 batch of the Post Graduate Program in Data Science & Engineering, Hyderabad.

Date: 14th Nov 2019

Place: Hyderabad.

# Declaration

I declare that the project entitled “Prediction of Online Shoppers Purchasing Intention” is a project work carried out by me under the supervision and guidance of Srikar Muppidi, for the award of degree PGP DSE, and this has not been previously submitted for the award of any Degree, Diploma or other similar title of any other University/ Institute.

Date:

Place: Hyderabad Group members:

Harish Chandanam

Shaik Shahul Ummar Sharief

Uday Kumar Sahukari

Tippani Sirisha

Ved Kumar Verma

# Acknowledgements

At the outset, we are indebted to our Mentor Mr. Srikar Muppidi for his time, valuable inputs and guidance. His experience, support and structured thought process guided us to be on the right track towards completion of this project.

We also thank all the course faculty of the DSE program for providing us a strong foundation in various concepts of analytics & machine learning.

Last but not the least, we would like to sincerely thank our respective families for giving us the necessary support, space and time to complete this project.

We certify that the work done by us for conceptualizing and completing this project is original and authentic.

Harish Chandanam

Uday Sahukari

Shaik Shahul Ummar Sharief

Tippani Sirisha

Ved Kumar Verma

Date: 14th Nov 2019

Place: Hyderabad

# Table of Contents

[CUSTOMER VALUE ANALYSIS IN AUTO INSURANCE INDUSTRY 1](#_Toc24718037)

[Abstract 2](#_Toc24718038)

[Certification of Completion 3](#_Toc24718039)

[Declaration 4](#_Toc24718040)

[Acknowledgements 5](#_Toc24718041)

[Table of Contents 6](#_Toc24718042)

[Executive Summary 8](#_Toc24718043)

[Chapter 1 - Project Introduction 9](#_Toc24718044)

[Problem Statement 9](#_Toc24718045)

[Dataset Description 10](#_Toc24718046)

[Chapter 2 – Literature Review 11](#_Toc24718047)

[Regression Models and Their Explanation 11](#_Toc24718048)

[Model Performance Evaluation Metrics for Regression: 13](#_Toc24718049)

[Chapter 3 - Data Cleaning 14](#_Toc24718050)

[Missing value Treatment: 14](#_Toc24718051)

[Outlier Treatment: 15](#_Toc24718052)

[Correction of Inconsistent Data: 15](#_Toc24718053)

[Chapter 4 - Exploratory Data Analysis 16](#_Toc24718054)

[Visual EDA: 16](#_Toc24718055)

[Statistical EDA: 16](#_Toc24718056)

[Chapter 5 – Base Model Validation 23](#_Toc24718057)

[Chapter 6 - Feature Engineering 24](#_Toc24718058)

[Forward Selection: 25](#_Toc24718059)

[Backward Elimination: 25](#_Toc24718060)

[Binning: 26](#_Toc24718061)

[One-Hot Encoding: 26](#_Toc24718062)

[Chapter 7 – Model’s Evaluation After Feature Engineering 28](#_Toc24718063)

[Finding the best sample by random state for each model: 28](#_Toc24718064)

[Chapter 8 – Recommendations and Actionable Insights 30](#_Toc24718065)

[Chapter 9 - Bibliography 31](#_Toc24718066)

Abbreviations used



|  |  |
| --- | --- |
| **Abbreviation** | **Expansion** |
| LR | Linear Regression |
| R | Ridge Regression |
| L | Lasso Regression |
| DT | Decision Tree Regressor |
| ET | Extra Tree Regressor |
| RFR | Random Forest Regressor |
| ETR | Extra Trees Regressor |
| ABR | Ada Boost Regressor |
| GBR | Gradient Boosting Regressor |
| Bag\_DT | Bagging Decision Trees |
| Boost\_DT | Boosting Decision Tree |
| CLTV | Customer Life Time Value |



# Executive Summary

**Background & need for study**: Customer Lifetime Value represents a customer’s value to a company over a period of time. It’s a competitive market for insurance companies in 2019, and insurance premium isn’t the only determining factor in a customer’s decisions. CLV is a customer-centric metric, and a powerful base to build upon to retain valuable customers, increase revenue from less valuable customers, and improve the customer experience overall. Finding your Customer Lifetime Value will make you think, not just about the sale of policy, but about the full customer journey: when, where, why, for how much, and how often do your customers make a purchase of policy. Using CLV effectively can improve customer acquisition and customer retention, prevent churn, help you plan your marketing budget, measure the performance of your ads in more detail, and much more.

**Scope & Objectives**: Once you’ve performed a CLV analysis on all your current customers, you’ll know how much it makes sense to spend on acquisition of new customers. You’ll also know which acquisition channels produce the highest value customers, and can repeat the strategies you used to find them. Once you have a solid data profile of what characteristics your VIPs have, you can use predictive analytics to get a strong idea of which new customers will likely be future VIPs, and focus on these customers with personalized messaging and offers You can also make use of a look-alike model, to target similar profiles. Now that you now know which premium and number of policies a customer takes; the door is open for personalized messaging to send the right offer to the right person at the right time.

**Approach & Methodology:** After processing the dataset. Various regression algorithms are used to predict customer lifetime value based on set of independent variables like income, monthly premium, months since last claim, months since policy inception, number of open complaints, coverage, renewal offer type and vehicle class. Various statistical tests are used to understand which independent features influence the CLTV and removed the features that do not influence the CLTV. RMSE and R^2 are the metrics used to evaluate the model. Base model is implemented using the linear regression algorithm, and looking forward to apply ensemble methods and other regression techniques for a better model.

**Key learnings:**

**Recommendations & Actionable insights:**

We could see that CLTV of existing customers is high and affording new customers would make the company invest more. In order to attract more customers to Auto insurance company, below actions need to be taken into consideration.

* Most of the customers are purchasing the policy personal l3 type indicating we can expect to get more customers of such type.
* Most of our customers are coming from Sub Urban regions.
* Most of our customers have opted for policy are having Medium size - Four door car indicating we can market on new customers of this behavior in future followed by SUV and 2 door cars.
* Number of policies getting expired per day is 150 +- 20 on avg. So, should target to make more than 150 renewals per day on avg then we can stay in profits.



# Chapter 1 - Project Introduction

## Problem Statement

The Auto Insurance industry is data-intensive with typically massive graveyards of unused and unappreciated processing data. As insurance companies face increasing pressure to stay profitable, understanding customer needs and preferences becomes a critical success factor. New models of proactive risk management are being increasingly adopted by major insurance companies and financial institutions. Through Data mining and advanced analytic techniques, Auto insurance companies are better equipped to manage market uncertainty, minimize fraud, and control exposure risk. But in order to discover the set of critical success factors that will help insurance companies reach their strategic goals, they need to move beyond standard business reporting and sales forecasting. By applying data mining and predictive analytics to extract actionable intelligent insights and quantifiable predictions, insurance companies can gain insights that encompass all types of customer behavior, including channel transactions, account opening and closing, default, fraud and customer departure.

The best auto insurance customer retention strategy for existing customers is to classify each type of customer (silent attrition, ideal and unhappy) and create appropriate initiatives to change their behavior. For instance, customers in “silent attrition” are those that have reduced or stopped using a product(policy), but where the account is still open. You must understand why they are no longer using your product policies and create initiatives to change their behavior.

Customers that are exiting are those customers that have started the process of moving their business to another company or are in the process of considering that move. The first step in creating auto insurance customer retention strategies for exiting customers is to identify which customers are in each camp. For customers in the process of moving their business you will need to understand the product drop cycle, i.e. the order in which customers drop your products before leaving. With this information you can create effective customer retention strategies to target those customers.

Hence by working on the provided dataset we will be reducing the Customer Churn and provide suggestions.

## Dataset Description

This dataset has 24 columns (features) and rows (records).

|  |  |
| --- | --- |
| **Numerical Feature** | **Feature Description** |
| Income | Customer annual income in USD |
| Monthly Premium Auto | Monthly Premium for auto insurance |
| Total Claim Amount | Amount claimed till data |
| Months Since Last Claimed | Number of months since the customer last claimed his insurance |
| Months Since Policy Inception | Number of months before the policy commenced |
| Number of Open Complaints | Numbers of unresolved complaints made by customer |
| Number of Policies | Number of policies taken by the customer |
| Customer Life time value – (Y) | CLV of the customer for the auto insurance company |

|  |  |
| --- | --- |
| **Categorical Feature** | **Category Description** |
| State | 6 different States |
| Response | Whether the customer responded to the marketing. |
| Coverage | Basic, Premium and Extended coverage |
| Education | Maximum qualification |
| Employment Status | Employed/Unemployed/  Medical leave/disabled/retired |
| Gender | Male/Female |
| Location Code | Sub-urban/urban/rural |
| Marital Status | Single/married/divorced |
| Policy Type | Corporate/personal/Special |
| Renew Offer Type | 4 different offers |
| Sales Channel | Web/Agent/Call Center/Branch |
| Vehicle Class | Two -Door/Four-Door/Luxury SUV/Luxury Car/Sports Car/SUV |
| Vehicle Size | Midsize/Small/Large |

|  |  |
| --- | --- |
| **Object Feature** | **Description** |
| Customer | Unique Customer ID |
| Effective to date | Expiry date of Policy |

# Chapter 2 – Literature Review

The source dataset received has been prepared to ensure that the fields are cleaned up, the values are suitable for model building and the variable names are self-explanatory. The broad approach for data preparation can be outlined as:

Table 4 – Data pre-processing steps

|  |  |  |
| --- | --- | --- |
| **Data Encoding** | **Outlier Treatment** | **Standardization** |
| All categorical variables which do not have any order are converted into numerical by one hot encoding. For categorical data with certain order are converted to numerical using Label Encoding. | Box plot is drawn for Independent features against Target variable and outlier had been detected. Since the outliers are legitimate, we have decided to retain them in data. | Standard Scalar function from Sci-kit learn library since the numerical variable are of different scale in order to obtain better performance. |

## Regression Models and Their Explanation

Various regression algorithms have been used to analyze customer lifetime value and to identify the extent to which each independent variable is significant.

The model building exercise has also considered cross validation and tuning techniques to ensure that the models built perform well when used for prediction.

The regression algorithms used for Commercial intent prediction include,

* Linear regression
* Decision Tree Regressor
* Random Forest Regressor
* Extra Tree/Trees Regressor
* AdaBoost Regressor
* Bagging Regressor

Linear Regressor: Linear regression is a statistical approach for modelling relationship between a dependent

variable with a given set of independent variables.

Note: In this article, we refer dependent variables as response and independent variables as features for simplicity.

In order to provide a basic understanding of linear regression, we start with the most basic version of linear regression, i.e. Simple linear regression.

Simple Linear Regression

Simple linear regression is an approach for predicting a response using a single feature.

It is assumed that the two variables are linearly related. Hence, we try to find a linear function that predicts the response value(y) as accurately as possible as a function of the feature or independent variable(x)

Decision Tree Regressor:

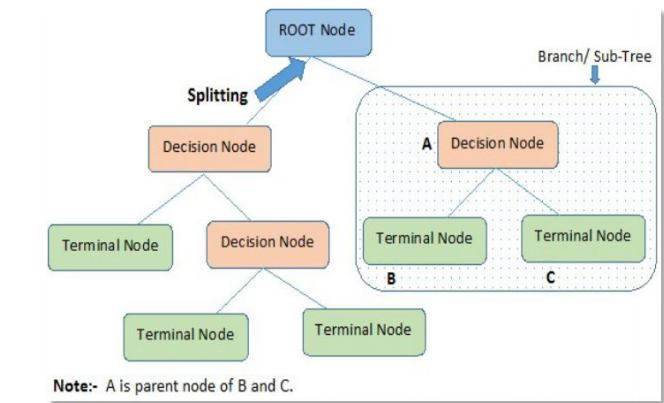
Decision Tree is a decision-making tool that uses a flowchart-like tree structure or is a model of decisions and all of their possible results, including outcomes, input costs and utility. Decision-tree algorithm falls under the category of supervised learning algorithms. It works for both continuous as well as categorical output variables.

The branches/edges represent the result of the node and the nodes have either:

Decision Tree Regression:

Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. Continuous output means that the output/result is not discrete, i.e., it is not represented just by a discrete, known set of numbers or values.

Continuous output example: A profit prediction model that states the probable profit that can be generated from the sale of a product.



## Model Performance Evaluation Metrics for Regression:

The various models built, must be evaluated based on certain model performance measures to identify the most robust models. Model accuracy alone may not be enough to evaluate a model. Hence the following model performance measures have been used to evaluate the models, MSE, R2\_Score, MAPE.

MSE: Mean Squared Error basically measures average squared error of our predictions. For each point, it calculates square difference between the predictions and the target and then average those values.

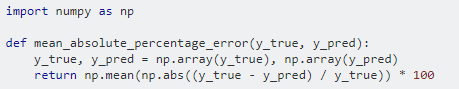
Considering the Customer Lifetime Value (CLTV) as the target variable, the MSE for our base linear regression model is 40782604.59607423 for Training data and 35832657.97238506 for testing data. Testing error is marginally lower than the training error.

R2\_Score: The coefficient of determination R2, it is closely related to MSE, but has the advantage of being scale-free — it doesn’t matter if the output values are very large or very small, the R² is always going to be between -∞ and 1. When R² is negative it means that the model is worse than predicting the mean.

Similarly, r2\_score for training data is 0.1751 and for testing data r2\_score is 0.14558.

MAPE: Mean Absolute Percentage Error.

This is a function which is not defined In Sci-kit Learn library by default. We have to define the function and run.



MAPE's formula involves dividing by the actual observed value.

MAPE ... puts a heavier penalty on forecasts that exceed the actual than those that are less than the actual. For example, the MAPE is bounded on the low side by an error of 100%, but there is no bound on the high side’.

The quoted (Makridakis, 1993) paper gives a nice example for the asymmetry, when the predicted value is 150 and the forecast is 100, MAPE is |150−100150|=33.33%, while when the predicted value is 100 and the forecast is 150 MAPE is |100−150100|=50% despite the fact that both forecasts are wrong by 50 units!

What the above references, and the number of other sources, show is that if you use MAPE as a criterion for selecting your forecasts, this would lead to biased and underestimated results. Moreover, you run into problems when the predicted value is equal to zero.

# Chapter 3 - Data Cleaning

## Missing value Treatment:

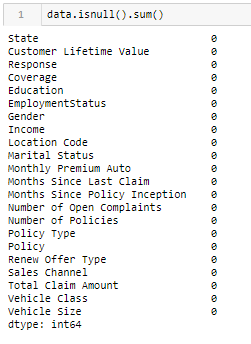
In any real-world data set, there are usually few null values. It doesn’t really matter whether it is regression, classification or any other kind of problem no model can handle these NULL or NaN values on its own so we need to intervene. First of all we need to check whether we have null values in our dataset or not. We can do that using the isnull () method. There are various ways for us to handle this problem. The easiest way to solve this problem is by dropping the rows or columns that contain null values.

However, it is not the best option to remove the rows and columns from our dataset as it can lead to loss of valuable information. So, if you have 300K data points then removing 2–3 rows won’t affect your dataset, whereas if you only have 100 data points and out of which 20 have NaN values for a particular field then you can’t simply drop those rows. In real life datasets it can happen quite often that you have large number of NaN values for a particular field.

Suppose we are collecting the data from a survey, then it is possible that there could be an optional field which let’s say 20% people left blank. So, when we get the dataset then we need to understand that the remaining 80% data is still useful so rather than dropping these values we need to somehow substitute the missing 20% values. We can do this with the help of Imputation.

Imputation is simply the process of substituting the missing values of our dataset. We can do this by defining our own customized function or we can simply perform imputation by using the Imputer class provided by sklearn.In Imputer() you can pass on the axis and strategy. strategy could be mean, median etc.

There are NO MISSING VALUES are present in the given Dataset.



## Outlier Treatment:

In statistics, outliers are data points that don’t belong to a certain population. It is an abnormal observation that lies far away from other values. An outlier is an observation that diverges from otherwise well-structured data.

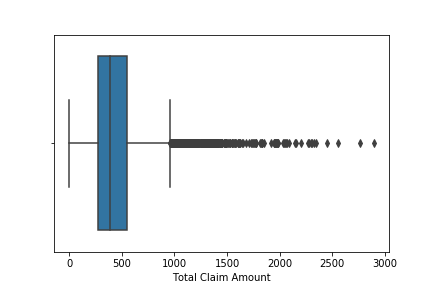
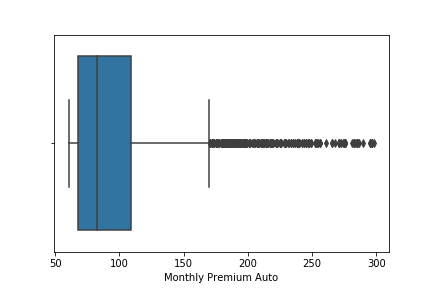
For Example, you can clearly see the outlier in this list: [20,24,22,19,29,18,4300,30,18] is 4300 making it right skewed.

It is easy to identify it when the observations are just a bunch of numbers and it is one dimensional but when you have thousands of observations or multi-dimensions, you will need more clever ways to detect those values.

Methods to Find and Treat Outliers:

1. Boxplot
2. IQR (Inter Quartile Range)
3. MICE

**There are a lot of outliers in each column of our dataset. But we didn’t consider removing or treating them as according to our domain knowledge they are considered as Extreme Values which are important.**



## Correction of Inconsistent Data:

Do some preliminary text pre-processing

For e.g.:

1. Making everything lower case.

2. Spell checks.

3. Data Duplicates.

4. Irregular Date format. (If any)



# Chapter 4 - Exploratory Data Analysis

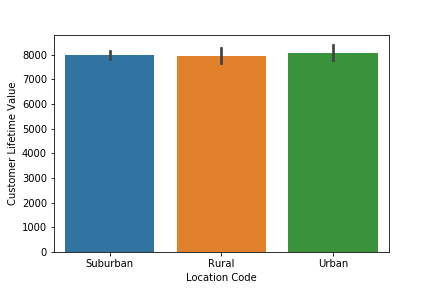
Here, we have done EDA both visually and statistically:

## Visual EDA:

The purpose of exploratory data analysis is two-fold:

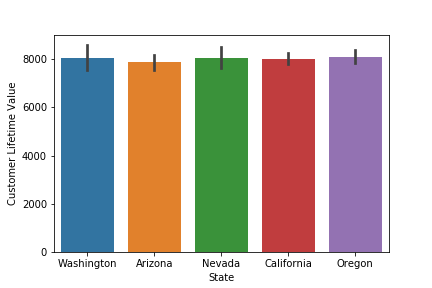
* To understand the data in terms of what features are significantly important to our target variable, Customer lifetime value.
* Get insights on various features.

**All variables W.R.T Customer Lifetime Value:**

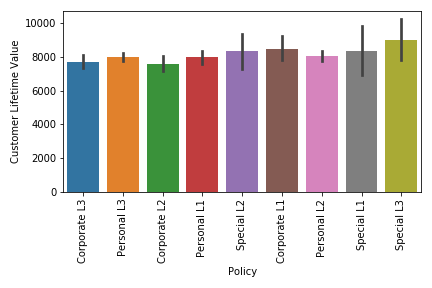
****

This Bar plot clearly shows that no matter what location code a person is from the average customer lifetime value is same.

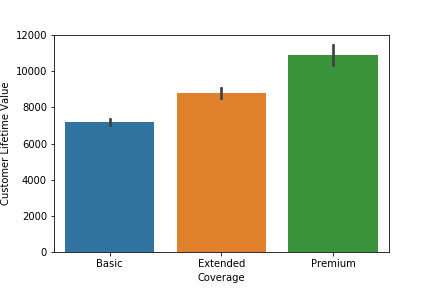
This is also proved using ANOVA statistical test.

****

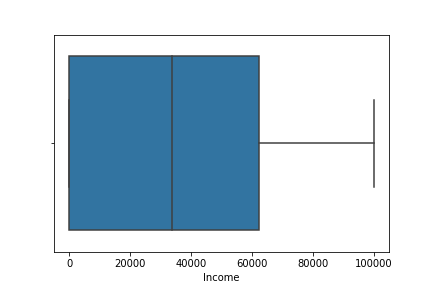
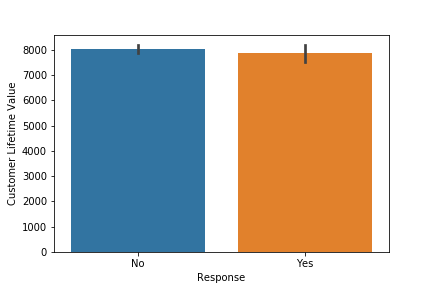
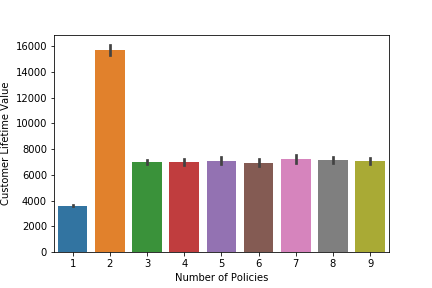
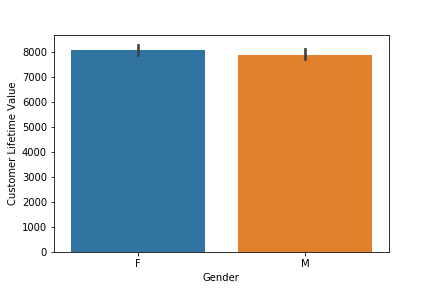
This Bar plot clearly shows that no matter which state a person is from he/she will have the same Customer lifetime value. This is also proved using ANOVA statistical test.



This is the CLTV Vs Policy, and we can clearly say that the average customer lifetime value is almost the same for every policy a customer has taken, and there is no pattern between policies to consider this feature as important.



This bar plot shows us that there is a relationship between CLTV and Coverage, the better coverage higher is the CLTV, in our case premium is considered as the best coverage and hence has the highest CLTV.

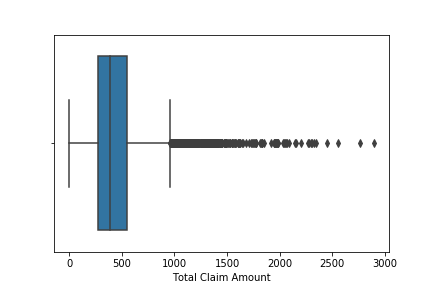
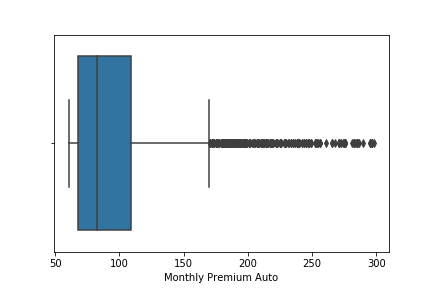


This Bar plot clearly shows that there is no bias in the data for male and female and both will have the same Customer lifetime value. This is also proved using ttest statistical test.

This bar plot tells us that people who have take policies>2 have similar trend and people who have taken only 1 policy have low CLTV comparatively and the CLTV is highest for people who have taken 2 policies.

In our data, response signifies whether a person has accepted the marketing call (YES) or he declined the marketing call (NO), so no matter what the response is we can say that the average CLTV of the customer is same.

Boxplots are useful to determine whether our data has outlier’s, we can see that the income independent variable/feature has no outliers.



This is the boxplot for the total claim amount a customer claimed w.r.t his/her premium, we don’t consider these values as outlier’s but we consider these as extreme values, which are quite important for us when modelling, removing these values will result in loss of information and we would end up creating a bad model.

This is the boxplot for the monthly premium amount a customer has to pay w.r.t his/her policy, we don’t consider these values as outlier’s but we consider these as extreme values, which are quite important for us when modelling, removing these values will result in loss of information and we would end up creating a bad model.

## Statistical EDA:

Considering CLTV (Customer Lifetime Value) as the target variable, we shall try to understand how each of the independent variables are contributing towards the target variable.

Since our target variable is a continuous variable, we will have to perform ANOVA to understand how significant are the independent variables towards target variable.

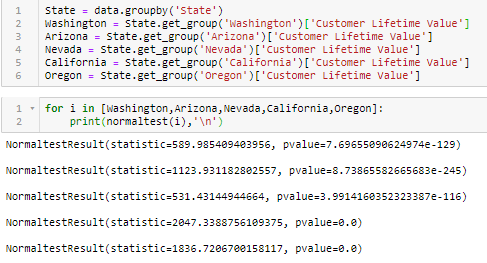
For ANOVA:

- Null hypothesis is that there is no significant difference among the groups.

- Alternative hypothesis is that there is at least one significant difference among the groups.

Here we have performed statistical test for 13 variables against CLTV. Below are the results.

# 1 State v/s Customer Lifetime Value:



\* ALL STATE HAVE SAME MEAN'S OF CLTV.

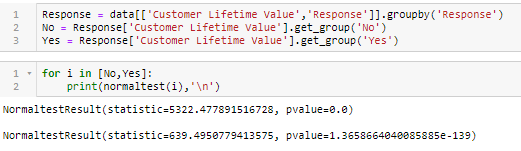
\* p value > 0.05 implies that there is no significant difference in the mean of target variable which means 'State' feature is not significant for predicting 'Customer Lifetime Value'.

CLTV of all the 'States' follow a normal distribution.

Hence, we can perform ANOVA test.



# 2 Customer Response to marketing calls v/s Customer Lifetime Value:



\*CLTV of all the 'Response' follow a normal distribution.

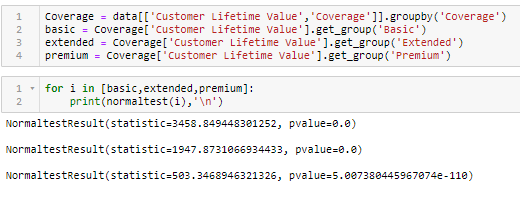
\*Hence, we can perform ANOVA test or test of mean for independent categories.



\* RESPONSE HAVE SAME MEAN'S OF CLTV

\* p value > 0.05 implies that there is no significant difference in the mean of target variable which means 'Response' feature is not significant for predicting 'Customer Lifetime Value'.

# 3 Coverage Type v/s Customer Lifetime Value:



\* MEAN'S ARE NOT SAME FOR COVERAGE

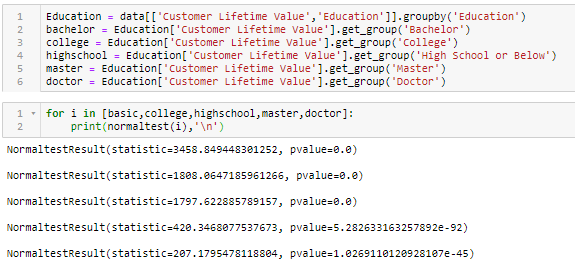
\* p value < 0.05 implies that there is significant difference in the mean of target variable for at least one group of 'Coverage' which means 'Coverage' feature can be a significant for predicting 'Customer Lifetime Value'.

\* CLTV of all the 'Coverage' follow a normal distribution.

Hence, we can perform ANOVA test.



# 4 Education v/s Customer Lifetime Value:





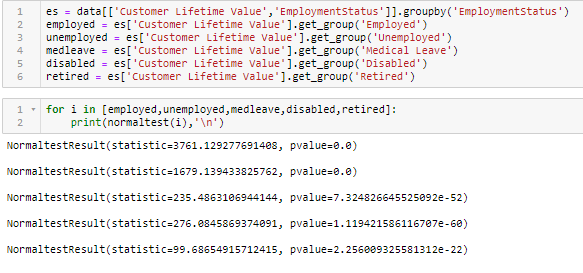
CLTV of all the categories of 'Education' follow a normal distribution.

Hence, we can perform ANOVA test.

CLTV of all the categories of 'Education' follow a normal distribution.

Hence, we can perform ANOVA test.

# 5 Employment Status v/s Customer Lifetime Value:





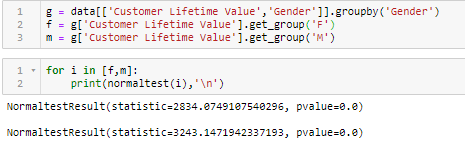
\* MEAN'S ARE NOT SAME FOR Employment Status

\* p value < 0.05 implies that there is significant difference in the mean of target variable for at least one group of 'Employment Status' which means 'Employment Status' feature can be a significant for predicting 'Customer Lifetime Value'.

CLTV of all the categories of 'Employment Status' follow a normal distribution.

Hence, we can perform ANOVA test.

# 6 Gender v/s Customer Lifetime Value:





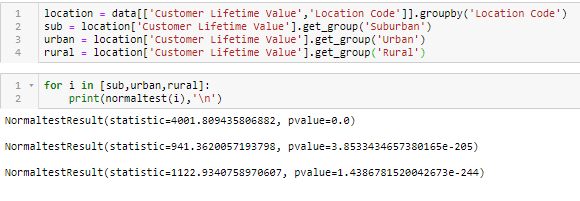
\* MEAN'S ARE SAME FOR GENDER

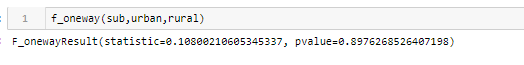
\* p value > 0.05 implies that there is no significant difference in the mean of target variable for 'Gender' which means 'Gender' feature is not significant for predicting 'Customer Lifetime Value'

CLTV of all the categories of 'Gender' follow a normal distribution.

Hence, we can perform ANOVA test or test of mean for independent features.

# 7 Location Code v/s Customer Lifetime Value:





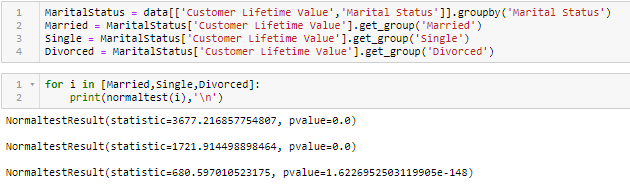
\* MEAN'S ARE SAME FOR LOCATION CODE

\* p value > 0.05 implies that there is no significant difference in the mean of target variable for 'Location Code' which means 'Location Code' feature is not significant for predicting 'Customer Lifetime Value'.

CLTV of all the categories of 'Location Code' follow a normal distribution.

Hence, we can perform ANOVA test.

# 8 Marital Status v/s Customer Lifetime Value:





\* MEANS ARE NOT SAME Marital Status

\* p value < 0.05 implies that there is significant difference in the mean of target variable for at least on Group of 'Marital Status' which means 'Marital Status' feature can be significant for predicting 'Customer Lifetime Value'.

CLTV of all the categories of 'Location Code' follow a normal distribution.

Hence, we can perform ANOVA test.

# 9 Marital Status v/s Customer Lifetime Value:



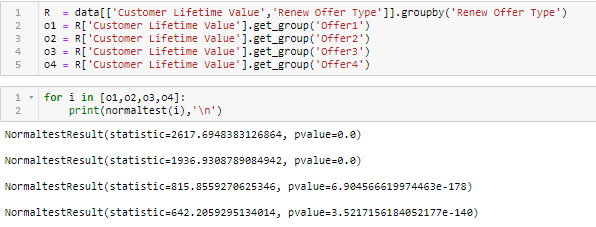
\* MEANS ARE SAME Marital Status

\* p value > 0.05 implies that there is no significant difference in the mean of target variable for 'Policy' which means 'Policy' feature is not significant for predicting 'Customer Lifetime Value'.

CLTV of all the categories of 'Policy' follow a normal distribution.

Hence, we can perform ANOVA test.

# 10 Renew Offer Type vs Customer Lifetime Value:





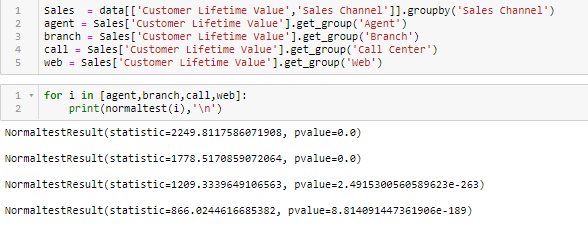
\* MEANS ARE NOT SAME FOR Renew Offer Type

\* p value < 0.05 implies that there is significant difference in the mean of target variable for at least on Group of 'Renew Offer Type' which means 'Renew Offer Type' feature can be significant for predicting 'Customer Lifetime Value'.

CLTV of all the categories of 'Policy' follow a normal distribution.

Hence, we can perform ANOVA test.

# 11 Sales Channel vs Customer Lifetime Value:





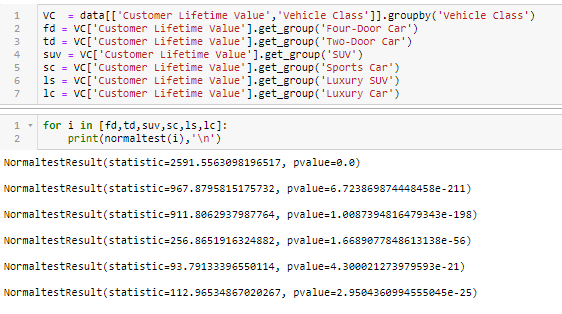
CLTV of all the categories of 'Sales Channel' follow a normal distribution.

Hence, we can perform ANOVA test.

\* MEANS ARE SAME for Sales Channel

\* p value > 0.05 implies that there is no significant difference in the mean of target variable for 'Sales Channel' which means 'Sales Channel' feature is not significant for predicting 'Customer Lifetime Value'.

# 12 Vehicle Class vs Customer Lifetime Value:





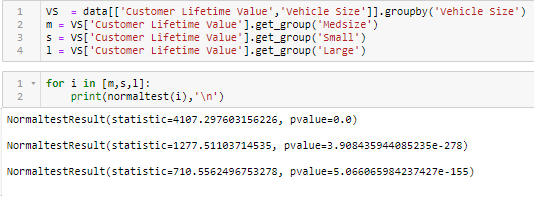
\* MEANS ARE NOT SAME FOR Vehicle Class.

\* p value < 0.05 implies that there is significant difference in the mean of target variable for at least on Group of 'Vehicle Class' which means 'Vehicle Class' feature can be significant for predicting 'Customer Lifetime Value'.

CLTV of all the categories of 'Vehicle Class' follow a normal distribution.

Hence, we can perform ANOVA test.

# 13 Vehicle Size v/s Customer Lifetime Value:





\* MEANS ARE SAME for Vehicle Size

\* p value > 0.05 implies that there is no significant difference in the mean of target variable for 'Vehicle Size' which means 'Vehicle Size' feature is not significant for predicting 'Customer Lifetime Value'.

CLTV of all the categories of 'Vehicle Size' follow a normal distribution.

Hence, we can perform ANOVA test.

# Chapter 5 – Base Model Validation

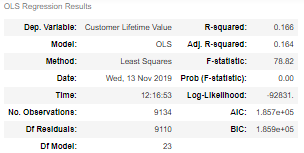
After Looking at the base model and the p-value of the feature's, we know that the Hypothesis for the feature's is:

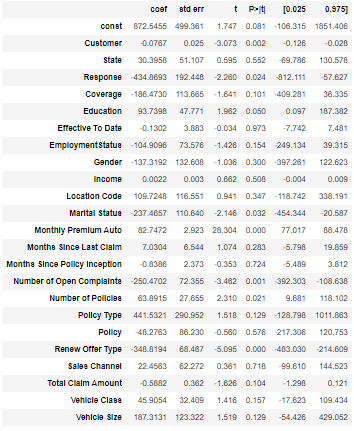
* H0: Feature is not significant
* Ha: Feature is significant
* But we just can’t conclude the significance of the feature's just by base model and also without using any of the feature engineering technique's we have at our disposal. So, we will first try to do the statistical test of the feature for the feature selection, we can also use the forward selection and backward elimination, we will use the Variance inflation factor.











# Chapter 6 - Feature Engineering

What are the fundamental techniques of Feature Engineering for machine learning?

What is a feature and why we need the engineering of it?

Basically, all machine learning algorithms use some input data to create outputs. This input data comprises features, which are usually in the form of structured columns. Algorithms require features with some specific characteristic to work properly. Here, the need for feature engineering arises. I think feature engineering efforts mainly have two goals:

1. Preparing the proper input dataset, compatible with the machine learning algorithm requirements.
2. Improving the performance of machine learning models.

The features you use influence more than everything else the result. No algorithm alone, to my knowledge, can supplement the information gain given by correct feature engineering.

List of Techniques Used:

1. Feature Selection.

2. Binning.

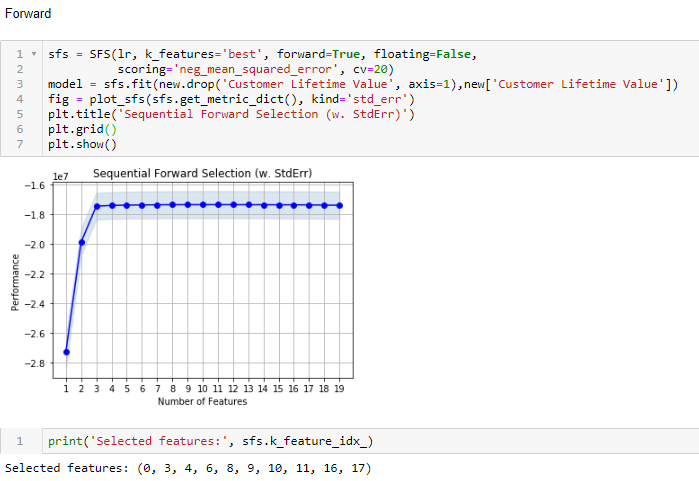
3. One-Hot Encoding.

**Feature selection** is the process of selecting a subset of relevant attributes to be used in making the model in machine learning. Effective feature selection eliminates redundant variables and keeps only the best subset of predictors in the model which also gives shorter training times. Besides this, it avoids the curse of dimensionality and enhance generalization by reducing overfitting.

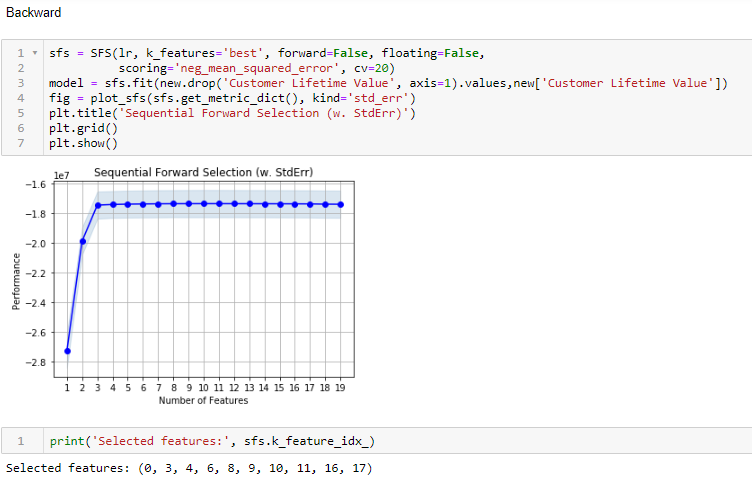
In this project, feature selection techniques are applied to improve the regression performance and/or scalability of the system. Thus, we aim to investigate if better or similar regression performance can be achieved with a smaller number of features. An alternative of feature selection is the use a feature extraction technique such as Principal Component Analysis for dimensionality reduction. However, in this case, the features in the reduced space will be the linear combinations of 6 numerical attributes, which do not have any liner relationship with each other. Therefore, it has been deemed appropriate to apply feature selection instead of feature extraction within the scope of this research. For feature ranking, instead of wrapper algorithms that require a learning algorithm to be used and consequently can result in reduced feature sets specific to that regressor. Correlation Attribute Evaluation, Forward selection, backward elimination, recursive feature elimination and feature importance with the help of extra tree regressor algorithm were used in our experiments. the aim is to maximize the relevance between the selected set of features and class variable while avoiding the redundancy among the selected features. Thus, maximum accuracy is aimed to be obtained with minimal subset of features.

Besides, considering the real-time usage of the proposed system, achieving better or similar regression performance with a smaller number of features will improve the scalability of the system since a smaller number of features will be kept track during the session.

## Forward Selection:



## Backward Elimination:



**- Surprisingly Both the forward and backward selection gave us the same features to select for our model, so we will be sticking to the same feature's.**

## Binning:

Binning or grouping data (sometimes called quantization) is an important tool in preparing numerical data for machine learning, and is useful in scenarios like these: To mitigate the bias in the model, you might transform the data to a uniform distribution, using the quantiles (or equal-height) method. Binning is the process of converting the numerical variable to categorical variable. We have created binning categorical variable which gives the generalized information about the variable for example in our dataset:

- Though the feature's months since policy inception, months since last claim, number of open complaints and number of policies are all numerical, but they are discrete number's and we will consider them as categorical feature's while preparing the model.

- Firstly, according to our EDA, we saw that the number of policies >= 3 have similar trend so we will group all of them as 3.



## One-Hot Encoding:

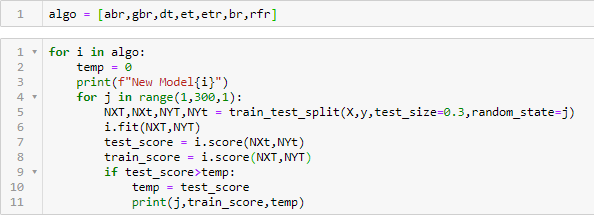
Second, when we convert the numerical feature's to categorical, our normal practice is label encoding for ordinal data and one hot for nominal data, but we can also use one hot encoding for ordinal data if there isn’t any curse of dimensionality, so we will convert the categorical to numerical with one-hot encoding / dummification.

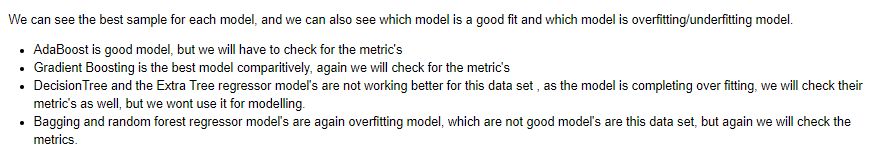


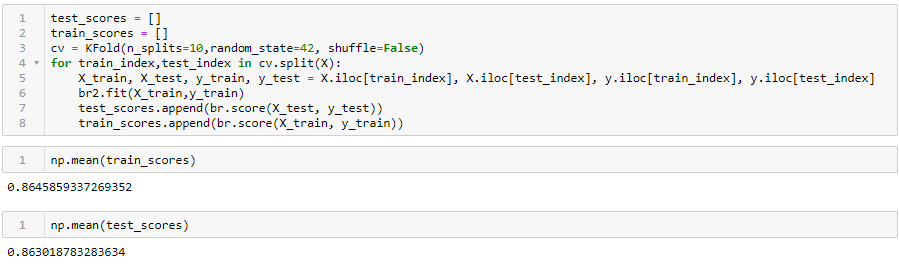
# Chapter 7 – Model’s Evaluation After Feature Engineering

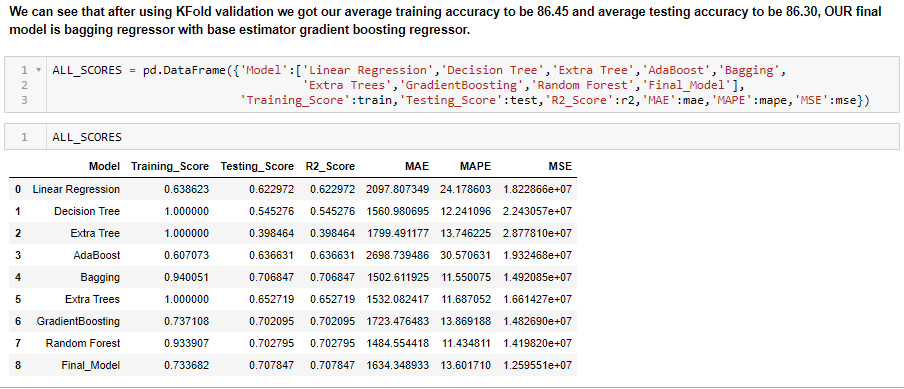
One of the purposes of this project is to get the analyses results of the customer’s lifetime value to determine if that particular customer is going to a good customer in the coming future. Linear regression, decision tree regressor, random forest regressor, boosting, bagging, extra tree regressor are used with 5fold CV. The accuracy, RMSE and MAE are presented for each regressor.

## Finding the best sample by random state for each model:









# Chapter 8 – Recommendations and Actionable Insights



We could see that CLTV of existing customers is high and affording new customers would make the company invest more. In order to attract more customers to Auto insurance company, below actions need to be taken into consideration.

* Most of the customers are purchasing the policy personal l3 type indicating we can expect to get more customers of such type.
* Surprisingly, we can see that in our customer base, doctors and people having masters are very less compared to college going, bachelors and people who are educated up to high school or below indicating highly educated people are not interested in buying policies.
* Most of our customers are coming from Sub Urban regions.
* Most of our customers have opted for policy are having Medium size - Four door car indicating we can market on new customers of this behavior in future followed by SUV and 2 door cars.
* Number of policies getting expired per day is 150 +- 20 on avg. So, should target to make more than 150 renewals per day on avg then we can stay in profits.
* Number of policies sold through the Agent are more compared to Branch, Call center and Web. Hence, the Agency should invest more on sales through Agent channel instead of spending on Web sales.
* Avg. Customer Life Time Value is high for people having Luxury Car & Luxury SUV irrespective of vehicle size, so the agency should focus on acquiring customer base having cars of these classes.



# Chapter 9 - Bibliography



1. Introduction to Auto Insurance , <https://www.investopedia.com/terms/a/auto-insurance.asp>

2. Nicholas Kashmerick, Learning To remove Internet Advertisement, 3rd Int. Conf. of

Autonomous Agent, 1999

3. Sakar, C.O., Polat, S.O., Katircioglu, M. et al. Neural Comput &amp; Applic (2018).

[Web

Link]

4. B.D. Davision. Recognizing Nepotistic links on the Web. Proceeding of AAAI 2000.

5. S. Paek and J. R. Smith, Detecting Image Purpose in World-Wide Web Documents,

SPIE/IS&amp;T Photonics West, Document Recognition, January, 1998.

6. Guohua Hu, Qingshan Zhao” Study to Eliminating Noisy Information in Web Pages

7. Kushmerick, 1999] Nicholas Kushmerick. Learning to remove Internet advertisements.

Agnets-1999, 1999..

8. Kao et al., 2002] Hung-Yu Kao, Ming-Syan Chen Shian-Hua Lin, and Jan-Ming Ho,

Entropy-Based Link Analysis for Mining Web Informative Structures. CIKM-2002, 2002.

9. H. Y. Kao, J. M. Ho, and M. S. Chen, Wisdom Web intrapage informative structure

mining based on document object model in IEEE Trans KDD, 2005.

10. YI L. et LIU B. (2003), “Web Page Cleaning for Web Mining through Feature

Weighting”, in Proceedings of Eighteenth International Joint Conference on Artificial

11. S. Debnath, P. Mitra, and C. L. Giles. Automatic extraction of informative blocks from

webpages. In ACM Symposium on Applied Computing, pages 1722–1726, 2005.

12. L. Yi, B. Liu, and X. Li. Eliminating noisy information in web pages for data mining. In

Proceedings of the International ACM Conference on Knowledge Discovery and Data

Mining, pages 296– 305, 2003.

13. Diao, Y., Lu, H., Chen, S., and Tian, Z., TowardLearningBased Web Query Processing,

In Proceedings of International Conference on Very Large Databases, 2000, pp. 317-328.

14. Kaasinen, E., Aaltonen, M., Kolari, J., Melakoski, S., and Laakko, T., Two Approaches

to Bringing Internet Services to WAP Devices, In Proceedings of 9th International

World-Wide Web Conference, 2000, pp. 231-246.

15. Wong, W. and Fu, A. W., Finding Structure and Characteristics of Web Documents for

Classification, In ACM SIGMOD Workshop on Research Issues in Data Mining and

Knowledge

Discovery (DMKD), Dallas, TX., USA, 2000.

16. Gupta, S., Kaiser, G., Neistadt, D. and Grimm, P., DOMbased Content Extraction of

HTML Documents, In the proceedings of the Twelfth World Wide Web conference(WWW

2003), Budapest, Hungary, May 2003.

17. Cai Deng, Yu Shipeng, and Wen Jirong, et al. VIPS: a Vision Based Page Segmentation

Algorithm[R] , Microsoft Technical Report: (MSR-TR-2003-79) ,2003.

18. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O.,

Blondel,

|  |
| --- |
| Appendix |
| Detailed data dictionary |

|  |  |
| --- | --- |
| **Numerical Feature** | **Feature Description** |
| Income | Customer annual income in USD |
| Monthly Premium Auto | Monthly Premium for auto insurance |
| Total Claim Amount | Amount claimed till data |
| Months Since Last Claimed | Number of months since the customer last claimed his insurance |
| Months Since Policy Inception | Number of months before the policy commenced |
| Number of Open Complaints | Numbers of unresolved complaints made by customer |
| Number of Policies | Number of policies taken by the customer |
| Customer Life time value – (Y) | CLV of the customer for the auto insurance company |

|  |  |
| --- | --- |
| **Object Feature** | **Description** |
| Customer | Unique Customer ID |
| Effective to date | Expiry date of Policy |

|  |  |
| --- | --- |
| **Categorical Feature** | **Category Description** |
| State | 6 different States |
| Response | Whether the customer responded to the marketing. |
| Coverage | Basic, Premium and Extended coverage |
| Education | Maximum qualification |
| Employment Status | Employed/Unemployed/  Medical leave/disabled/retired |
| Gender | Male/Female |
| Location Code | Sub-urban/urban/rural |
| Marital Status | Single/married/divorced |
| Policy Type | Corporate/personal/Special |
| Renew Offer Type | 4 different offers |
| Sales Channel | Web/Agent/Call Center/Branch |
| Vehicle Class | Two -Door/Four-Door/Luxury SUV/Luxury Car/Sports Car/SUV |
| Vehicle Size | Midsize/Small/Large |