Open IIT Data Analytics

2019-20

Amateur Analysts

INDEX

1. **Introduction**
    1.1 Auto Insurance
    1.2 Customer Lifetime Value
    1.3 Objectives

2. **Data Analysis**
    2.1 Data Description
    2.2 Feature Engineering
    2.3 Business Insights

3. **Model Training**
4. **Results**
5. **Conclusion**
6. **Annexures**

Introduction

Auto Insurance

The insurance industry is an integral part of the global economy. Insurance is a mechanism of risk-transfer from a customer to an insurance company, to protect personal finances in the unfortunate event of accidental loss or damage. It offers full or partial financial compensation in such a case, provided a fee called premium is paid by the customer.

When the entity to be insured is a vehicle, the insurance is called **auto insurance**. In order to mitigate expenses in case an auto accident occurs, people pay premiums to an auto insurance company and the company will pay to cover some, or all of the damages incurred during the accident.

Auto insurance premiums vary depending on age, gender, years of driving experience and accident history among other factors. A poor driving record or the desire for complete coverage will lead to higher premiums, as the likelihood and possible expenses of covering an accident increases for the company issuing auto insurance.

Customer Lifetime Value- A profit prediction

Customer Lifetime Value (CLV) is the total worth to the auto insurance company of a customer over the whole period of their relationship. It’s an important metric in estimating a reasonable cost of acquisition, as it often costs less to keep existing customers than it does to acquire new ones. When applied to the customers of a company, the Pareto Principle (also known as the 80-20 rule), the rule of thumb is that 80% of your sales come from 20% of the customers. CLV is important in zoning in on those high-value customers.

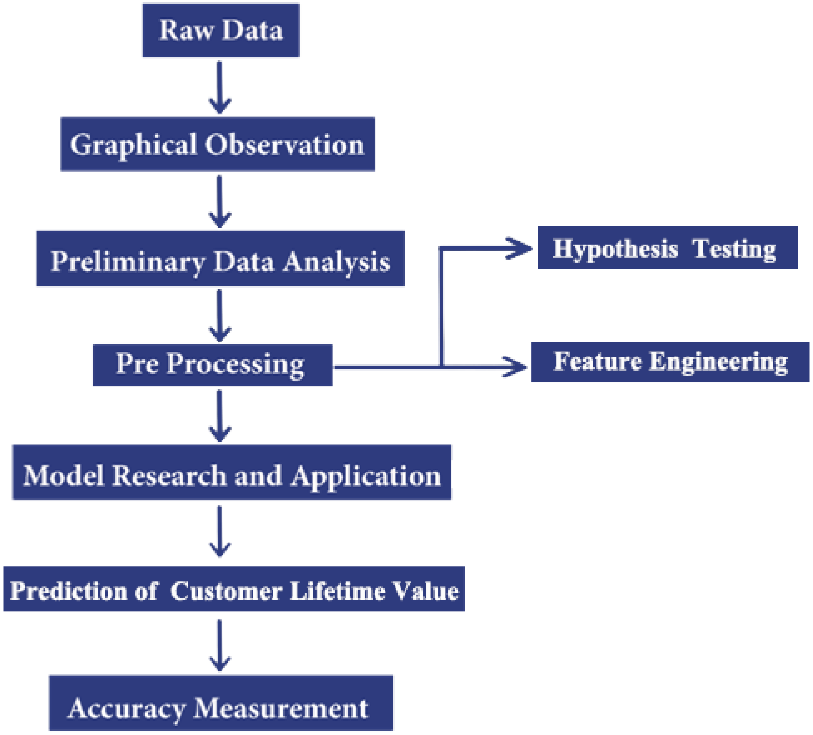


Fig1:- Workflow Chart

Objectives

General Objectives (Problem Statement):

- To predict Customer Life-time Value for an Auto Insurance company
- Find out the types of customers that would generally give the company more revenue.

Specific Objectives:

- To aggregate features to create relevant features and to find out which explanatory variables have a significant effect on CLV calculation.
- To test different models and choose the best model based on accuracy

Data Analysis

Description of the data

There are 24 variables explaining various characteristics of the customer and the customer-client relationship. The dataset has potential to gain insight into any of the 9134 customers enlisted. Descriptions of the following variables are:-

- **Customer Lifetime Value** is a quantitative variable that which describes the financial value of a particular customer's lifetime relationship with the insurance company. It is an indication of the predicted difference between all the projected premiums the customer will pay and all incurred costs by the company when dealing with a particular customer.

- **Customer** is a qualitative variable which is the Unique Customer ID of a particular customer holding an auto insurance policy under the company.

- **State** is a categorical variable which shows which state the customer lives in, namely Arizona, California, Nevada, Washington and Oregon.

- **Coverage** is a qualitative variable which shows the level of coverage of the insurance scheme chosen by a customer. The three levels of coverage are basic, extended and premium.

- **Education** is a qualitative variable describing the highest level of education of the customer, varying between four levels namely high school and below, college, bachelor's degree and master's degree.

- **Effective to Date** indicates the date till which a policy taken by the customer is valid.

- **Employment status** is a categorical variable which indicates whether the customer is currently employed, unemployed, on medical leave or disabled.

- **Gender** is a qualitative variable which shows if the customer identifies as male or female.

- **Income** is a quantitative variable that captures the monthly earnings of a customer.

- **Location code** is a qualitative variable that shows the kind of areas they have settled in, namely urban, suburban and rural.

- **Marital status** is a qualitative variable indicating whether the customer is married, single or divorced.

- **Monthly Premium Auto** is a quantitative variable which is the value of the premium paid by the customer each month towards the auto insurance policy.

- **Months since last claim** are a quantitative variable that is the number of months that have passed since the customer has last claimed the auto insurance.

- **Months since policy inception** is a quantitative variable that is the number of months elapsed since the customer has taken a particular insurance policy, to the data of collection of data.

- **Number of open complaints** is a quantitative variable that is the number of complaints the customer has registered with the company that have not yet been resolved.

- **Number of policies** is a quantitative variable which shows the number of auto insurance policies that a particular customer has taken.

- **Policy Type** is a categorical variable which shows the type of auto insurance policy taken by the customer. The different policy types are corporate auto, Personal auto and Special auto.

- **Policy** is a categorical variable which shows the specific auto insurance policy taken by the customer. With three policy types and three subdivisions each (L1, L2 and L3) there are a total of nine auto insurance policies.

- **Renew Offer Type** is a categorical variable which shows which offer type has been chosen to renew the customer's auto insurance once the validity expires. The different renewal offers are Offer 1, Offer 2, Offer 3 and Offer 4.

- **Sales Channel** is a qualitative variable which shows the medium through which the company contacts customers for policy related matters. It can be through any of the four mediums, namely agent, call center, web and branch.

- **Total claim amount** is a quantitative variable which is a measure of the total amount of money a customer can claim in case of an accident.

- **Vehicle class** is a qualitative variable describing the different classes of vehicles covered under auto insurance policies. There are six classes namely, two-door cars, four-door cars, SUV, luxury SUV, luxury car and sports car.

- **Vehicle size** is a categorical variable with three different levels of vehicle size, namely small, midsize (medium sized), and large.

Feature Engineering

Feature Creation

"$\text{CLV}_{\text{complex}}$ = PV(Monthly Premium Auto ,Maximum(Months Since Policy Inception, Months since Last Claim),r)* Number of policies - PV(Total Claim Amount, Months Since Last Claim-Maximum(Months Since Policy Inception, Months since Last Claim) ,r"

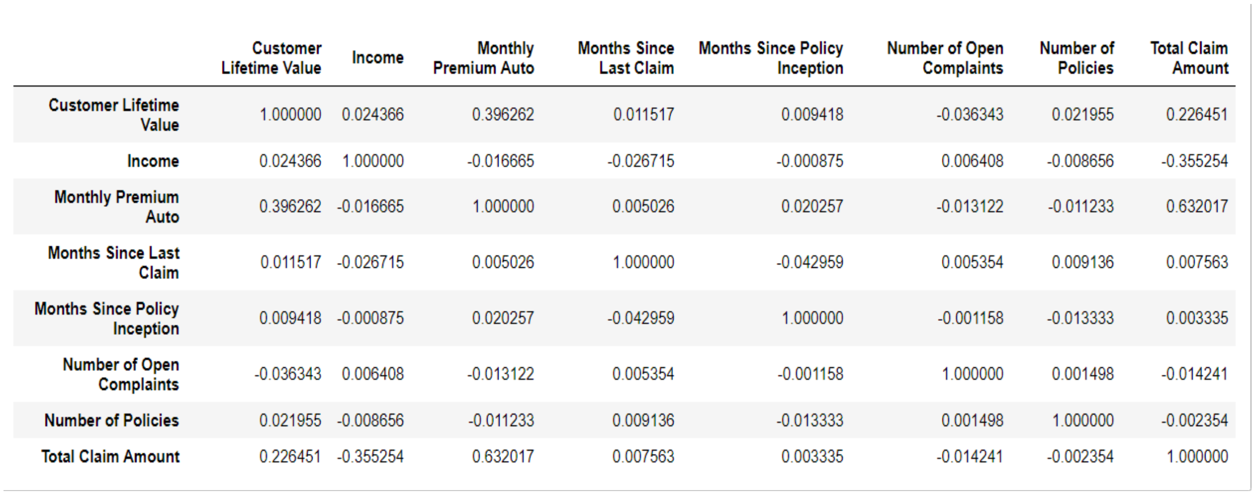
$\text{CLV}_{\text{simple}}$ = Monthly Premium Auto*Maximum(Months Since Policy Inception, Months since last claim)* Number of Policies - Total Claim Amount

The features showed a correlation of **0.25** and **0.241** with Customer Life Time Value and a correlation of 0.92 with each other .Hence, only $\text{CLV}_{\text{complex}}$ is used to avoid multicollinearity issues.

Features Selection

1. For continuous variables

   Correlation matrix and heat map are used to find the significant features



	Customer Lifetime Value	Income	Monthly Premium Auto	Months Since Last Claim	Months Since Policy Inception	Number of Open Complaints	Number of Policies	Total Claim Amount
Customer Lifetime Value	1.000000	0.024366	0.396262	0.011517	0.009418	-0.036343	0.021955	0.226451
Income	0.024366	1.000000	-0.016665	-0.026715	-0.000875	0.006408	-0.008656	-0.355254
Monthly Premium Auto	0.396262	-0.016665	1.000000	0.005026	0.020257	-0.013122	-0.011233	0.632017
Months Since Last Claim	0.011517	-0.026715	0.005026	1.000000	-0.042959	0.005354	0.009136	0.007563
Months Since Policy Inception	0.009418	-0.000875	0.020257	-0.042959	1.000000	-0.001158	-0.013333	0.003335
Number of Open Complaints	-0.036343	0.006408	-0.013122	0.005354	-0.001158	1.000000	0.001498	-0.014241
Number of Policies	0.021955	-0.008656	-0.011233	0.009136	-0.013333	0.001498	1.000000	-0.002354
Total Claim Amount	0.226451	-0.355254	0.632017	0.007563	0.003335	-0.014241	-0.002354	1.000000

Fig 2: Correlation Matrix

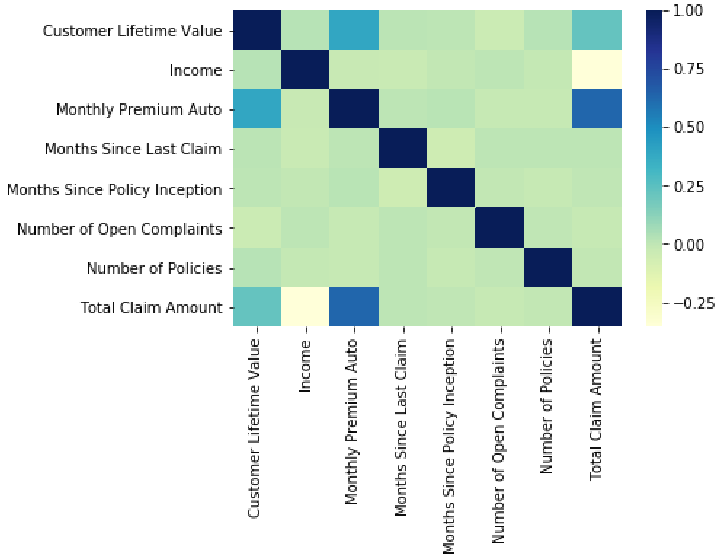


Fig 3: Heat Map

2. For categorical variables

The significant features are selected by first using Data Visualization Techniques like box plot, histogram and scatter-plot. After visualizing the graphical representation, we used hypothesis testing to get concrete proof for the same. The hypothesis tests done are namely:-

A. Since there appeared to be no significant difference between the CLV of the two genders, the impact of gender on CLV was considered minimal (refer: Annexure Fig 1.11). To confirm, T-test was performed on the feature Gender, the results of which are shown below.

$$Null\ Hypothesis\ (H0): 1 - 2 = 0$$
$$Alternate\ Hypothesis\ (Ha): 1 - 2\ 0$$

(where 1 and 2 are the means of clients having gender male and female respectively)

Result: - **p-value=0.1933**
Since, the p-value > alpha (significance value = 0.05)

Hence, the Null Hypothesis (both of the sample come from the same population) is accepted and feature Gender was dropped.

B. We used ANOVA test over CLV in case of multicategorical variables. The results of ANOVA after segregation of CLV on the basis of a particular feature are as follows:-

Variable for which ANOVA-test was applied	p- Value
Education	0.0460
Coverage	6.01* 10^(-58)
Renew Offer Type	1.23* 10^(-16)
Location Code	0.820
State	0.895
Employment	0.0042
Vehicle new	1.21* 10^(-266)
Policy Type	0.1126
Policy	0.304
Sales Channel	0.4502
Number of Open Complaints	0.0005

3. Both categorical and continuous variables

We used Extra trees classifier (for ranking of significance of categorical and non-categorical features combined).(refer: Annexure Fig 1.31)

Features Merging

- The features Vehicle Class and Vehicle Size were clubbed together to form 18 categories
- The features Policy Type appears to be a subclass of feature Policy and since we could see significant difference between two Policies having the same Policy Type, we decided to drop Policy Type
- The Location Code and State are merged together to create 15 categories

Features Analysis

- The features like CLV appear to be very skewed. Hence, we use different transformations like logarithmic, square root, cube root and selected the cube root as it gives back the most standardized distribution of the CLV values

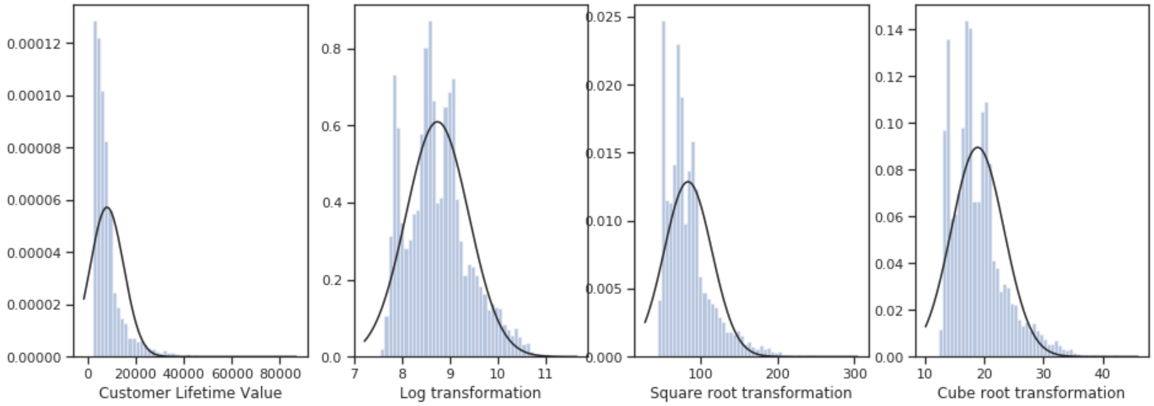
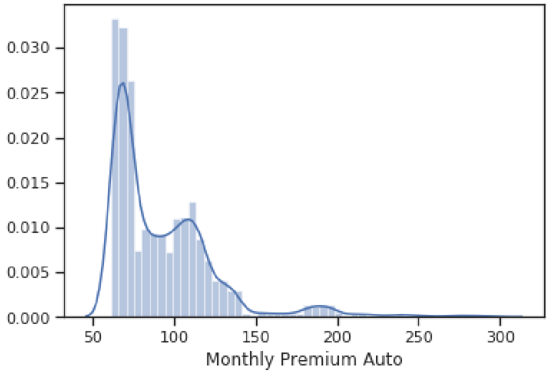


Fig 4

- The feature Monthly Premium Auto also appeared to be skewed; hence we used Box Cox transformation for the same to standardize the data.



- However, the skewness of Income isn't apparent. Hence, Welch Two Sample t-test was performed on Income. And p-value came out to be 0.762, So, null hypothesis was accepted (because p-value>0.05) and it was concluded that income was not skewed.

- The columns Months Since Last Claim and Months Since Policy Inception clearly showed no kind of skewness hence, no transformation was used for the same.

So, finally the features included and their significance from a business perspective are as follows(The plots for the same have been added in the annexure) :-

(i.) Income (ii.) Monthly Premium Auto (iii.) Marital Status (Iv.) Coverage (v.) Education (vi.)Employment Status (vii.) Months since last claim (viii.) Months since policy inception (ix.)Number of complains (x.)Number of policies (xi.) Policy (xii.) Renew Offer Type (xiii.) Sales Channel (xiv.) Total Claim Amount (xv.)CLV_complex (xvi.) Vehicles (xvii.) Location State

Business Insights

From our analysis on the data we find few of the features very important from the business perspective as these are important in calculating Customer Lifetime Value and hence determining the profits of the Auto Insurance Company. The following features were the most important:

Monthly Premium Auto- We find that there is a high correlation between Monthly Premium paid by customers to their lifetime value. The more premium a person pays, the more the company profits. Hence, the company should ask its customers to make policies which have high premiums. It provides more coverage as well as profits of the company.

Employment Status - It can be noted from the plots that Employed people tend to have higher lifetime value. Employed people have a fixed source of income and hence can afford cars and therefore have an auto insurance. Thus the insurance companies' main target should be people who are employed and having a high source of income.

Coverage - We find a relation between the coverage of the insurance and the lifetime value of the customer who has the insurance. "Premium" coverage has the highest average Customer Lifetime Value whereas "Basic" coverage has the least. The better the coverage the more risks it undertakes and also the more the company profits in the long run. Hence, the company should make their customers realize the need for a better coverage.

Vehicles - The more luxurious the car is the more concerned the owner are and hence they tend to take a better coverage. Thus the insurance companies profit more from such insurances and hence the customers tend to have a higher Customer Lifetime value.

We used both label encoding and one-hot encoding for categorical variables and it was seen that one-hot encoding had better accuracy over both training and validation set.

Model Training

Eight different models were tested, namely:

- Elastic Net Regression
- Lasso Regression
- Ridge Regression
- Random Forest Regression
- Support Vector Regression
- Gradient Boosting Regression
- XGBoost Regressor
- LightGBM Regressor

After training these models, the models that performed decently were:

- Random Forest Regression
- XGBoost Regression

The pros and cons of the models are as follows:-

Random Forest Regression

Pros:-
a. It is robust to outliers (which are many in number in our case)
b. It is also indifferent to non-linear data
c. Each Decision Tree has high variance, but low bias. But because we average over all trees, we have a low bias and moderate variance model

Cons:-

a) It can tend to overfit and hence needs parameter tuning

XGBoost Regression

Pros :-
a. It has inbuilt lasso and ridge regression and hence prevents overfitting
b. It has inbuilt capability to handle missing values

Cons:-
a. Model interpretability:- It appears to be a blackbox and is very difficult to be interpreted

We also tried using PCA by selecting the number of clusters on the basis of their explained variance. But that resulted in poorer accuracy in both training and validation set.

Note: - We used both label encoding and one-hot encoding for categorical variables and it was seen that one-hot encoding had better accuracy over both training and validation set.

Results

Model Name	R2-Score		Adjusted R2-Score		Mean Absolute Percentage Error (MAPE)	
	(Training Set)	(Validation Set)	(Training Set)	(Validation Set)	(Training Set)	(Validation Set)
Random Forest Regression	0.94	0.78	0.94	0.66	3.35	8.59
XGBoost Regressor	0.82	0.75	0.82	0.62	7.37	9.40

Conclusions

After studying the various features the following conclusions were derived:

- From the Business point of view, people under the Extended Coverage or availing Offer Type1 have higher value of CLV followed by Premium, therefore are more profitable for the company.

- According to the data, the frequency of people having no complaints was quite high and were having highest CLV, thus the number of complaints shows an inverse relation with CLV.

- Customers having Small Luxury Car are the most profitable and have the highest value of CLV.

- After Clustering we can easily segregate the customers into 3 groups and they were clearly different from other groups

- The three groups were named as best, medium and worst in terms of the value added to the company.(Plots added in Annexure)(Fig -1.33, 1.34,1.35,1.36)

- The huge number of no complaints show that the clients are satisfied with the company and hence gives an indication towards lesser number of claims the company has to provide

- Concluding that Random Forest Regressor is stable in order to predict the CLTV.

ANNEXURE

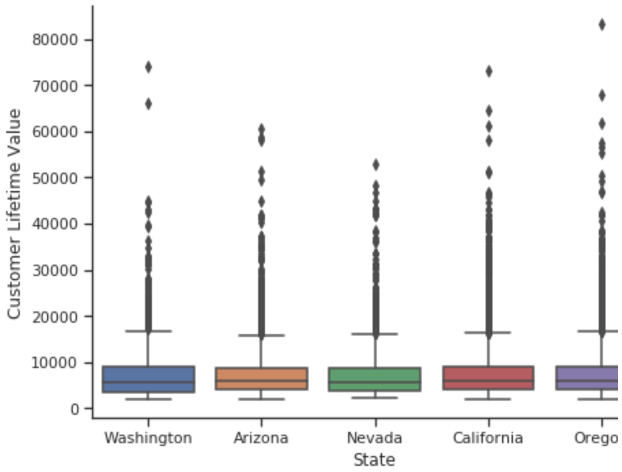


Fig 1.1

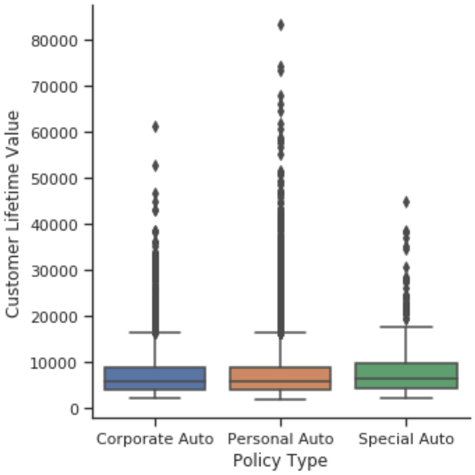


Fig 1.2

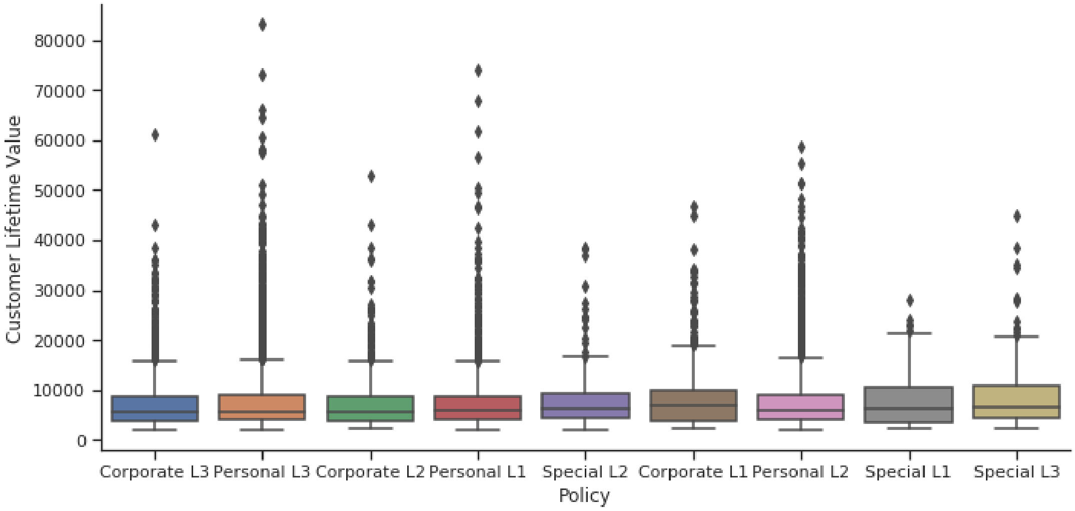


Fig 1.3

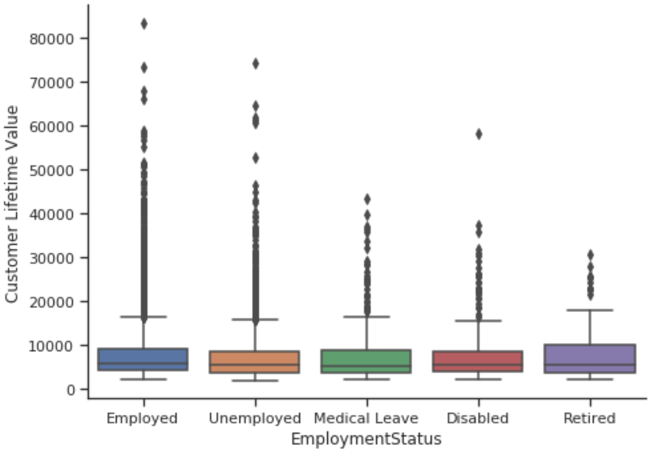


Fig 1.4

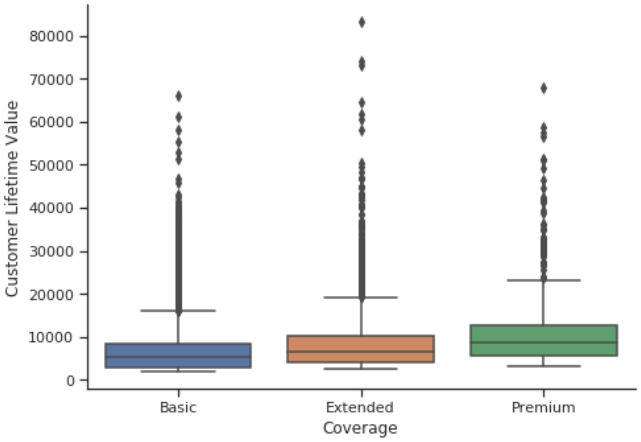


Fig 1.5

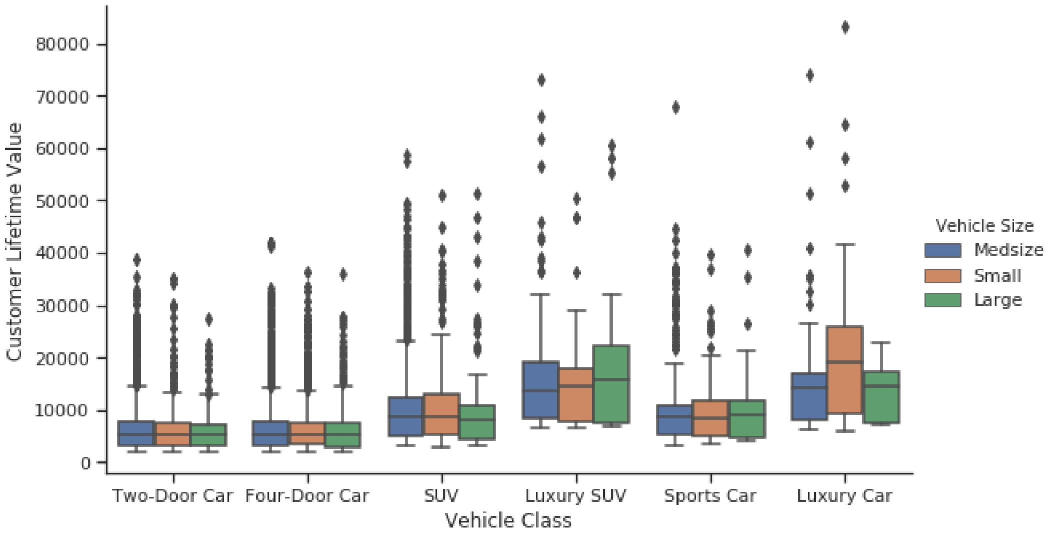


Fig 1.6

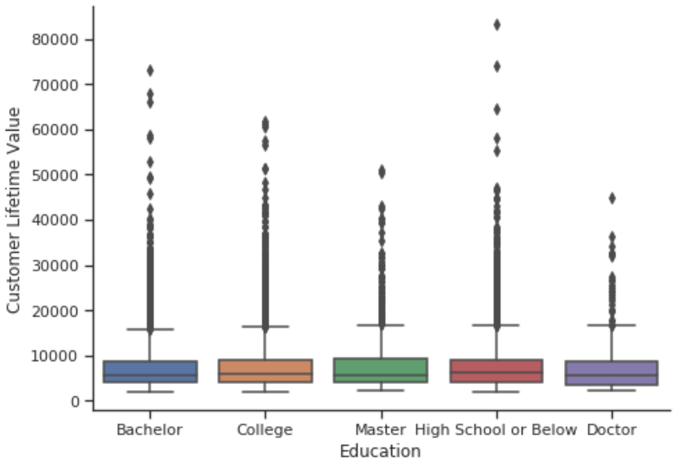


Fig 1.7

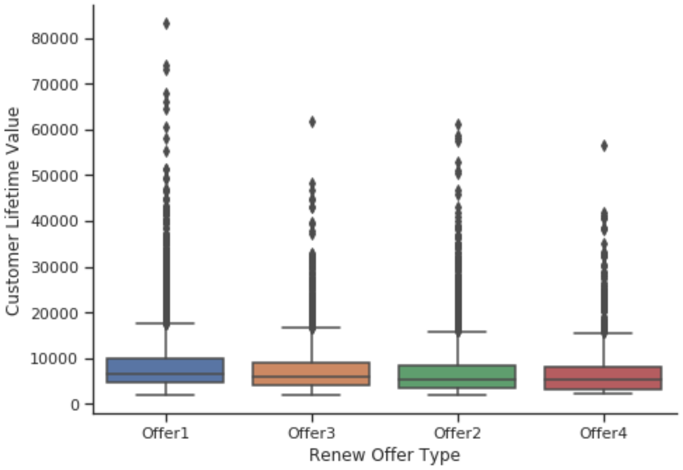


Fig 1.8

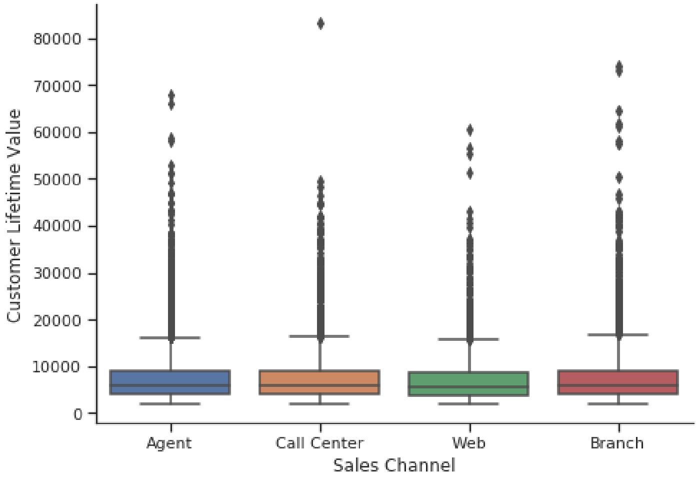


Fig 1.9

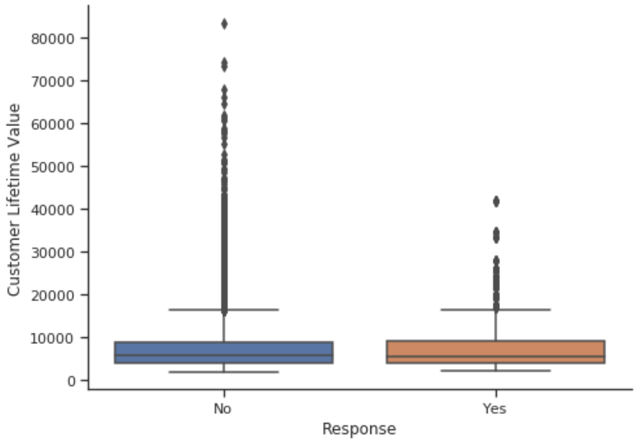


Fig 1.10

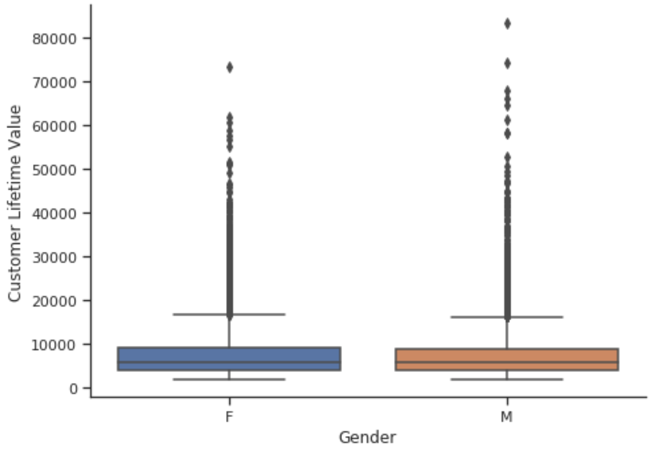


Fig 1.11

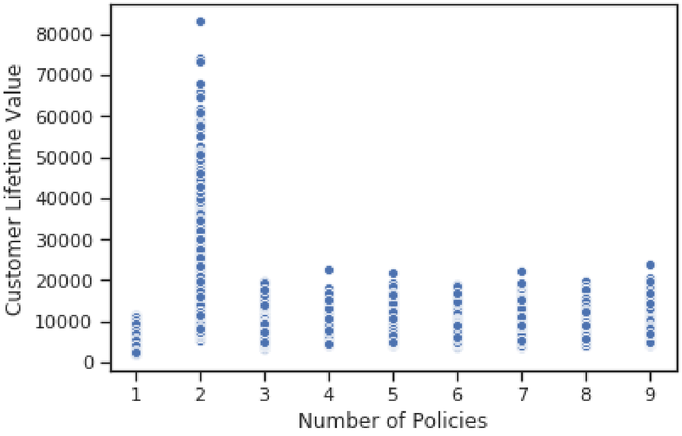


Fig 1.12

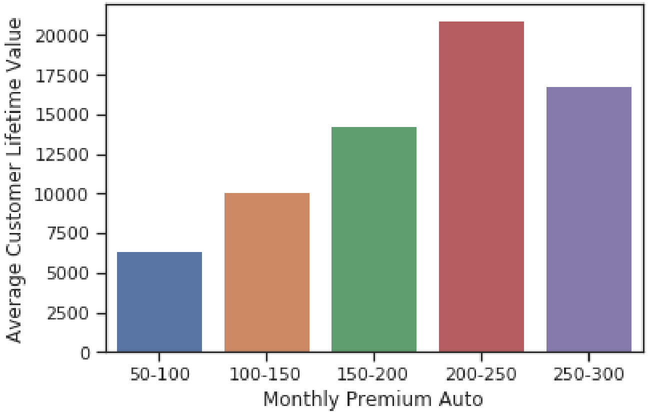


Fig 1.13

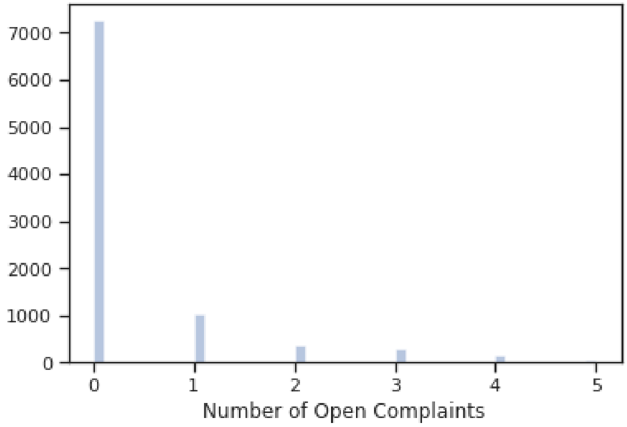


Fig 1.14

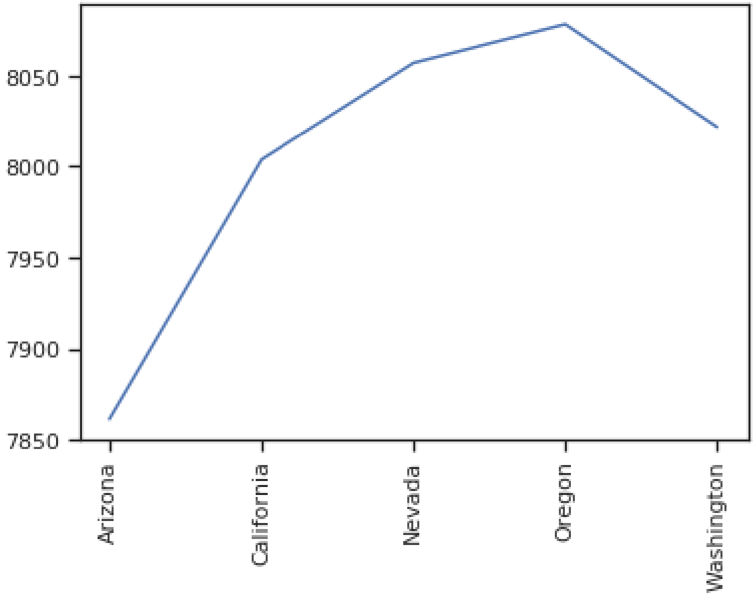


Fig 1.15

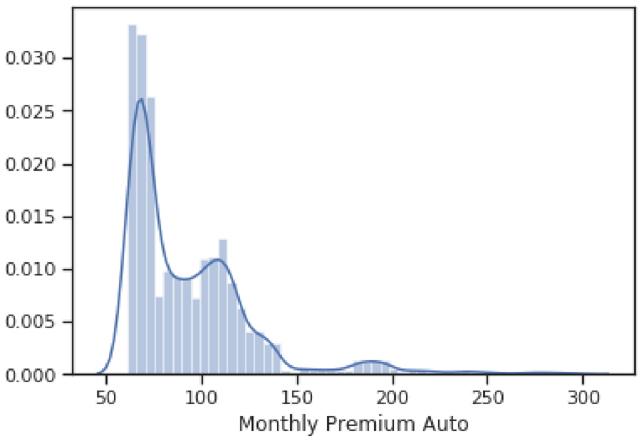


Fig 1.16

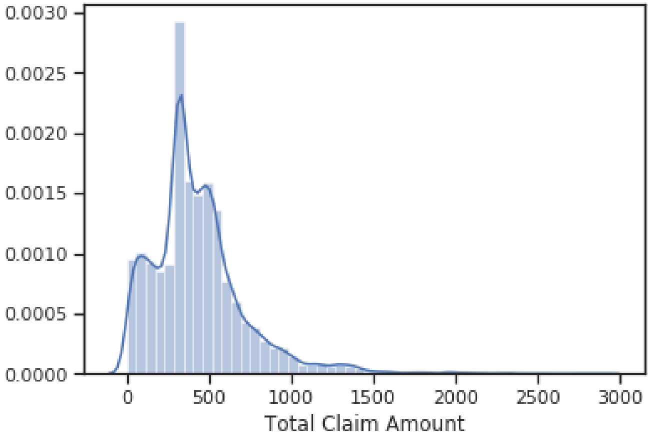


Fig 1.17

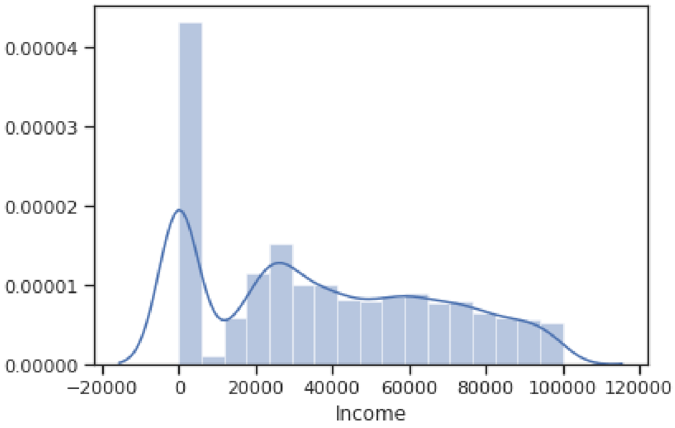


Fig 1.18

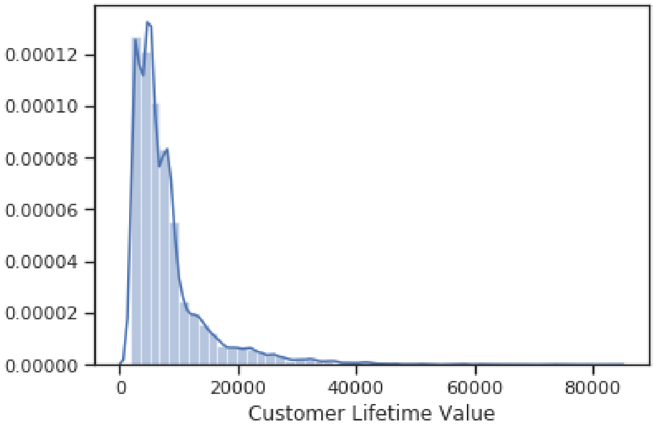


Fig 1.19

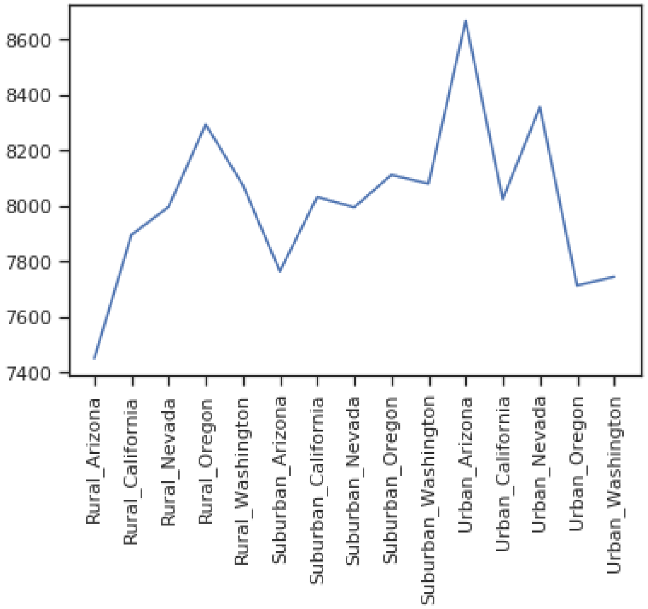


Fig 1.20

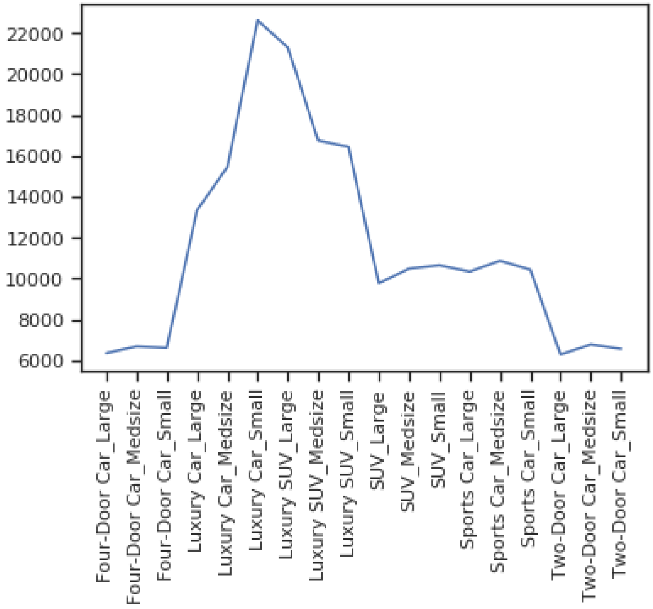


Fig 1.21

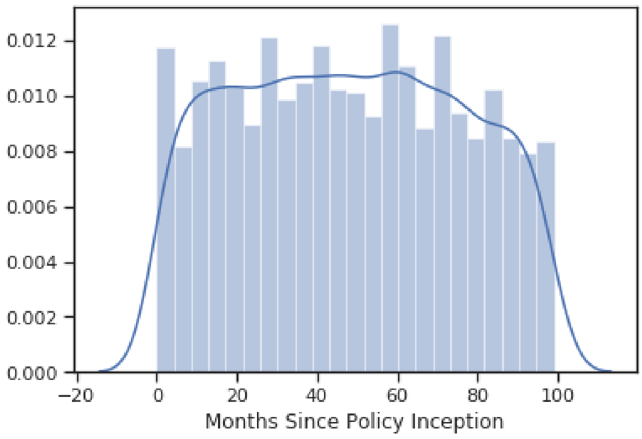


Fig 1.22

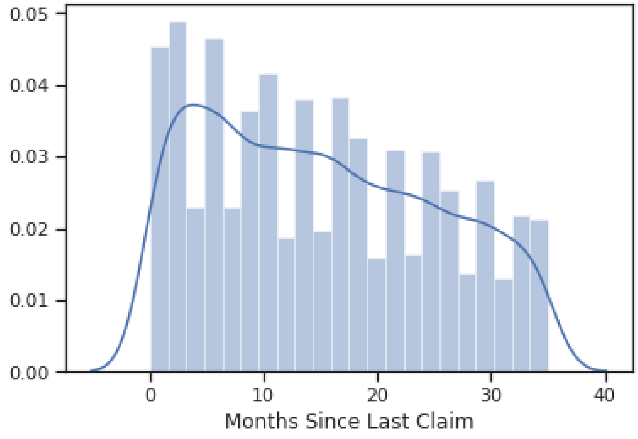


Fig 1.23

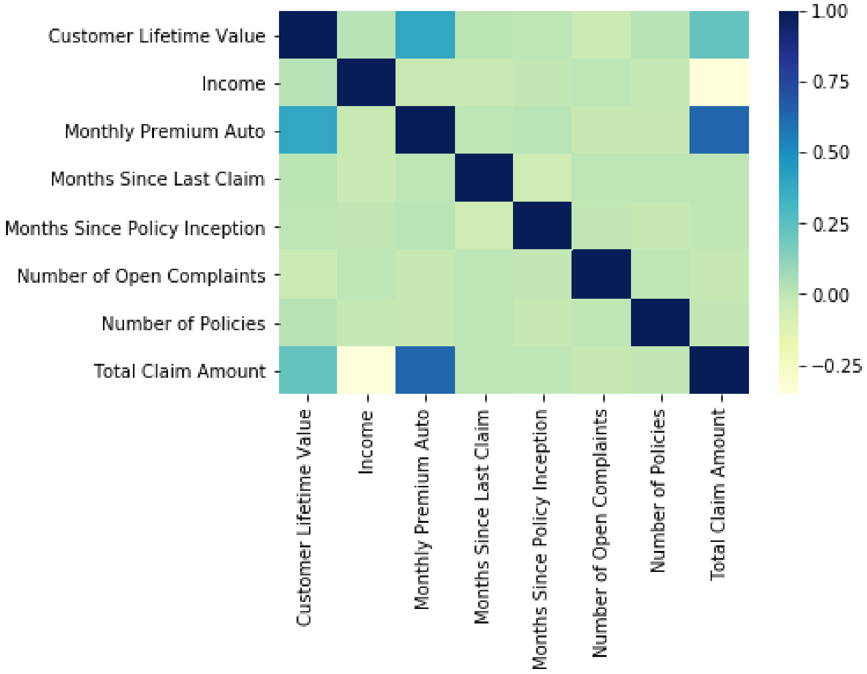
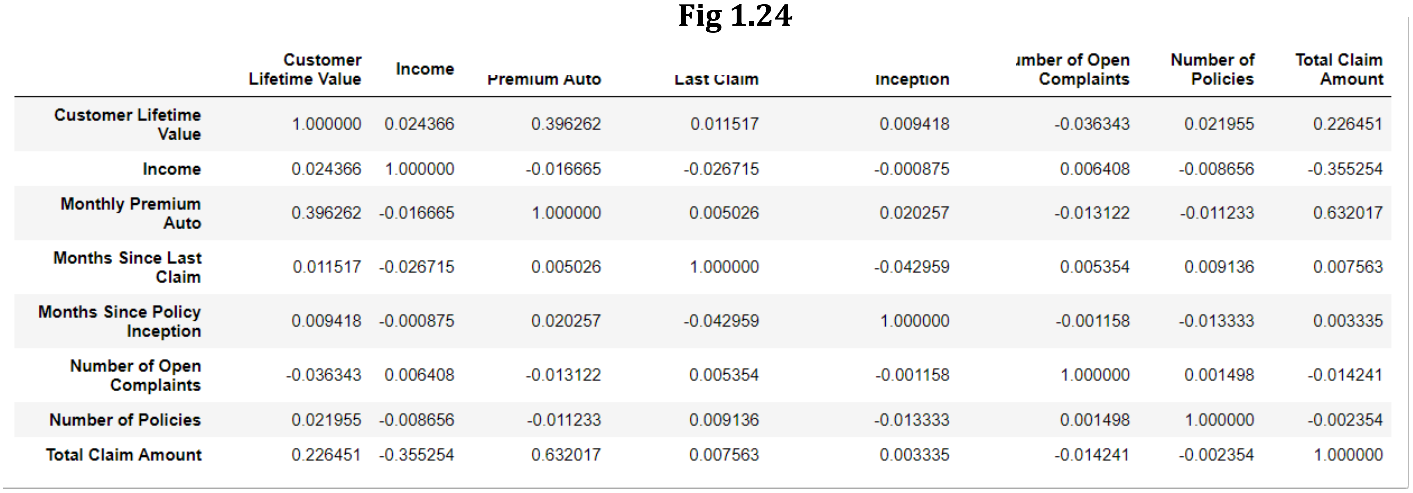


Fig 1.24



	Customer Lifetime Value	Income	Premium Auto	Last Claim	Inception	umber of Open Complaints	Number of Policies	Total Claim Amount
Customer Lifetime Value	1.000000	0.024366	0.396262	0.011517	0.009418	-0.036343	0.021955	0.226451
Income	0.024366	1.000000	-0.016665	-0.026715	-0.000875	0.006408	-0.008656	-0.355254
Monthly Premium Auto	0.396262	-0.016665	1.000000	0.005026	0.020257	-0.013122	-0.011233	0.632017
Months Since Last Claim	0.011517	-0.026715	0.005026	1.000000	-0.042959	0.005354	0.009136	0.007563
Months Since Policy Inception	0.009418	-0.000875	0.020257	-0.042959	1.000000	-0.001158	-0.013333	0.003335
Number of Open Complaints	-0.036343	0.006408	-0.013122	0.005354	-0.001158	1.000000	0.001498	-0.014241
Number of Policies	0.021955	-0.008656	-0.011233	0.009136	-0.013333	0.001498	1.000000	-0.002354
Total Claim Amount	0.226451	-0.355254	0.632017	0.007563	0.003335	-0.014241	-0.002354	1.000000

Fig 1.25

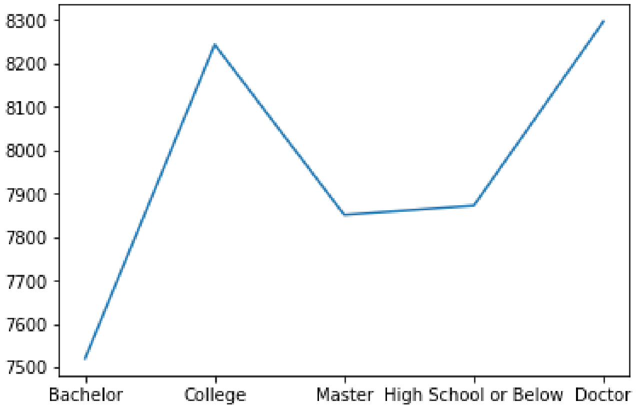


Fig 1.26

Mean CLV vs Education

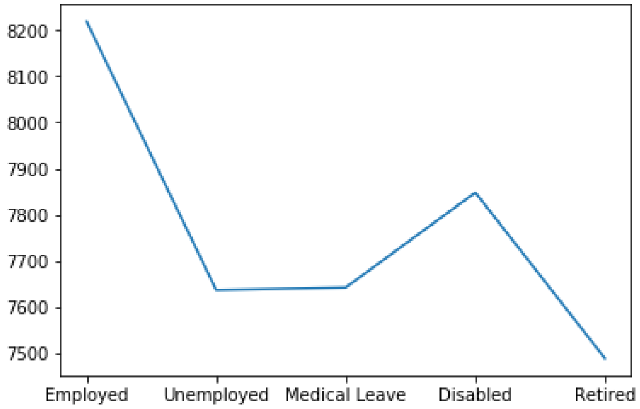


Fig 1.27

Means CLV vs Employment Status

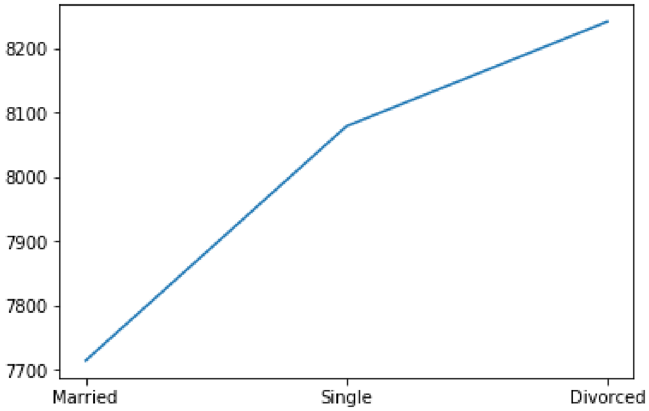


Fig 1.28

Mean CLV vs Marital Status

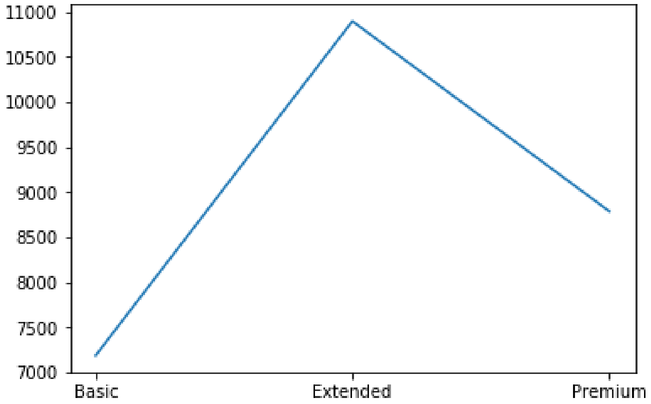


Fig 1.29

Mean CLV vs Coverage

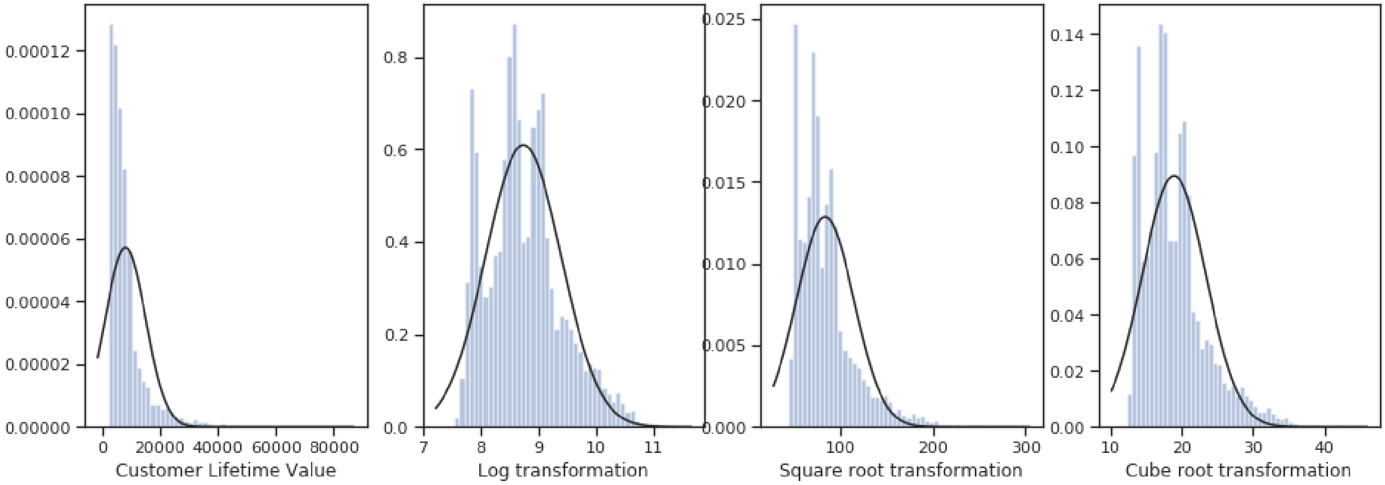


Fig 1.30

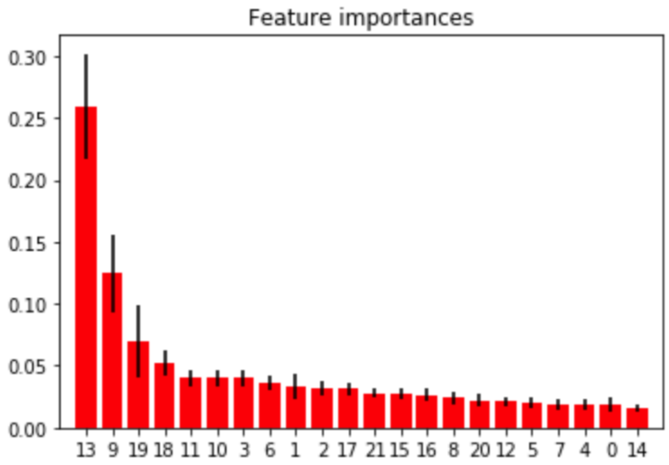


Fig 1.31

Extra Tree Classifier

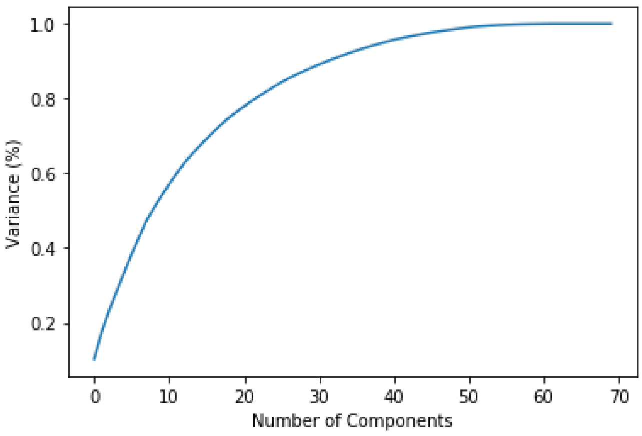


Fig 1.32

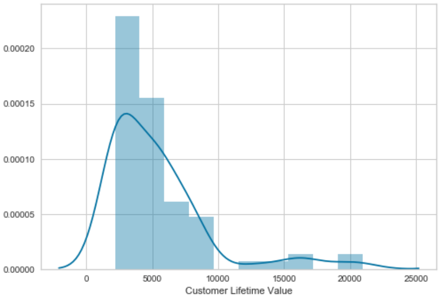


Fig 1.33

Cluster Worst

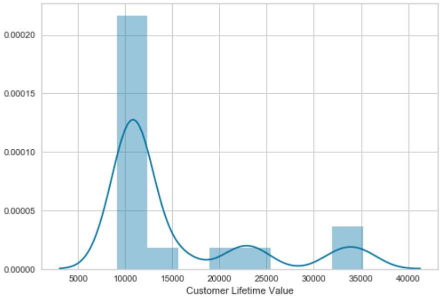


Fig 1.34

Cluster Medium

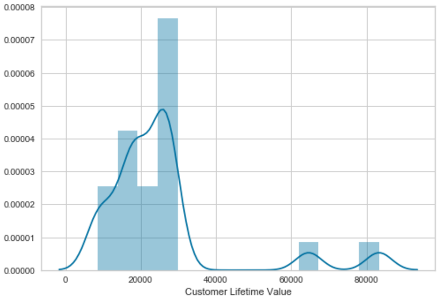


Fig 1.35

Cluster Best

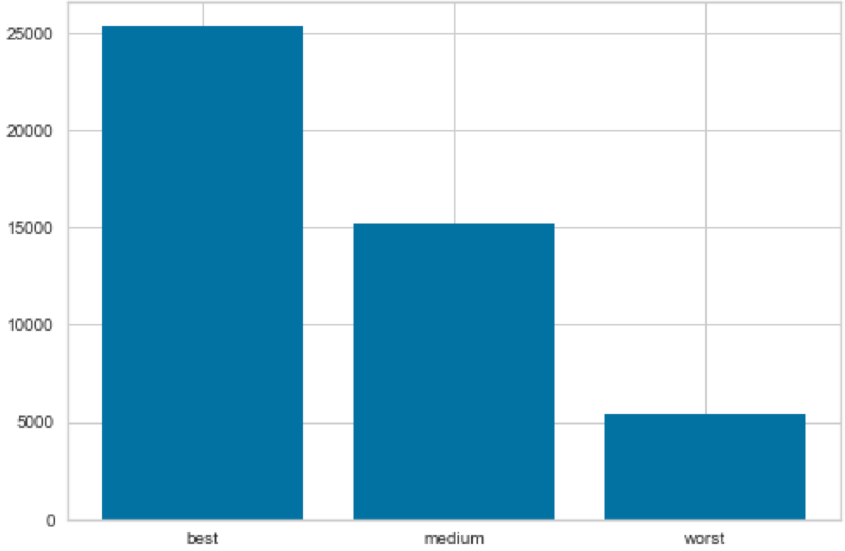


Fig 1.36

Mean CLV of clusters