Open IIT Data Analytics 2019

TEAM NAME : DataMates

Team Members

Onkar Sabnis	**(18CH30018)**
Rohit	**(18EE10045)**
Harshal Dupare	**(18MA20015)**
Devansh Kar	**(18AR10007)**

Programming language used : Python and R

INDEX:

- **Overview of Our Approach**
- **Understanding Auto-Insurance Industry and CLV**
  - Variables affecting CLV.
- **Applying the approach to Datasets**
  - Considering determining variables.
  - Feature engineering (NPV).
  - Suggestions for removal of rudimentary variables in the dataset.
- **Refining Dataset**
  - Removing correlated variables and variables which don’t affect CLV.
- **Applying Various Models**
  - Linear Regression
  - Random Forest
  - XGBoost
- **Cluster Analysis**
  - Properties of clusters
- **Business Suggestions and Strategies**

OVERVIEW OF APPROACH :

Our approach to solve the problem will follow the following steps.

1. First we analyse the industry and find out the major factors determining it
2. On basis of analysis of industry we point out significant variables determining CLV, and suspect the variables which are not that significant
3. we remove the variables by observing various statistics.
4. then we try various models and choose the one which fits best.
5. then decide the best model.
6. then we apply the clustering to segment the customer into sets.
7. we observe character of cluster and give suggestions on that basis.

1. Understanding Auto-Insurance Industry and CLV

Customer Lifetime value (CLV) is an important outcome-based metric. This metric represents the value of a customer for the period of time during which this customer has a relationship with your company. As a result it is one of the best indicators of the health of an organization.

Basic way to calculate :

CLV = P - C_costs - A_costs

CLV = Customer Lifetime Value

P = Premium over lifetime

C_costs = Total cost of claims over lifetime

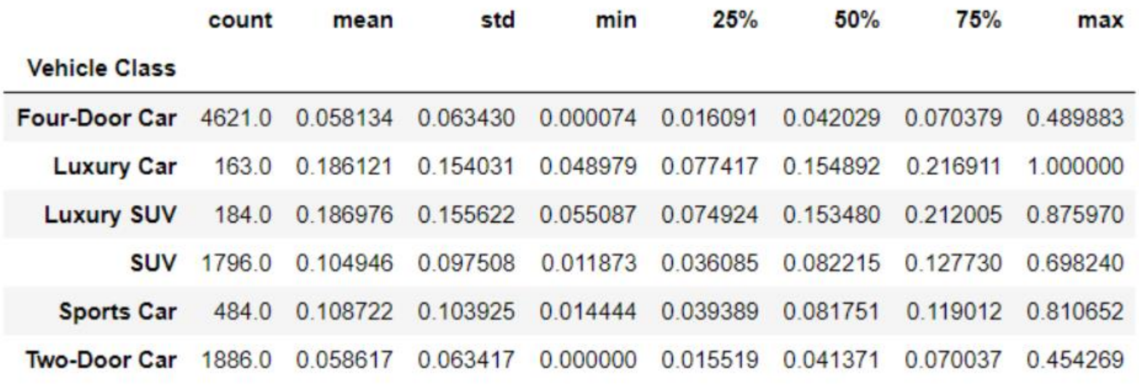
A_cost = Activity based cost

Basic analysis of industry tells us that **premium over lifetime** depends on the policy type, renew offer type, number of policies, coverage, vehicle type other factors which determines the premium and length of policy. **Total cost of claims over lifetime** depends on the type of vehicle which affects the probability of claim being made.

2. Applying the approach to dataset

Based on analysis in part (**1**) we observe significant variables in dataset and plot them against CLV to see the variation. we observe various statistics like mean, variance,standard deviation and other parameters like quartile ranges. This helps us in suggesting which variables to consider for model building.

Few Statistical Insights:

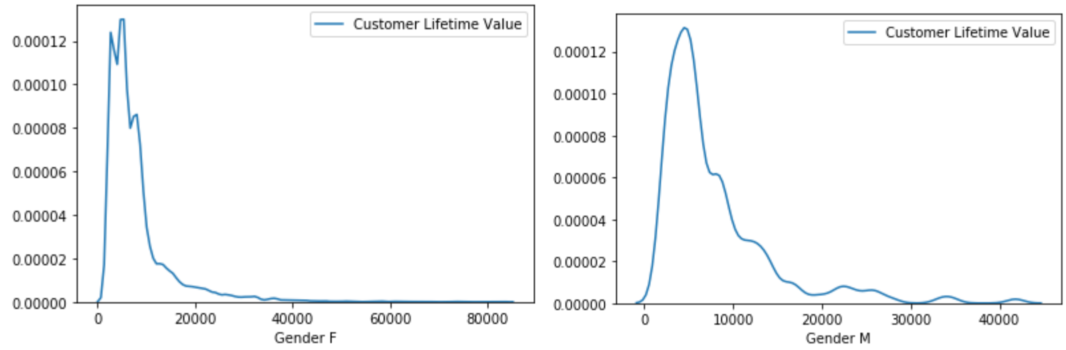


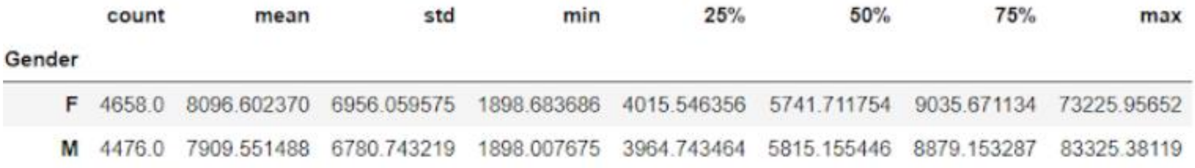
Vehicle Class	count	mean	std	min	25%	50%	75%	max
Four-Door Car	4621.0	0.058134	0.063430	0.000074	0.016091	0.042029	0.070379	0.489883
Luxury Car	163.0	0.186121	0.154031	0.048979	0.077417	0.154892	0.216911	1.000000
Luxury SUV	184.0	0.186976	0.155622	0.055087	0.074924	0.153480	0.212005	0.875970
SUV	1796.0	0.104946	0.097508	0.011873	0.036085	0.082215	0.127730	0.698240
Sports Car	484.0	0.108722	0.103925	0.014444	0.039389	0.081751	0.119012	0.810652
Two-Door Car	1886.0	0.058617	0.063417	0.000000	0.015519	0.041371	0.070037	0.454269

Fig: The statistical analysis of CLV grouped over Vehicle Class shows significant difference in there variance and mean making it an important variable in determination

3. Refining Dataset

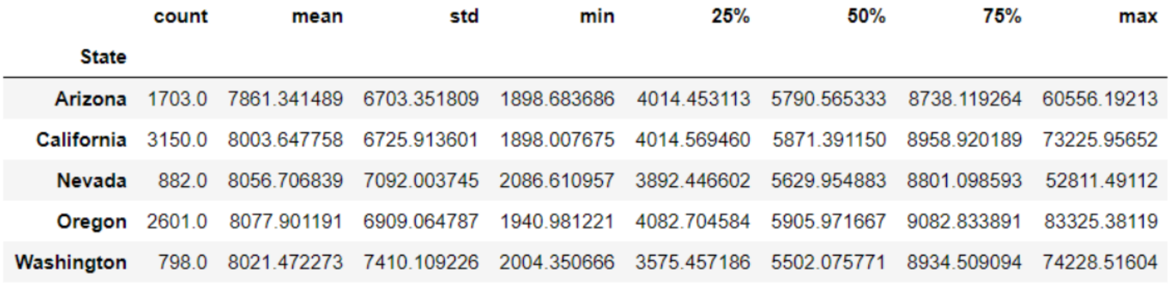
After considering various important variables which are suggested by analysis there are variables left in dataset like gender, sales channel, important accordingly we see various plots and stats like mean, variation, correlation and justify the elimination of variables





	count	mean	std	min	25%	50%	75%	max
Gender								
F	4658.0	8096.602370	6956.059575	1898.683686	4015.546356	5741.711754	9035.671134	73225.95652
M	4476.0	7909.551488	6780.743219	1898.007675	3964.743464	5815.155446	8879.153287	83325.38119

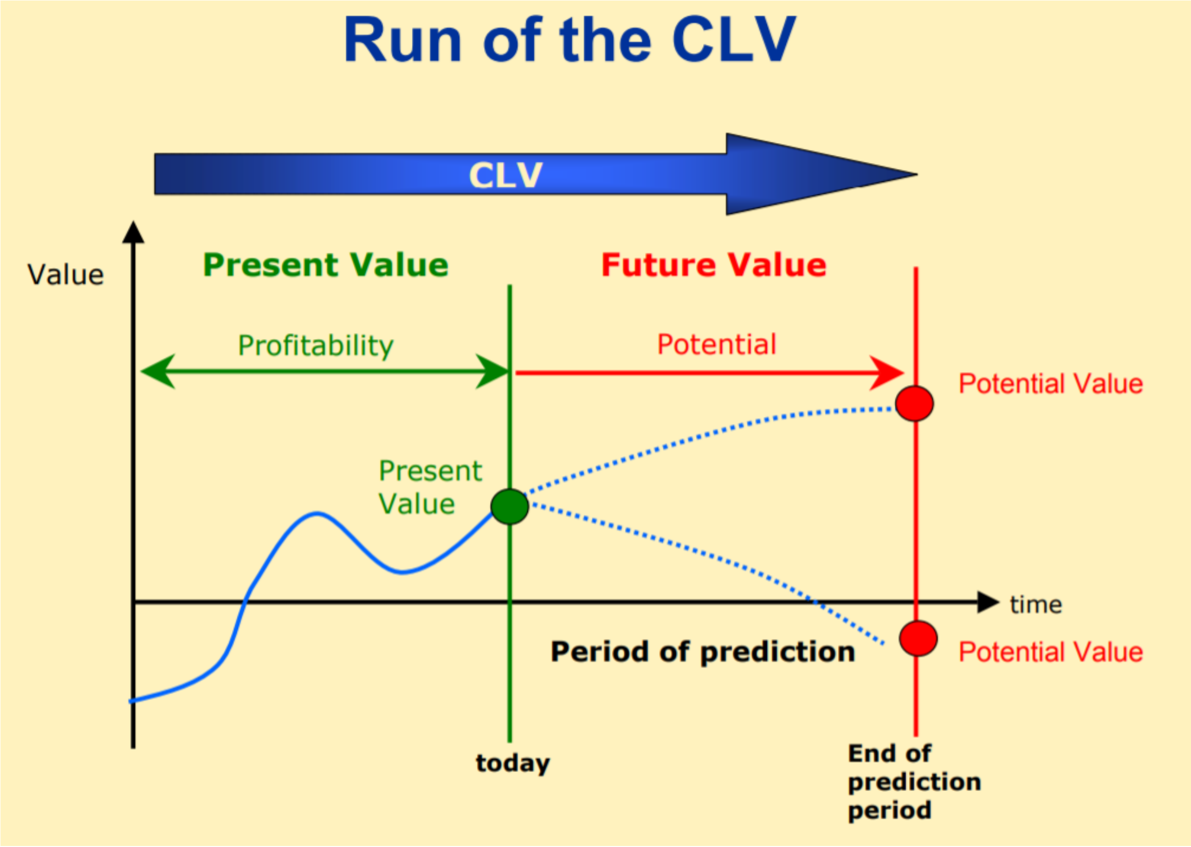
Fig: The statistical analysis of CLV grouped over Gender gives similar inferences, thus Gender can be neglected from our features.



	count	mean	std	min	25%	50%	75%	max
State								
Arizona	1703.0	7861.341489	6703.351809	1898.683686	4014.453113	5790.565333	8738.119264	60556.19213
California	3150.0	8003.647758	6725.913601	1898.007675	4014.569460	5871.391150	8958.920189	73225.95652
Nevada	882.0	8056.706839	7092.003745	2086.610957	3892.446602	5629.954883	8801.098593	52811.49112
Oregon	2601.0	8077.901191	6909.064787	1940.981221	4082.704584	5905.971667	9082.833891	83325.38119
Washington	798.0	8021.472273	7410.109226	2004.350666	3575.457186	5502.075771	8934.509094	74228.51604

Fig:The statistical analysis of the ‘State’ features shows similarity in its distributions and thus won’t be helpful in determining the variations in our target variable.

In industry analysis and CLV we find a good predictor of CLV suggested by research, that is Net Present Value (NPV).We make use of this feature in model and replace it by the variables in its formula.



Feature Engineering:

$$\textbf{Present Value of Customer} = \textbf{Monthly Premium Amount} * \textbf{Months since policy Inception} - \textbf{Total Amount Claim}$$

4. Applying Various Models:

After cleaning ,pre-processing,statistical analysis and feature engineering over dataset we arrived at the model building part.

A]**Linear Model**

We built a linear model with the aid of linear regression and used R2_Score for evaluation.The R2 were computed based on Simple Linear regression model.It was found to be close to 0.18(magnitude) which suggests model isn’t performing well over target variable.

The R2 is less because data is widely spread across the linear regression line.We tried various transformations over target variables such as Log(x),x^2,sqrt(x) and (1/x).The model gave optimum value in transformation (1/x) over others.

B]**XGBoost Model:**

We analyzed the dataset through Random Forests and Decision Trees algorithm but couldn’t get any good results.So we switched to XGBoost model.The XGBoost model was giving stable R2 over train and test dataset.The tuning parameters were tuned in order to get the various R2 value.

5.Cluster Analysis

By Applying both KMeans and Hierarchical Clustering we concluded that Hierarchical Clustering was successful in delivering appropriate clusters.

After applying dendrogram we decided the optimal number of clusters to be three.

Three groups are classified as:

1. Group with LOW CLV ,will rarely take insurance.
2. Group with average CLV can be bought in group 3 by applying ads.
3. Group with high CLV are most profitable group.

6.Suggestions & Recommendations:

From the cluster analysis and Feature Engineering we reached the following conclusions.

- As the Customer who is using Four Door Car and Two Door car more inclined to coverages and even the Loss incurred by them is more.
- After Clustering we can easily segregate the customers into 3 groups and they where clearly different with other groups.
- Group 3 Customers are the most profitable customer in comparison with the customer life time value but the no. of customer count is less.
- Even Group 2 Can be Bought in to Group 3 by giving offers to them and even new attracting policies.

ANNEXURE:

1.https://us.generaliglobalassistance.com/blog/insurance-cost-customer-lifetime-value/

2.http://www.sascommunity.org/seugi/SEUGI2003/SEYERLE_LifetimeValue.pdf2.https://www.insurancejournal.com/maga**zines/mag-mindyourbiz/2003/07/21/31007.htm**