



DATA ANALYTICS REPORT

Open IIT 2019-20

DARK Knights

CONTENTS

- **PROJECT OVERVIEW**
 - **CUSTOMER LIFETIME VALUE(CLV).**
 - **PROBLEM STATEMENT.**
 - **DATA DESCRIPTION.**
 - **EVALUATION CRITERIA.**
 - **ROOT MEAN SQUARE ERROR**
 - **MEAN ABSOLUTE PERCENTAGE ERROR**
- **ANALYSIS**
 - **DATA EXPLORATION AND ANALYSIS.**
 - **FEATURE ENGINEERING.**
- **ALGORITHMS AND TECHNIQUES.**
 - **OVERVIEW**
 - **MODELS USED**
 - **LINEAR REGRESSION MODEL**
 - **XG BOOST MODEL**
 - **SUPPORT VECTOR MACHINE MODEL**
 - **RANDOM FOREST REGRESSION MODEL**
 - **DATA PREPROCESSING**
 - **LABEL ENCODING**
 - **ONE-HOT ENCODER**
 - **GET_DUMMIES**
 - **STANDARD SCALER**
- **RESULT AND CONCLUSION**
- **BUSINESS RECOMMENDATION**
- **PRECAUTIONS**
- **APPENDIX**

PROJECT OVERVIEW

Customer Lifetime Value:

The lifetime value of a customer, or customer lifetime value (CLV), represents the total amount of money a customer is expected to spend in your business, or on your products, during their lifetime. This is an important figure to know because it helps you make decisions about how much money to invest in acquiring new customers and retaining existing ones.

In the big picture, CLV is a gauge of the profit associated with a particular customer relationship, which should guide how much you are willing to invest to maintain that relationship.

Problem Statement


1. For an Auto Insurance Company, predict the customer lifetime value (CLV). CLV is the total revenue the client will derive from the entire relationship with a customer. Because we don't know how long each customer relationship will be, we make a good estimate and state CLV as a periodic value - that is, we usually say "that this customer's 12 months (or 24 months, etc) CLV is \$x."
2. The client also wants to know the types of customers that would generally give us more revenue.

Data Description:-

24 total features are given for 9134 customers in the dataset.

Here is the description of each column:-

- 1) **Customer:** Unique Customer ID for each customer.
- 2) **State:** The state in which the customer lives.

- 
- 3) **Customer Life-Time Value:** The total revenue the client will derive the entire relationship with a customer.
 - 4) **Response:** Whether to renew or not.
 - 5) **Coverage:** The amount of risk or liability that is covered for an individual or entity by way of insurance services.
 - 6) **Education:** Highest educational qualification of a customer.
 - 7) **Effective To Date:** The date to which the insurance is effective.
 - 8) **Employment Status:** Employment Status of the customer.
 - 9) **Gender:** Gender of the customer.
 - 10) **Income:** Income of the customer.
 - 11) **Location Code:** The region of the customer i.e. urban, suburban or rural.
 - 12) **Marital Status:** Marital Status of a person, that is, whether a person is married, single, divorced, etc.
 - 13) **Monthly Premium Auto:** Monthly premium of the insurance.
 - 14) **Months Since Last Claim:** Number of months it has been since a customer last claimed their insurance.
 - 15) **Months Since Policy Inception:** Number of months since the policy was taken by the customer.
 - 16) **Number of Open Complaints:** Number of complaints filed by the customers for claiming insurance.
 - 17) **Number of Policy:** Number of policies availed by a certain customer.
 - 18) **Policy Type:** The category of policy taken i.e. corporate, personal, or special.
 - 19) **Policy:** A policy is a statement of intent, and is implemented as a procedure.
 - 20) **Renew Offer Type:** Type of renew offers.
 - 21) **Sales Channel:** A method of distribution used by a business to sell its products, that is agent, call center, etc.
 - 22) **Total Claim Amount:** Total amount of insurance money claimed by the customer.
 - 23) **Vehicle Class:** Class of the vehicle i.e. two-door, four-door, SUV, luxury SUV, Sports or Luxury.
 - 24) **Vehicle Size:** The size of each vehicle of the customer insured by the company.

Evaluation Criteria:

1. Root Mean Square Error:

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are. RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. Root mean square error is

commonly used in climatology, forecasting, and regression analysis to verify experimental results. The formula for RMSE is as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (Predicted_i - Actual_i)^2}{N}}$$

Here we calculate the predicted value for each test data set. Then we subtract the actual value of the objective from the predicted value and square it. Then we divide by the total number of training data sets. Then evaluate the square root of the result which gives us the RMSE.

2. Mean Absolute Percentage Error:

The mean absolute percentage error (MAPE), also known as mean absolute percentage deviation (MAPD), is a measure of prediction accuracy of a forecasting method in statistics, for example in trend estimation, also used as a loss function for regression problems in machine learning. It usually expresses accuracy as a percentage, and is defined by the formula:

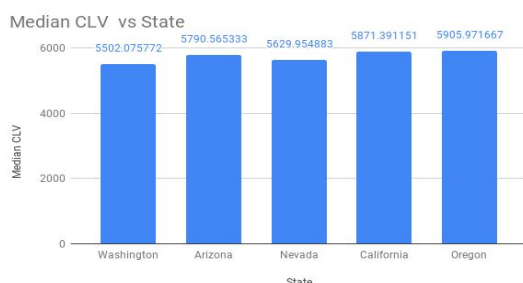
$$MAPE = \frac{100\%}{n} \sum \left| \frac{y - \hat{y}}{y} \right|$$

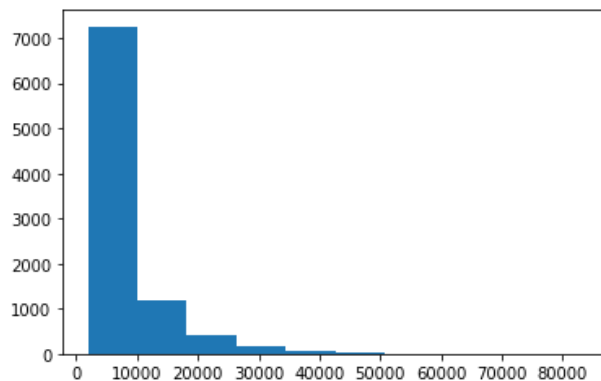
Multiplying by 100% converts to percentage
 The residual
 Each residual is scaled against the actual value

ANALYSIS

Data Exploration and analysis:

1) Median CLV v/s State:





The above graph shows that the median CLV of all the states is almost constant. But Oregon has a slightly higher median.

2)CLV Distribution:

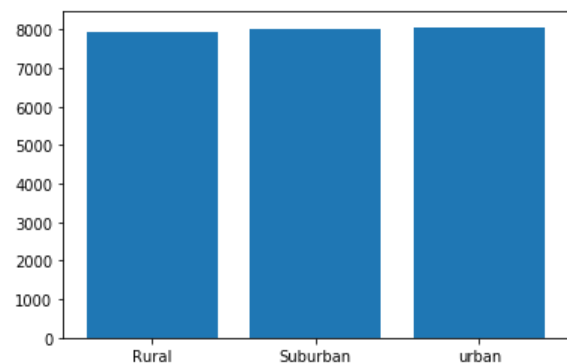
Number of People Vs CLV

The above graph shows that the graph of

CLV is skewed that is the majority of people have CLV under 10,000\$.

3)Region Redundancy:

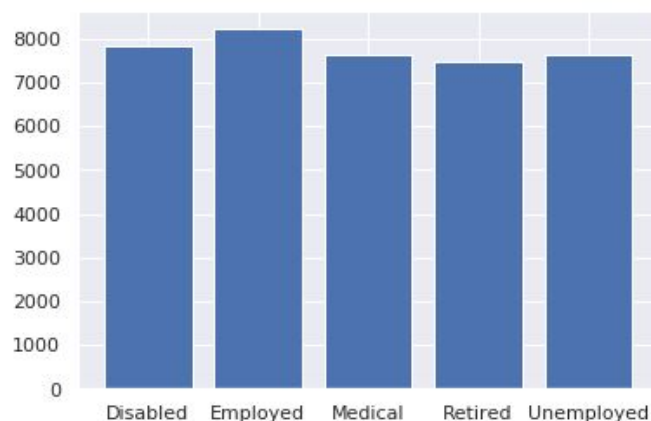
The total CLV is the same across the different location codes.



4) Avg. CLV Vs Location Code

The above graph shows that the average CLV spread across the different location codes is about constant. So we can avoid the field.

5)Average CLV Vs Employment Status

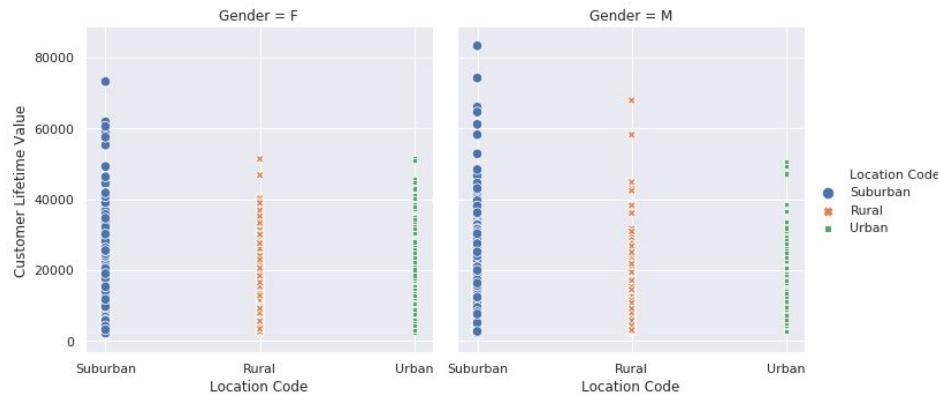


a)Unemployed=7600

b)Employed= 8216

The above graphs show that the average CLV of the Employed class is higher than the rest. And the average CLV of the other classes is about constant.

6) Scatter plot of Gender for CLV Vs Location Code



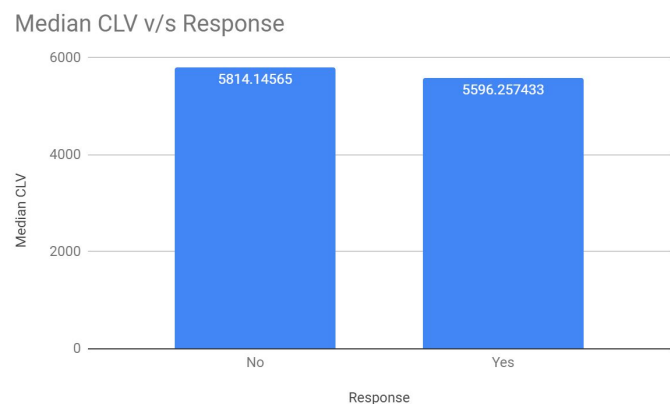
a) Suburban
Male=3023
Suburban
Female=2756
b) Rural Male=769
Rural
Female=1004
c) Urban Male=684
Urban
Female=898

Urban Female=898

CLV for both males and females is more or less the same across all location codes. Max CLV was more in sub-urban areas for both males and females. Also, more no. of customers were from suburban areas regardless of the gender.

7) Median CLV Vs Response:

Median CLV of the responses having no as the answer is higher than that of having yes as an answer.



FEATURE ENGINEERING:

The data set provided contained 24 features. Each data was assessed and its relationship with CLV was analyzed with the help of graphs plotted for features against the result corresponding to that feature.

After studying the correlation matrix, certain features having relatively less influence in CLV determination were attempted at merging by concatenating the columns and deleting the less significant one. The features that were merged were :

- Renew offer Type and Policy
- Policy Type and Customer ID
- State and Location Code

- Vehicle class and Vehicle Size
- Employment Status and Income

The result of this merging was not as significant as expected and negligible changes in RMSE and MAPE were observed and hence feature combination was not a fruitful attempt towards improving our hypothesis/ model accurate.

Next, we attempted feature elimination in order to decrease insignificant feature in the hypothesis, for that again we took the help of the correlation matrix and results of data exploration.

The features omitted in hypothesis formation were :

- 1) **Gender:** As the average CLV vs Gender graph was almost constant.
- 2) **Location Code:** Average CLV vs Location Code graph was almost constant.
- 3) **Policy Type:** Since the policy provided a more general field of the policy type.
- 4) **Customer:** Since it is unique for each customer it does not affect the CLV.

ALGORITHMS AND TECHNIQUES

OVERVIEW:-

Models implemented for obtaining the evaluation matrix were linear regression, polynomial regression, SVM regression, Random Forest Regression.

Random Grid search was also implemented on a random forest regressor.

The data provided was divided into test and train data with a test: original data set ratio = 0.25

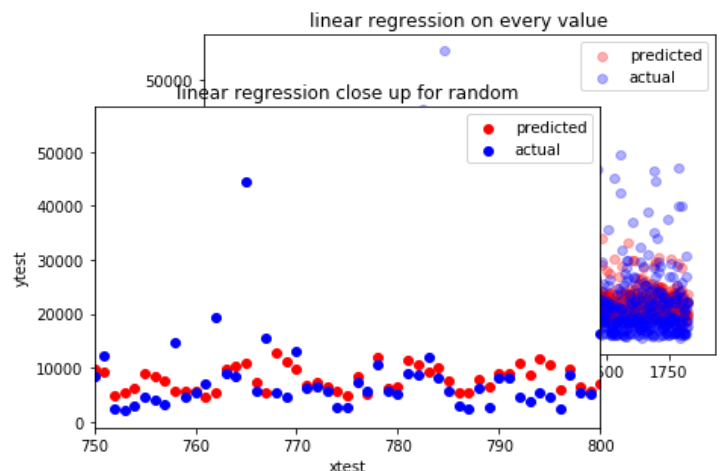
MODELS USED:-

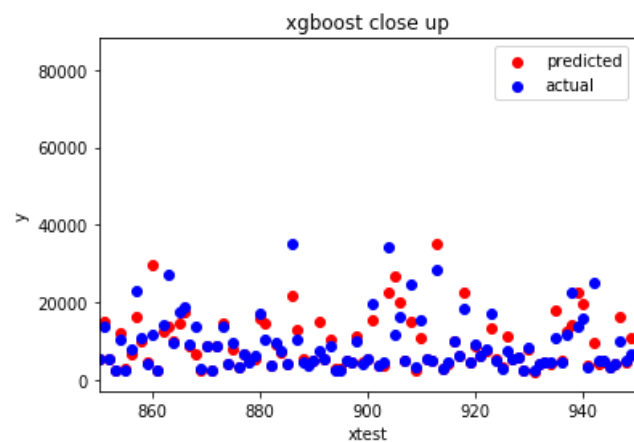
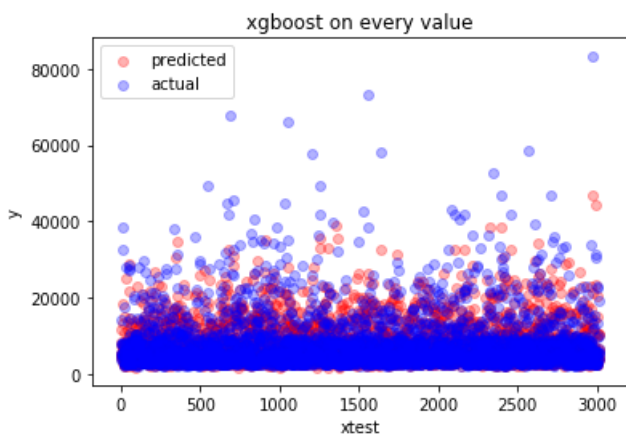
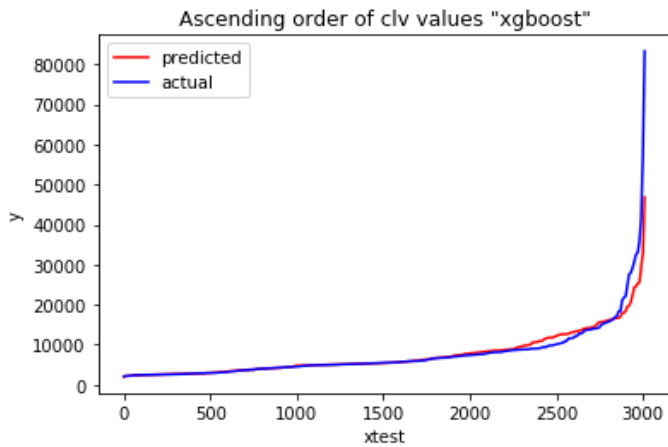
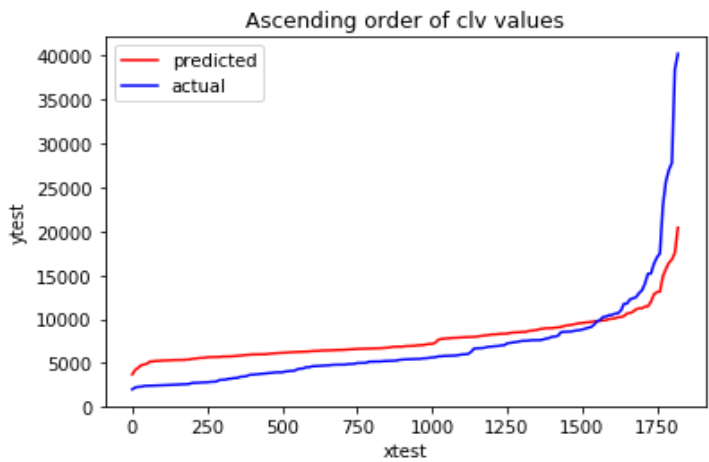
1.LINEAR REGRESSION MODEL

First, we used a linear regression model with label encoding as it is the most basic model. But the RMSE and MAPE values were too high thus the model was rejected.

2.XGBOOST MODEL:

XGBoost model was then used as it generally gives very accurate results by superposing models. This time one hot encoder was used. But it also gave too high RMSE and MAPE value so it was rejected.

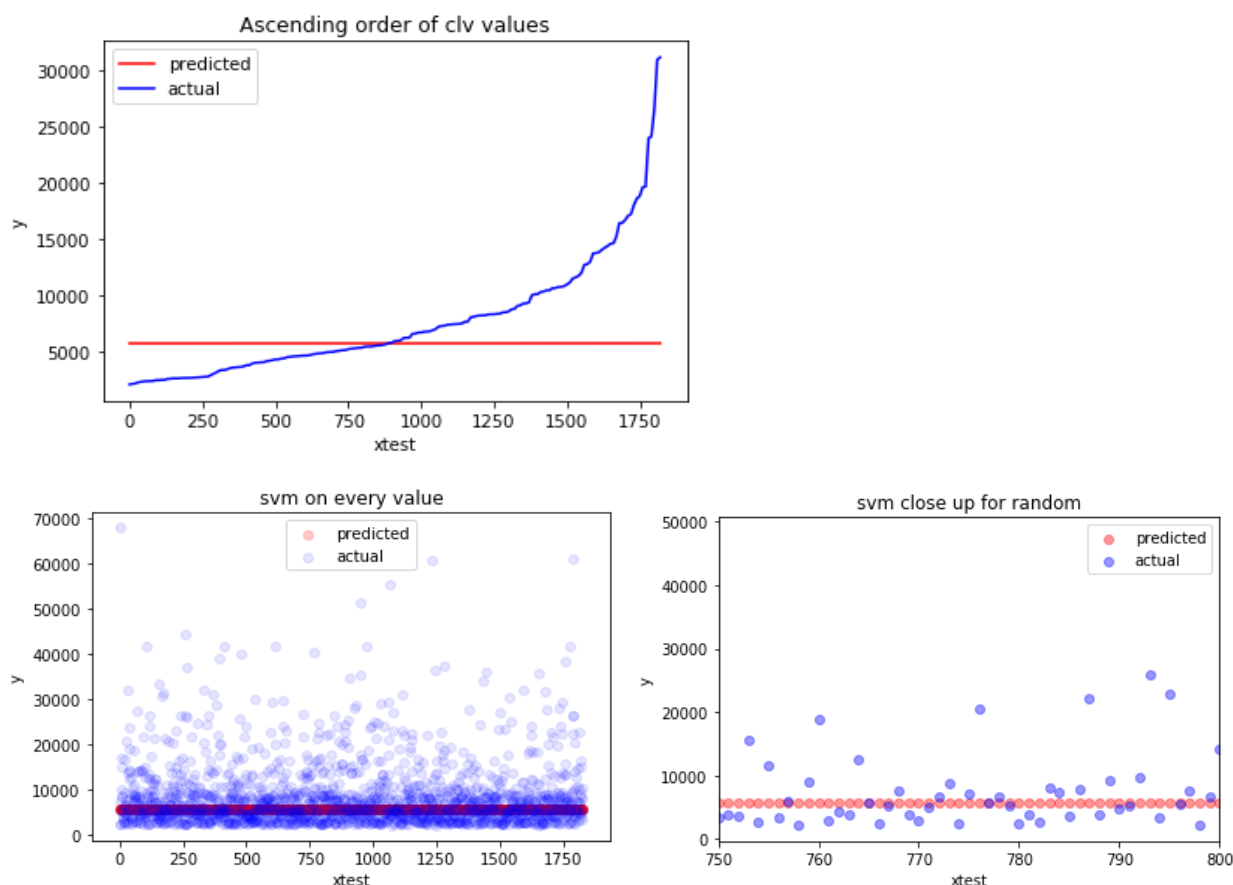




3.SUPPORT VECTOR MACHINE MODEL

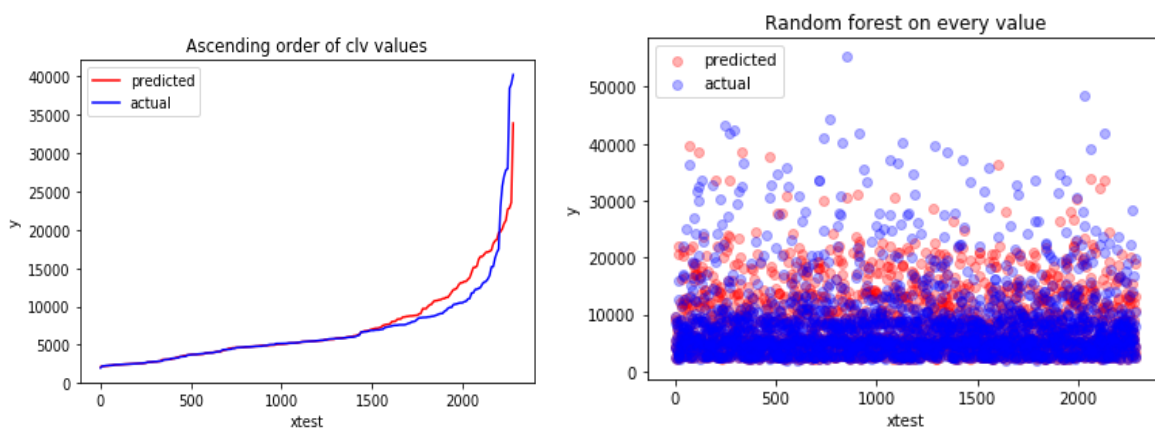
In simple regression, we try to minimize the error rate. While in SVR we try to fit the error within a certain threshold. The predicted data was limited to a range(within the boundary lines), and the actual test data is well spread across the xtest range and thus it was giving a high RMSE and

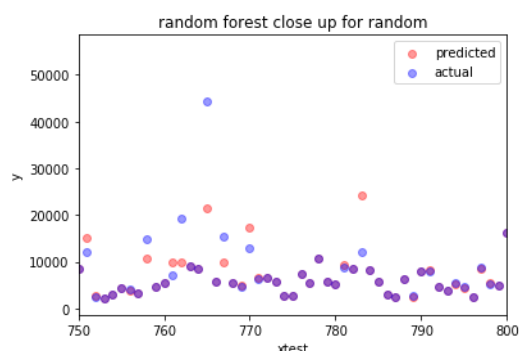
MAPE. The following graph helps to show that beyond the point of intersection the error increases at a larger rate and hence this model was rejected.



4.RANDOM FOREST REGRESSION MODEL

Random forest is a model that uses a number of decision trees to bring out the best decision tree to predict the given set of data. Random forest was first with one hot encoder. The RMSE and MAPE value was then reduced by encoding it with Get_dummies. Then we used a random search CV and cross-validation to reduce the RMSE and MAPE further to [3504](#).





MODELS USED	RMSE	MAPE
LINEAR REGRESSION MODEL	6079	60.03
SUPPORT VECTOR REGRESSION MODEL	7353	51.69
XGboost Model(objective= linear regression)	4073	13.89
RANDOM FOREST REGRESSION MODEL(random search and get dummies)	3504	10.84

DATA PRE-PROCESSING :-

1.Label encoding:-

Label encoder encodes labels with a value between 0 and $n_{\text{classes}}-1$ where n is the number of distinct labels. If a label repeats it assigns the same value as assigned earlier.

```
In [134]: from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
```

```
In [136]: for i in x.columns :
            if (x[i].dtype == 'object'):
                x[i]= le.fit_transform(x[i])
```

2.One-Hot Encoder:-

One Hot Encoding takes a column which has categorical data, which has been label encoded and then splits the column into multiple columns. The numbers are replaced by 1s and 0s, depending on which column has what value.

RMSE obtained on using One Hot Encoder in Random Forest Model is [3703](#).

```
In [54]: from sklearn.preprocessing import OneHotEncoder
onehotencoder = OneHotEncoder()
x = onehotencoder.fit_transform(x).toarray()
```

3. Get_dummies:-

Pandas `get_dummies` method is a very straight forward one-step procedure to get the dummy variables for categorical features. The advantage is you can directly apply it to the data frame and the algorithm inside will recognize the categorical features and perform get dummies operation on it. Another advantage is that it can operate on values other than integers (so you don't need the `LabelEncoder`) and returns a DataFrame with the categories as column names.

```
In [403]: x = pd.get_dummies(x)
```

```
In [404]: x.head()
```

```
Out[404]:
```

	Income	Monthly Premium Auto	Months Since Last Claim	Months Since Policy Inception	Number of Open Complaints	Number of Policies	Total Claim Amount	State_Arizona	State_California	State_Nevada	...	Sales Channel_Web	Vehicle Class_Four- Door Car	Clas
0	56274	69	32	5	0	1	384.811147	0	0	0	...	0	0	
1	0	94	13	42	0	8	1131.464935	1	0	0	...	0	1	
2	48767	108	18	38	0	2	566.472247	0	0	1	...	0	0	
3	0	106	18	65	0	7	529.881344	0	1	0	...	0	0	
4	43836	73	12	44	0	1	138.130879	0	0	0	...	0	1	

5 rows × 68 columns

4. Standard Scaler:-

The idea behind `StandardScaler` is that it will transform your data such that its distribution will have a mean value 0 and a standard deviation of 1. Given the distribution of the data, each value in the dataset will have the sample mean value subtracted and then divided by the standard deviation of the whole dataset.

```
In [ ]: from sklearn.preprocessing import StandardScaler

sc = StandardScaler()
x_train = sc.fit_transform(x_train)
x_test = sc.transform(x_test)
```

RESULT AND CONCLUSION

For this problem statement, we have tried various machine learning models, such as Linear Regression (most basic), Xgboost, Random Forest, and Support Vector Machine (SVM), to obtain the best compatible model for the given data.

The above-listed models gave us different values of error.

Out of the following four models, Random Forest was the one which provided the best fitting method, which is the least error which is about 11.25%(Mean Absolute Percentage Error) and hence providing an accuracy of 89%.

For the random forest model:-

Root Mean Square Error = 3504

MAPE = 10.84%

PRECAUTIONS

Whenever working on a data set to predict or classify a problem, we tend to find accuracy by implementing a design model on the first train set, then on the test set. If the accuracy is satisfactory, we tend to increase accuracy of data-sets prediction either by increasing or decreasing data feature or features selection or applying feature engineering in our machine learning model. But sometimes our model may be giving poor results.

The poor performance of our model may be because, the model is too simple to describe the target, or maybe the model is too complex to express the target.

In machine learning, we predict and classify our data in a more generalized way. So in order to solve the problem of our model that is overfitting and underfitting we have to generalize our model. Statistical speaking how well our model fit to data set such that it gives proper accurate results as expected.

BUSINESS RECOMMENDATIONS

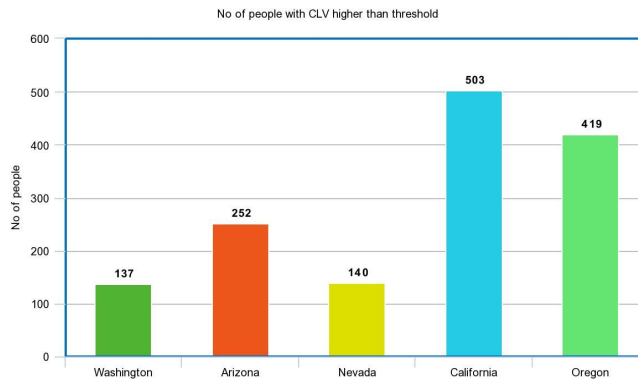
With the help of feature extraction and vivid data analysis, we got many insights into the working of the insurance agency.

A primitive observation is that about 70% of customers generate CLV less than \$12000, hence it can be considered as a threshold CLV.

For the purpose of achieving CLV above threshold value the analysis conducted can be consolidated to the following headers: -

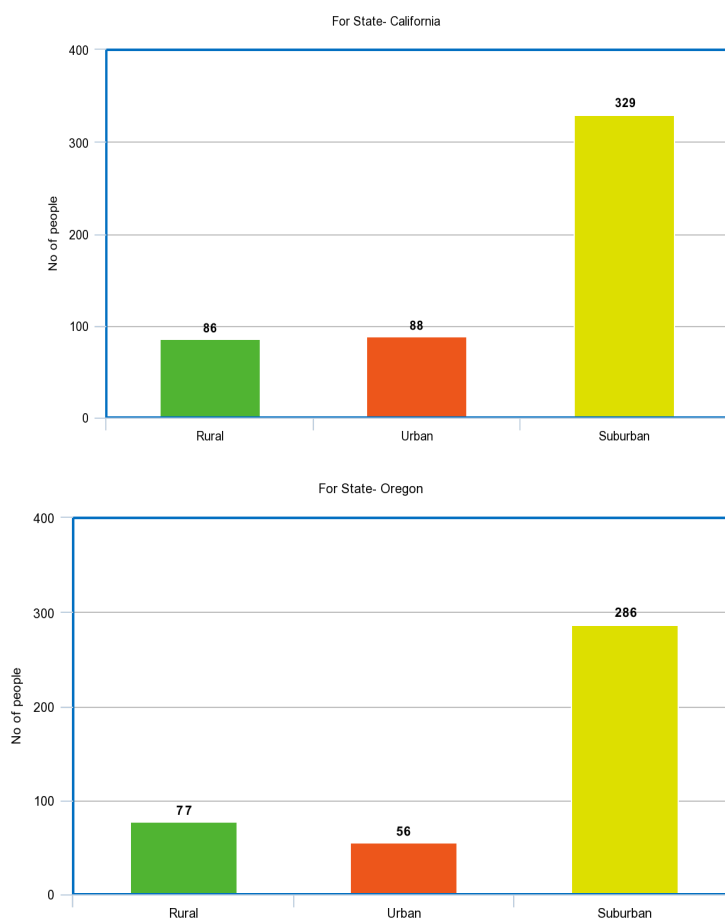
[Note: The graphs plotted contain CLV values above the set threshold of \$12000 plotted against the following parameters]

- **State-wise comparison to the number of customers having CLV above threshold:**



As evident from the graph, it would be profitable if we focus on the states of California and Oregon.

For the state of Oregon and California further statistics can be interpreted with the help of graphs below:

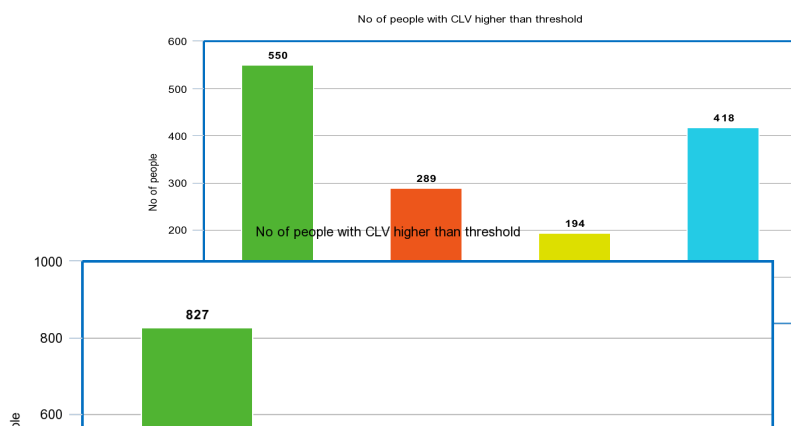


- ❖ Suburban regions generate relatively higher CLVs. Sub-urbans are under constantly under development which reflects in growing demand for automobiles and auto insurance. Hence focusing on suburban regions is advisable.

● Sales agent vs number of customers:

The old-school methodology of making insurances offline by going to branches and through sales agent services still prevails over modern mediums of web portals as well as call centre.

Focus should be on creating more media and online awareness through advertisements in online web and entertainment platforms.



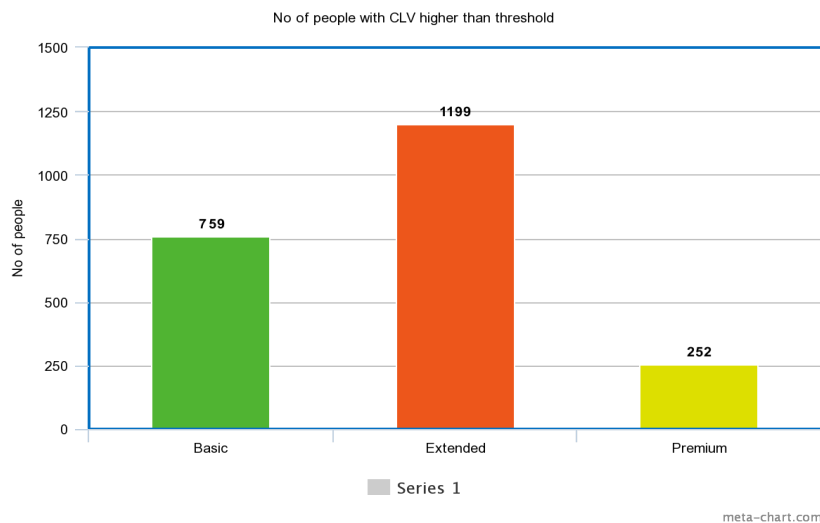
- Even focusing on married couples will be more useful and hence profitable.

A common notion among the youth and single population to be “care-free” and independent evidently does not convince them to play safe and consider taking vehicle insurance.

On the other hand married couples are more inclined to the sense of “safe being” and this reflects in them being the most profitable customer group for the company.

Meanwhile couples continue being the go to customers focus should be on reaching out to the youth and single population by reaching out to them through web awareness campaigns and online advertisements.

- **Most of the people are involved in extended coverage and hence it will be more beneficial to promote the extended coverage.**



APPENDIX

Software Used:

- 1) Anaconda Navigator
- 2) Jupyter Notebook


Libraries Used:

- 1) Matplotlib
- 2) Numpy
- 3) Pandas

- 4) Scikit Learn
- 5) Seaborn
- 6) XGBoost

References

- 1) <https://www.geeksforgeeks.org/python-linear-regression-using-sklearn/>
- 2) <https://stackabuse.com/random-forest-algorithm-with-python-and-scikit-learn/>
- 3) <https://www.geeksforgeeks.org/ml-label-encoding-of-datasets-in-python/>
- 4) <https://www.geeksforgeeks.org/ml-one-hot-encoding-of-datasets-in-python/>
- 5) <https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989>
- 6) https://www.geeksforgeeks.org/python-pandas-series-str-get_dummies/
- 7) https://en.wikipedia.org/wiki/Customer_lifetime_value
- 8) <https://www.shopify.in/encyclopedia/customer-lifetime-value-clv>
- 9) <https://www.statisticshowto.datasciencecentral.com/rmse/>
- 10) https://en.wikipedia.org/wiki/Mean_absolute_percentage_error
- 11) <https://www.kaggle.com/zhukov/simple-xgboost>
- 12) <https://gist.github.com/amanahuja/6315882>
- 13) <https://machinelearningmastery.com/gentle-introduction-xgboost-a-pplied-machine-learning/>
- 14) <https://medium.com/@sagar.rawale3/feature-selection-methods-in-machine-learning-eaeef12019cc>
- 15) <https://towardsdatascience.com/data-visualization-for-machine-learning-and-data-science-a45178970be7>
- 16) <https://medium.com/@sagar.rawale3/feature-selection-methods-in-machine-learning-eaeef12019cc>

- 
- 17) https://www.researchgate.net/publication/335491240_Hyperparameter_Tuning?fbclid=IwAR0Fy7uGoQda_sNDshcbrcB4IOcF0eGiMgL12doDU9diP2H8OqmmJ_fCbsQ
 - 18) <https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74>