P.Anand Sagar
11197418

Predicting Customer Lifetime Value (CLV)

Introduction/Background:

We know that every business need customers, as they are the source of revenue. Success of business is directly proportional to its ability to acquire customers, solve their issues , and make more money. But for this to happen, business has to identify the right potential customers. In today's world, customers have many options to choose. The business has to understand and plan for what the customers might do in the future. Simply, putting in words, they need to predict customer behavior. This project aims at this point. It uses customer data to analyze and understand how customers behave and react to marketing strategies, observe customer segmentation w.r.t various attributes, and finally predict customer Lifetime value. CLV is a monetary value that represents the amount of revenue a customer will spend on the business over the lifetime of their relationship. Tagging each customer with CLV helps a business focus on those customers who can bring most revenue in future.

Problem Statement:

Now-a-days competition is getting faster in the market, and customers have many options to choose, if they want to buy a product. Also, with this pandemic, many businesses are facing a tough time to retain their old customers and acquire new ones. So, I thought this area would be the most interesting topic for me to carry out my research investigation. I have selected Auto insurance customer data for my research and with my analysis, I would like to contribute to major questions for business:

1) Predict Customer Lifetime Value using regression models.
2) Who are the potential customers contributing for their business?
3) What are they expecting and the factors for it?
4) How are the customers making their decisions?
5) How does a business plan for their future?

Literature review:

Research in the area of CLV is being carried out which offers a useful framework where marketing activities are related to financial metrics. It represents how a change in CLV effects profitability to the firm. Firms are showing increased interest in the areas of customer management process and for this concept of CLV will be mandatory. Significant research has to be made in this field to explain accurately about the effects of CLV.

P.Anand Sagar
11197418

Objectives:

1) Customer Analytics (For e.g. With the given data, how are customers responding to the offers/ policies given by the company and visualizations to observe current market needs from customer perspective)
2) As CLV is the dependent variable, observed what are the factors that contribute to higher CLV. For this to achieve, made use of hypothesis testing, correlation analysis and various visualization plots and carried out feature selection.

Hypothesis:

Null hypothesis: Response has significant effect on CLV.

Alternate hypothesis: Response has no effect on CLV, it is only by random chance.

3) Applied feature engineering methods and built two models; Linear regression model and Decision Regressor model to predict CLV, that will help the business to identify and nurture these top customers for a steady revenue. Compared the accuracies of both the models.

Data Description:

This is the customer data from an auto insurance company, which is publicly available data set prepared by IBM, and acquired from Kaggle.

It includes 24 features, and nearly 10000 records.

The features include customer id, CLV, Response(whether he is responding to offers or not), coverage of insurance policy, months since policy inception, customer attributes (Income, location, marital status), monthly premiums, policy type, Renew offers, Sales channel, vehicle type and vehicle size.

Research design and Methodology:

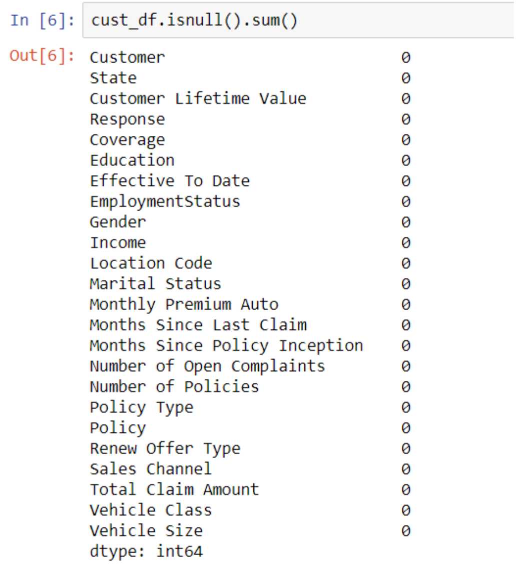
After acquiring this data, I planned to breakdown this research methodology into three steps. My first approach is data preparation such as handling the null values, and unusual data. Next step is to conduct Exploratory data Analysis and to find out key features that effect CLV. I have used python visualizations, and tableau visualizations in this step to find out the most important features. Also, to use SAS enterprise miner to carry out correlation analysis. The final step is to build a linear regression and Decision regressor models to predict Customer Lifetime Value.

Data Preparation:

a) With the initial data being in CSV format, column “Vehicle size” has got some unnatural value #name. To replace this, I have observed the frequency of the values, out of which vehicle size medium has appeared around 7000 times out of 9500 rows. So, replaced those with Medium.

b) Also, in this phase, made note of those columns like “ effective date”, and “Location” etc. which has nothing to do with our target variable and to remove them later.

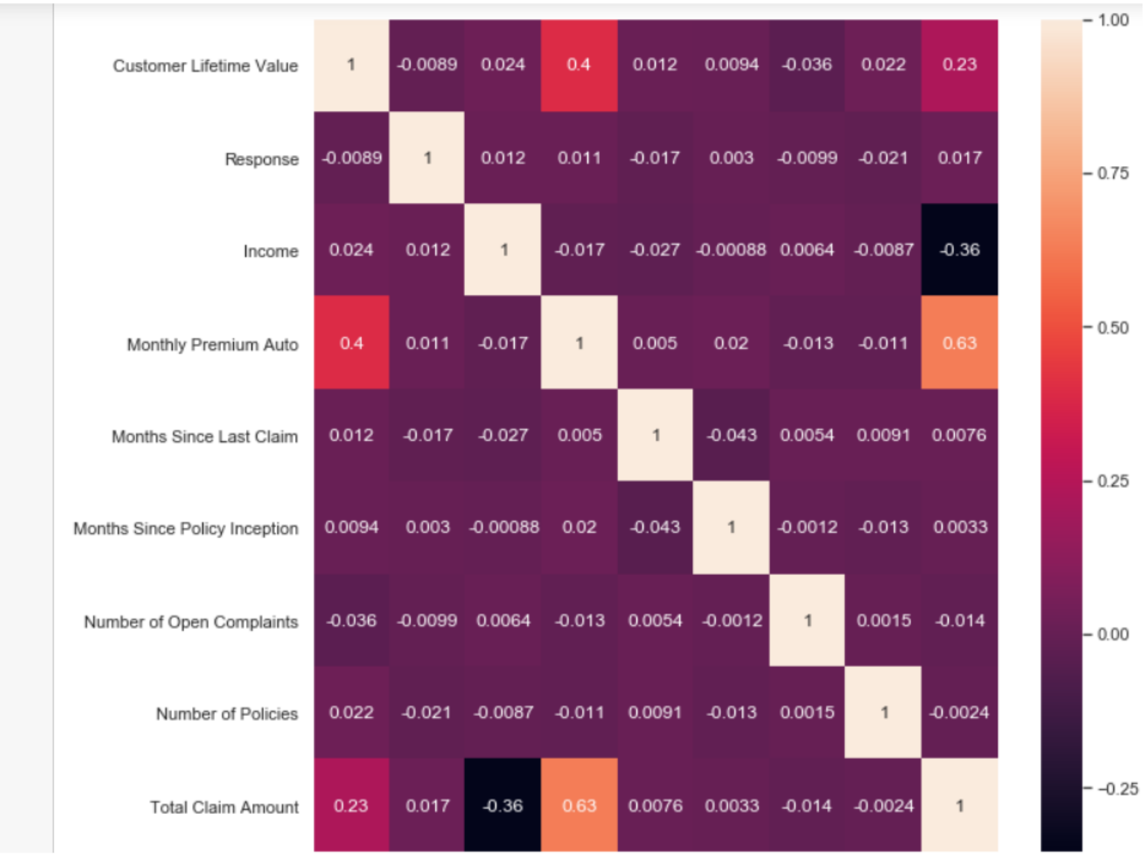
c) With the cleaned data being loaded into Jupyter notebook, null value check is performed.

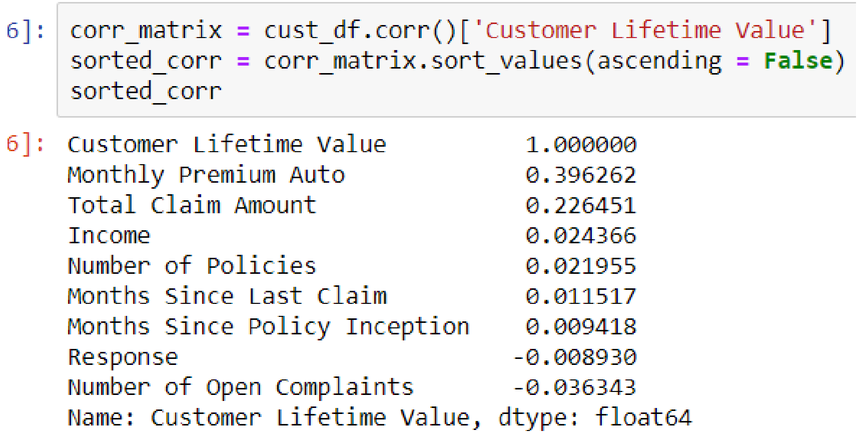


```
In [6]: cust_df.isnull().sum()

Out[6]: Customer                         0
        State                            0
        Customer Lifetime Value          0
        Response                         0
        Coverage                         0
        Education                        0
        Effective To Date                0
        EmploymentStatus                 0
        Gender                           0
        Income                           0
        Location Code                    0
        Marital Status                   0
        Monthly Premium Auto             0
        Months Since Last Claim          0
        Months Since Policy Inception    0
        Number of Open Complaints        0
        Number of Policies               0
        Policy Type                      0
        Policy                           0
        Renew Offer Type                 0
        Sales Channel                    0
        Total Claim Amount               0
        Vehicle Class                    0
        Vehicle Size                     0
        dtype: int64
```

Exploratory Data Analysis:

- As there were both numerical as well as categorical columns in my data, I have focused my evaluation on numerical data by finding correlation with the target variable using correlation matrix and heat map.





```
6]: corr_matrix = cust_df.corr()['Customer Lifetime Value']
    sorted_corr = corr_matrix.sort_values(ascending = False)
    sorted_corr
```

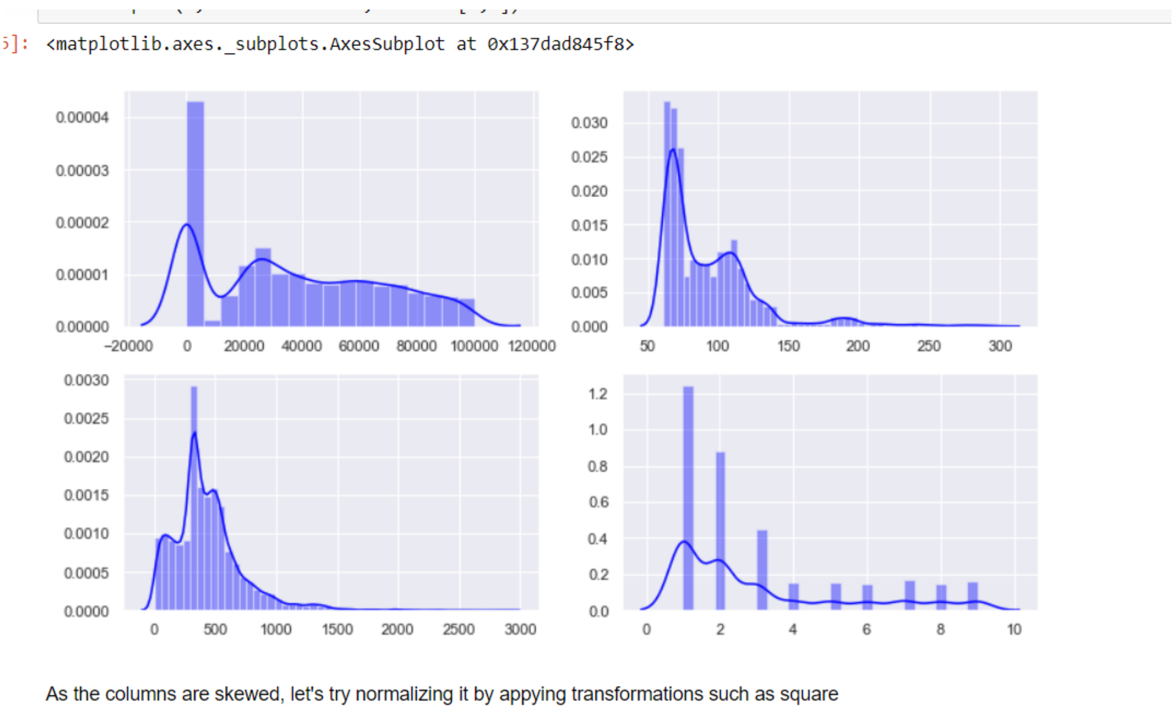
```
6]: Customer Lifetime Value          1.000000
    Monthly Premium Auto             0.396262
    Total Claim Amount               0.226451
    Income                           0.024366
    Number of Policies               0.021955
    Months Since Last Claim          0.011517
    Months Since Policy Inception    0.009418
    Response                        -0.008930
    Number of Open Complaints       -0.036343
    Name: Customer Lifetime Value, dtype: float64
```

- It reveals that Monthly premium Auto, Income, total claim amount has major effect on CLV, with being number of policies and number of open complaints has minor effects.

For these columns,

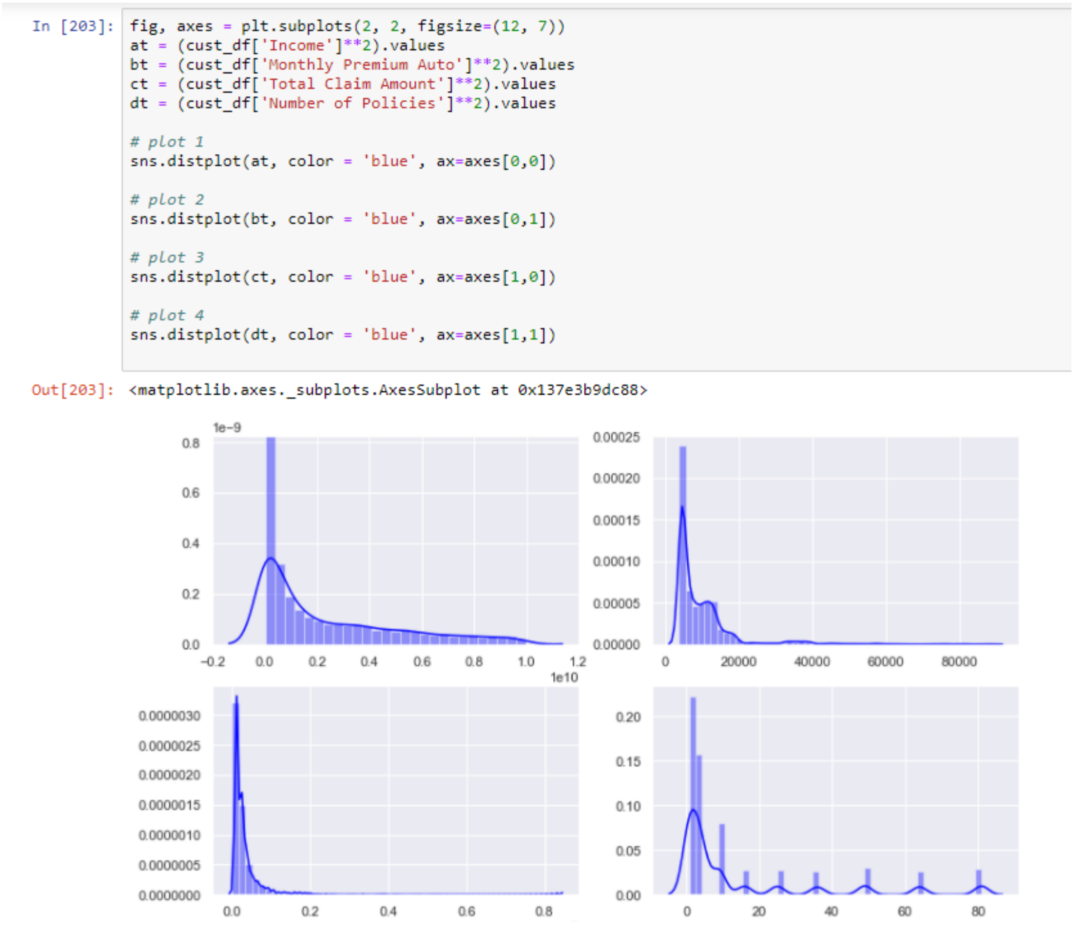
a) Firstly, observed the skewness and tried to normalize it by using the square transformations.

P.Anand Sagar
11197418



```
5]: <matplotlib.axes._subplots.AxesSubplot at 0x137dad845f8>
```

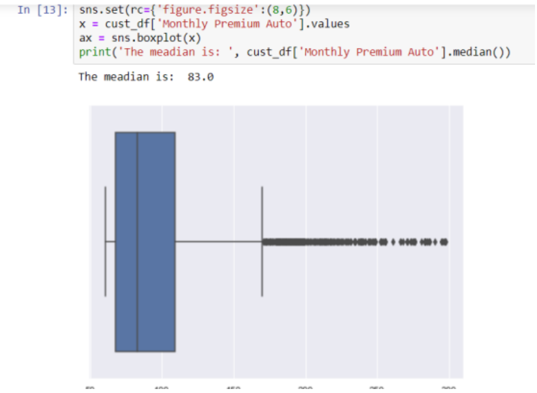
As the columns are skewed, let's try normalizing it by appying transformations such as square



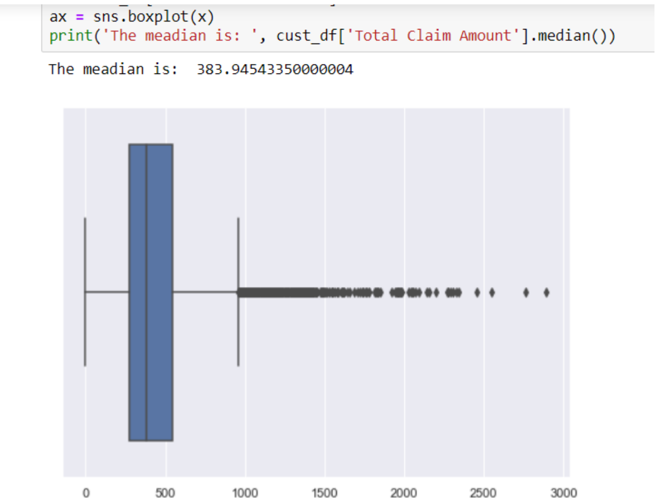
```
In [203]: fig, axes = plt.subplots(2, 2, figsize=(12, 7))
          at = (cust_df['Income']**2).values
          bt = (cust_df['Monthly Premium Auto']**2).values
          ct = (cust_df['Total Claim Amount']**2).values
          dt = (cust_df['Number of Policies']**2).values

          # plot 1
          sns.distplot(at, color = 'blue', ax=axes[0,0])

          # plot 2
          sns.distplot(bt, color = 'blue', ax=axes[0,1])

          # plot 3
          sns.distplot(ct, color = 'blue', ax=axes[1,0])

          # plot 4
          sns.distplot(dt, color = 'blue', ax=axes[1,1])

Out[203]: <matplotlib.axes._subplots.AxesSubplot at 0x137e3b9dc88>
```

But it resulted in more no peaks and increased skewness.

b) Observed the outliers too,



```
In [13]: sns.set(rc={'figure.figsize':(8,6)})
         x = cust_df['Monthly Premium Auto'].values
         ax = sns.boxplot(x)
         print('The meadian is: ', cust_df['Monthly Premium Auto'].median())

The meadian is:  83.0
```



```
ax = sns.boxplot(x)
print('The meadian is: ', cust_df['Total Claim Amount'].median())

The meadian is:  383.94543350000004
```

- As this in insurance company data, removing the outliers would lead to losing potential customers, so decided to remain with the original data.
- With this EDA, Monthly premium auto (i.e., which we pay in installments to insurance provider for coverage), and total claim amount(i.e., the amount which can be claimed from insurance provider), and customers income has impact on CLV, which seems meaningful.

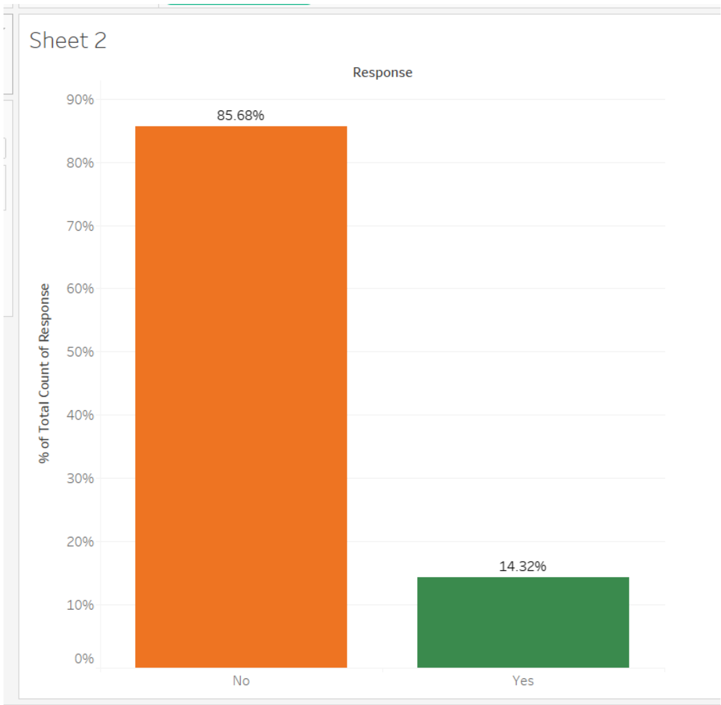
Before working on important categorical variables, and model prediction lets see some visualizations which helps the company to see customer behavior in the market and helps improving decision making.

Data Visualization:

" Response ", the column which explains whether the customer is responding to a particular offer, or by what means they are responding etc. Let's see the engagement rate of the customers by observing response vs other columns.

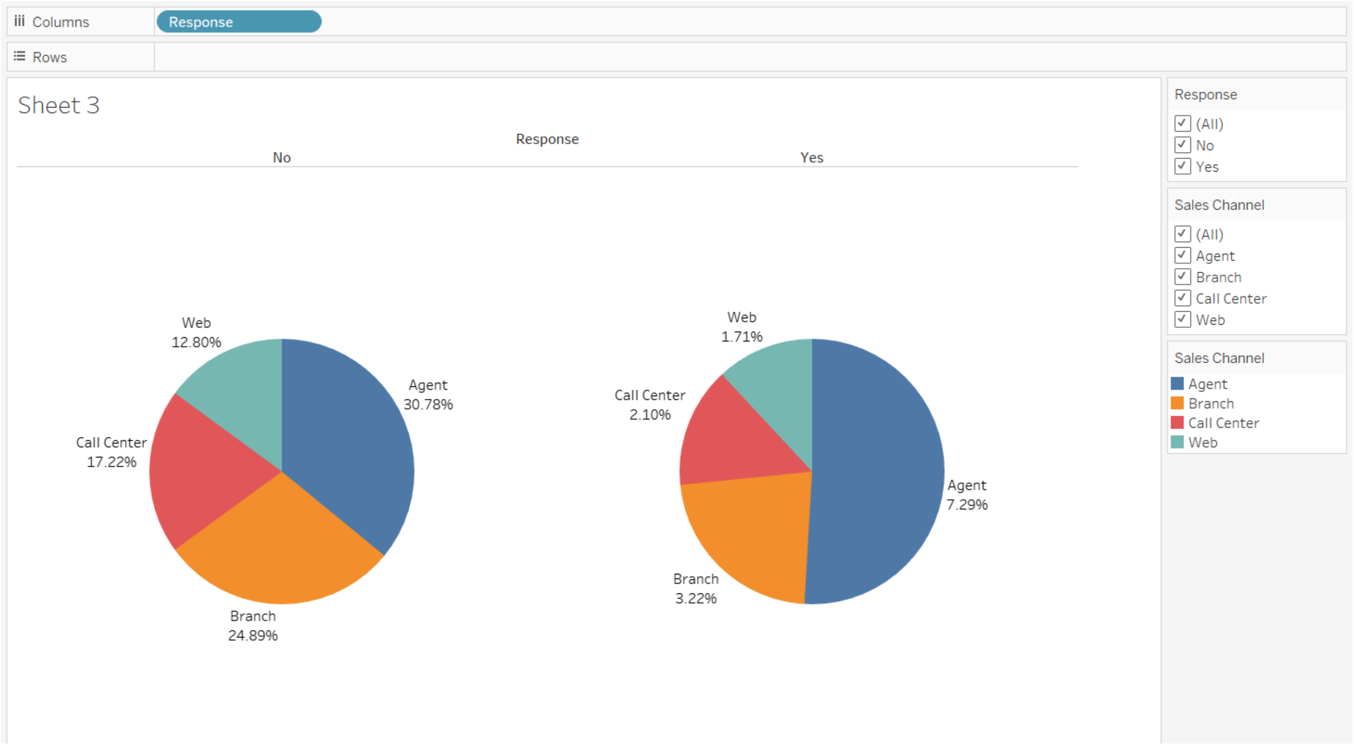
Made use of Tableau and python sea born packages.

1) Engagement Overall



As it is clear that, for this insurance company the overall response rate is low, and key changes have to be made to make the customers respond to company's strategies.

2) By what means insurance company is reaching to customers?(Engagement rate w.r.t sales channel)


Columns: Response

Rows

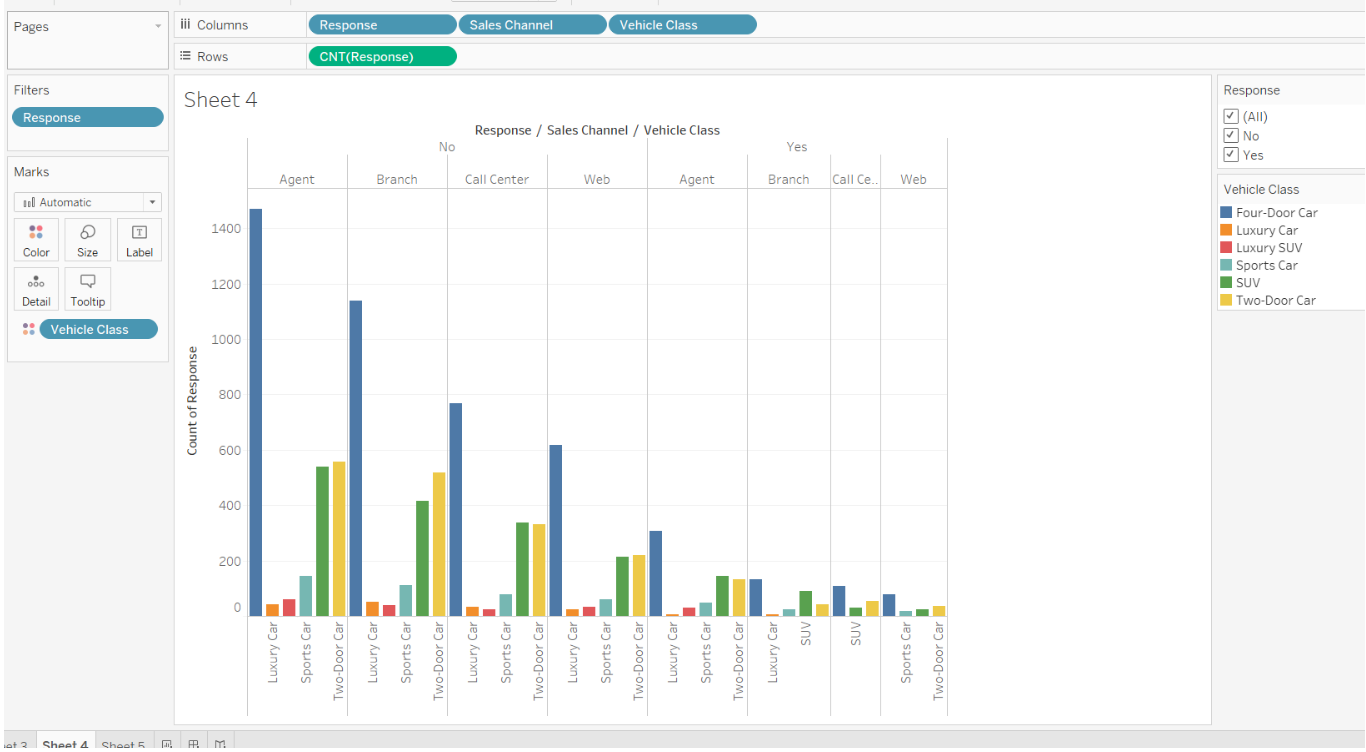
Response
☑ (All)
☑ No
☑ Yes

Sales Channel
☑ (All)
☑ Agent
☑ Branch
☑ Call Center
☑ Web

Sales Channel
🟦 Agent
🟧 Branch
🟥 Call Center
🟩 Web

If we observe the yes part, least percentage of response is through web and call center, and similarly if we see people who do not respond there is considerable percentage who doesn't respond through both these sales channel. So, my key finding here is that if the company takes steps by reaching customers by in person (Agent and branch) rather than virtual(web and call center), it can definitely observe increase in response rate.

3) Going further, the red, yellow bars indicate Luxury cars, have no significant response through any sales channel. So, the blue bar and green bars which says customers with four door and SUV, respond higher when contacted through Agent and branch.



Pages

Columns: Response | Sales Channel | Vehicle Class

Rows: CNT(Response)

Filters: Response

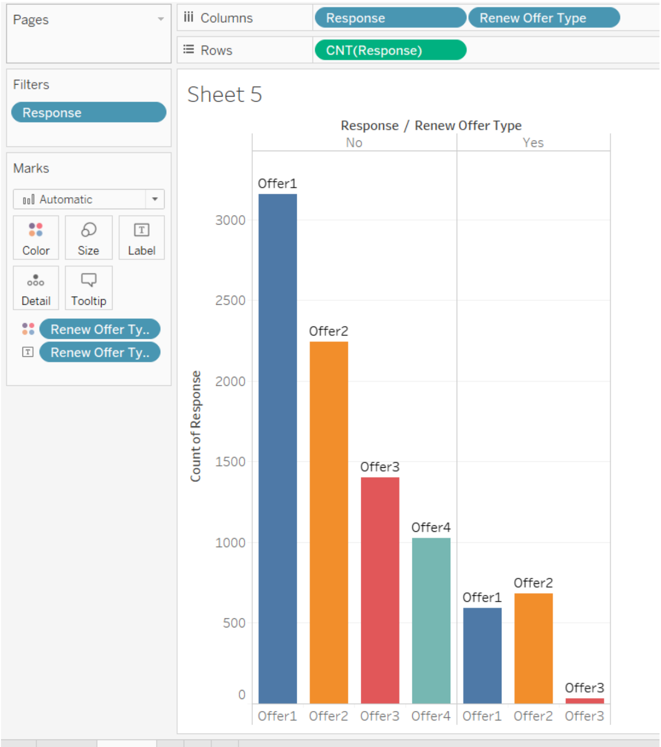
Marks: Automatic

Color: Vehicle Class

Response
☑ (All)
☑ No
☑ Yes

Vehicle Class
🟦 Four-Door Car
🟧 Luxury Car
🟥 Luxury SUV
🟩 Sports Car
🟩 SUV
🟨 Two-Door Car

4) Engagement rate by offer type: There are certain offers which company offers to its customers. As offer 3 and offer 4 are negligible, offering more on 1 and 2 will bring a significant impact.



Columns: Response, Renew Offer Type

Rows: CNT(Response)

Filters: Response

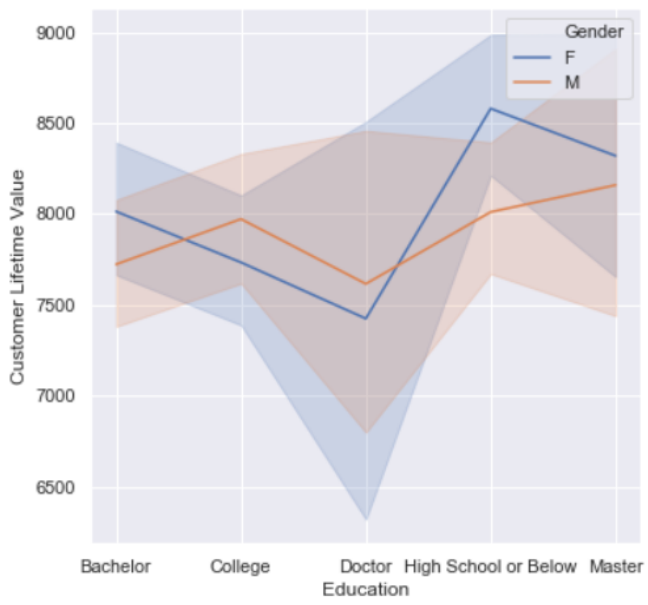
Marks: Automatic

Renew Offer Ty..

Renew Offer Ty..

5) We can say that people who are educated(doctor) have much less customer lifetime value, when compared to the people who studied high school or below. And also, CLV is slightly high in case of females, than males.

```
In [20]: ax = sns.lineplot(y='Customer Lifetime Value', x='Education',hue = 'Gender',data = cust_df)
```



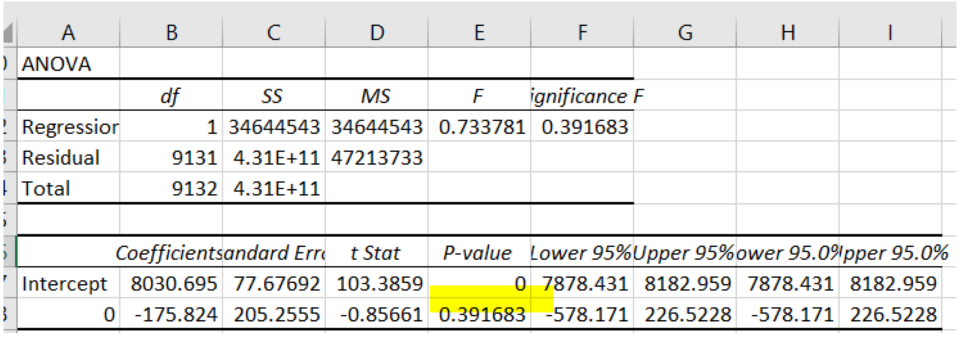
Feature Selection for Categorical Variables:

As we have sorted out important numerical variables, we need to find out the categorical variables which have impact on target variable. Used Anova techniques and bar plots to see the variations.

1) Response

Null hypothesis: Response has significant effect on CLV.

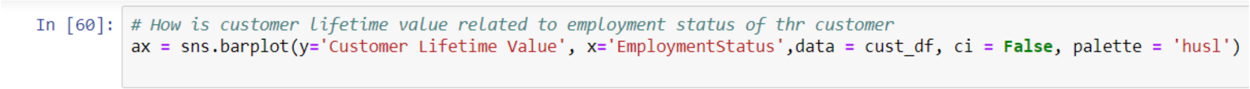
Alternate hypothesis: Response has no effect on CLV, it is only by random chance.



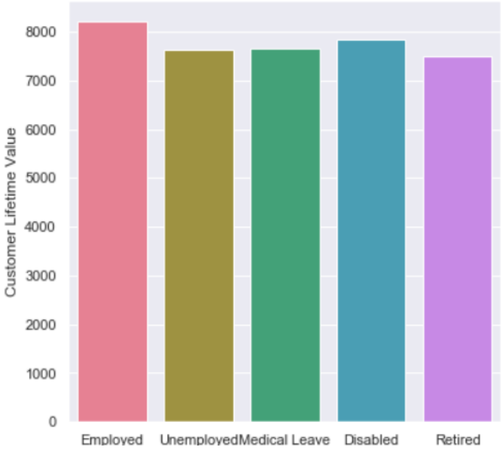
A	B	C	D	E	F	G	H	I
ANOVA								
	df	*SS*	*MS*	*F*	*ignificance F*			
Regressior	1	34644543	34644543	0.733781	0.391683			
Residual	9131	4.31E+11	47213733					
Total	9132	4.31E+11						
	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	8030.695	77.67692	103.3859	0	7878.431	8182.959	7878.431	8182.959
0	-175.824	205.2555	-0.85661	0.391683	-578.171	226.5228	-578.171	226.5228

It is clear that, p-value is greater that 0.05 and hence the column is not significant.

2) Employment Status

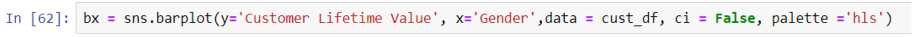


```
In [60]: # How is customer lifetime value related to employment status of thr customer
         ax = sns.barplot(y='Customer Lifetime Value', x='EmploymentStatus',data = cust_df, ci = False, palette = 'husl')
```

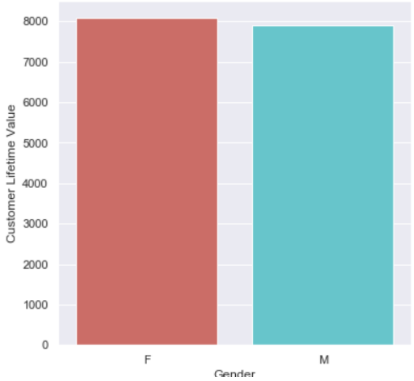


No significant variation in Employment Status variable. So we can ignore this.

3)Gender

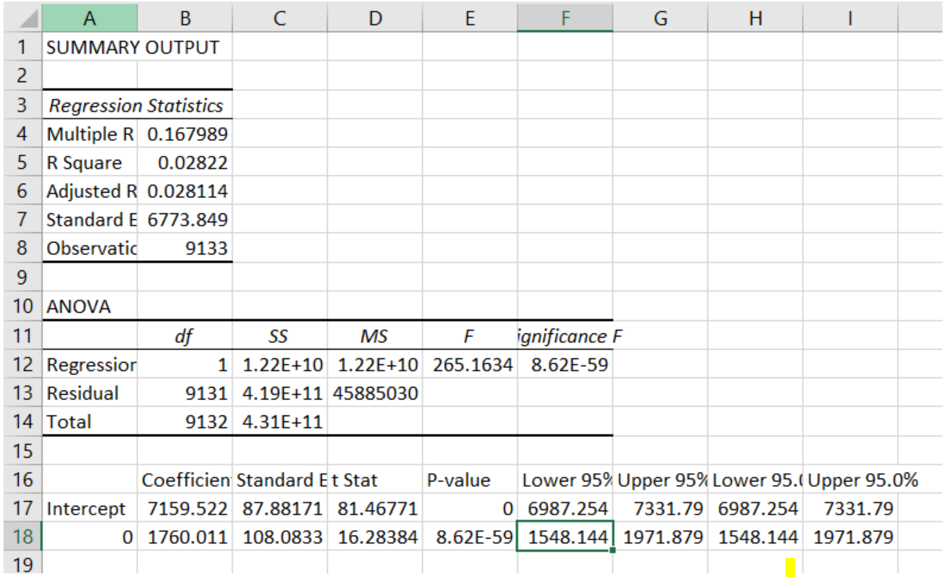


```
In [62]: bx = sns.barplot(y='Customer Lifetime Value', x='Gender',data = cust_df, ci = False, palette ='hls')
```



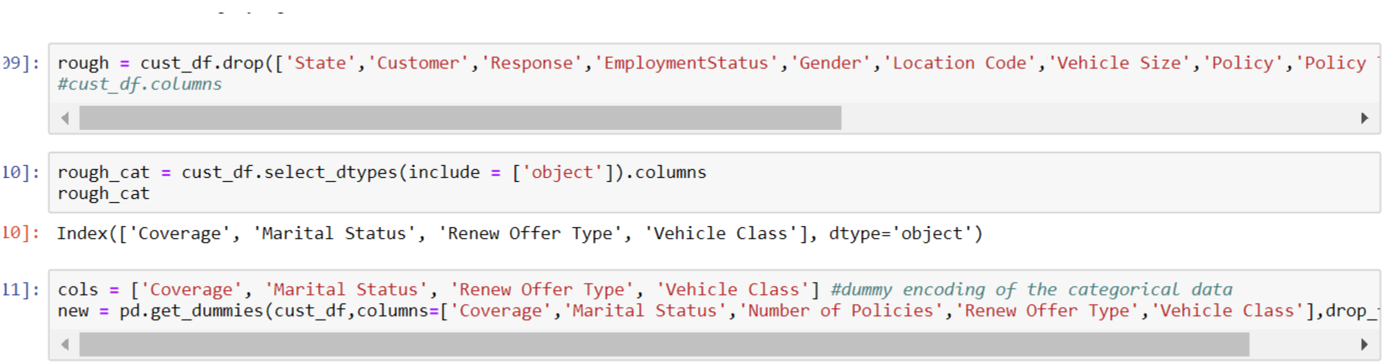
No much Variation in Gender variable w.r.t CLV

3) Coverage – Here p-value is less than 0.05 which means that coverage is a significant variable to predict CLV.



	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.167989							
5	R Square	0.02822							
6	Adjusted R	0.028114							
7	Standard E	6773.849							
8	Observatic	9133							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*ignificance F*			
12	Regressior	1	1.22E+10	1.22E+10	265.1634	8.62E-59			
13	Residual	9131	4.19E+11	45885030					
14	Total	9132	4.31E+11						
15									
16		Coefficien	Standard E	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.(	Upper 95.0%
17	Intercept	7159.522	87.88171	81.46771	0	6987.254	7331.79	6987.254	7331.79
18	0	1760.011	108.0833	16.28384	8.62E-59	1548.144	1971.879	1548.144	1971.879
19									

After finding out the effective columns, removed all the unwanted columns and applied dummy encoding for the categorical variables.

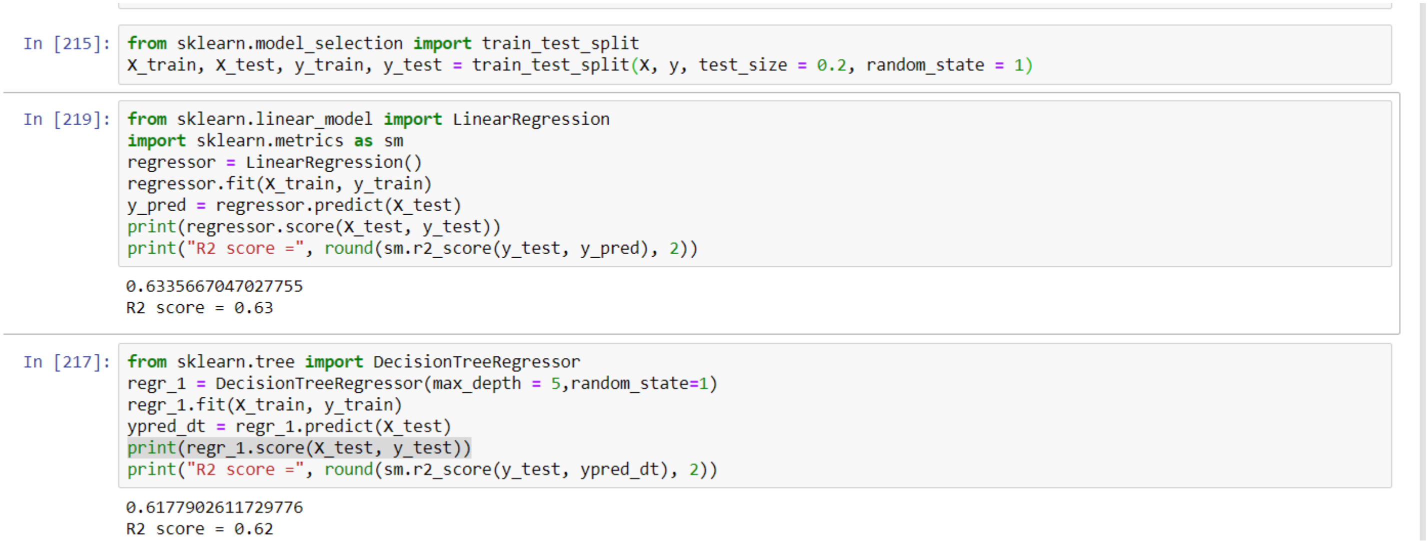


```
99]: rough = cust_df.drop(['State','Customer','Response','EmploymentStatus','Gender','Location Code','Vehicle Size','Policy','Policy
     #cust_df.columns

10]: rough_cat = cust_df.select_dtypes(include = ['object']).columns
     rough_cat

10]: Index(['Coverage', 'Marital Status', 'Renew Offer Type', 'Vehicle Class'], dtype='object')

11]: cols = ['Coverage', 'Marital Status', 'Renew Offer Type', 'Vehicle Class'] #dummy encoding of the categorical data
     new = pd.get_dummies(cust_df,columns=['Coverage','Marital Status','Number of Policies','Renew Offer Type','Vehicle Class'],drop_
```

Model Selection:

Linear Regression model and Decision tree Regressor model:

- Linear regression is a ML model to observe the relation between a dependent variable and one or more independent variables by fitting a linear equation of the data.
- With the selected features, trained the Linear Regression model and achieved a score of 63.3%.
- Decision tree regressor is a predictive model that uses binary rules and predict a target value where it consists of branches, nodes and leaves.
- Built a Decision tree regressor (max depth = 5) model with a score of 61.1%



```
In [215]: from sklearn.model_selection import train_test_split
          X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 1)

In [219]: from sklearn.linear_model import LinearRegression
          import sklearn.metrics as sm
          regressor = LinearRegression()
          regressor.fit(X_train, y_train)
          y_pred = regressor.predict(X_test)
          print(regressor.score(X_test, y_test))
          print("R2 score =", round(sm.r2_score(y_test, y_pred), 2))

          0.6335667047027755
          R2 score = 0.63

In [217]: from sklearn.tree import DecisionTreeRegressor
          regr_1 = DecisionTreeRegressor(max_depth = 5,random_state=1)
          regr_1.fit(X_train, y_train)
          ypred_dt = regr_1.predict(X_test)
          print(regr_1.score(X_test, y_test))
          print("R2 score =", round(sm.r2_score(y_test, ypred_dt), 2))

          0.6177902611729776
          R2 score = 0.62
```

The two models resulted almost the similar accuracy rate, with Linear Regression being slightly high.

Conclusion:

The main motive behind choosing this project is to help business see the trends in the market that what customers are looking for, and predict a monetary value CLV, that helps them to plan for the future, in terms of investing amount on a particular customer. It helps them to identify the issues and retain their customers. With CLV, they can assume whether their best customer in the past going to be best in the future**. Because it costs 10 times less to sell to an existing customer than to find a new customer.**

Related Research Work:

There's a lot of research being carried out in this field, in every organization, as this analysis is the key factor for the businesses to run. Few articles, which I came through for this work are:

https://www.sciencedirect.com/science/article/abs/pii/S109499680570058X

https://www.intechopen.com/books/data-mining/estimating-customer-lifetime-value-using-machine-learning-techniques

https://www.researchgate.net/publication/4752546_Predicting_Customer_Lifetime_Value_in_Multi-Service_Industries

https://addepto.com/how-ltv-and-machine-learning-can-optimise-your-marketing-expenses-and-increase-roi/

https://www.researchgate.net/publication/4752546_Predicting_Customer_Lifetime_Value_in_Multi-Service_Industries