TEAM SCHEDIASTIKS

INDEX

INTRODUCTION

The report outlines the analysis done on the customer data collected from the five different cities of United States. The objective of the analysis is to determine **Customer Lifetime Value (CLV).** CLV is the total revenue the client will derive from their entire relationship with a customer. CLV is determined to find the duration of a customer relationship, i.e. how long a customer relationship will be. This would also help the company to predict the potential customers that would generally give them more revenue.h

The dataset given for the competition contains customer's data including customer ID, State, Education, Employment, Income, Marital Status, Monthly premium auto, Policy type, Vehicle type, Vehicle size, etc.

The dataset was found not to be linear. Data set was stacked into K folds and each fold of the data set passes through some base models like Random forest, XGBoost, Cat Boost,Tabular Learner with Fast AI,etc.

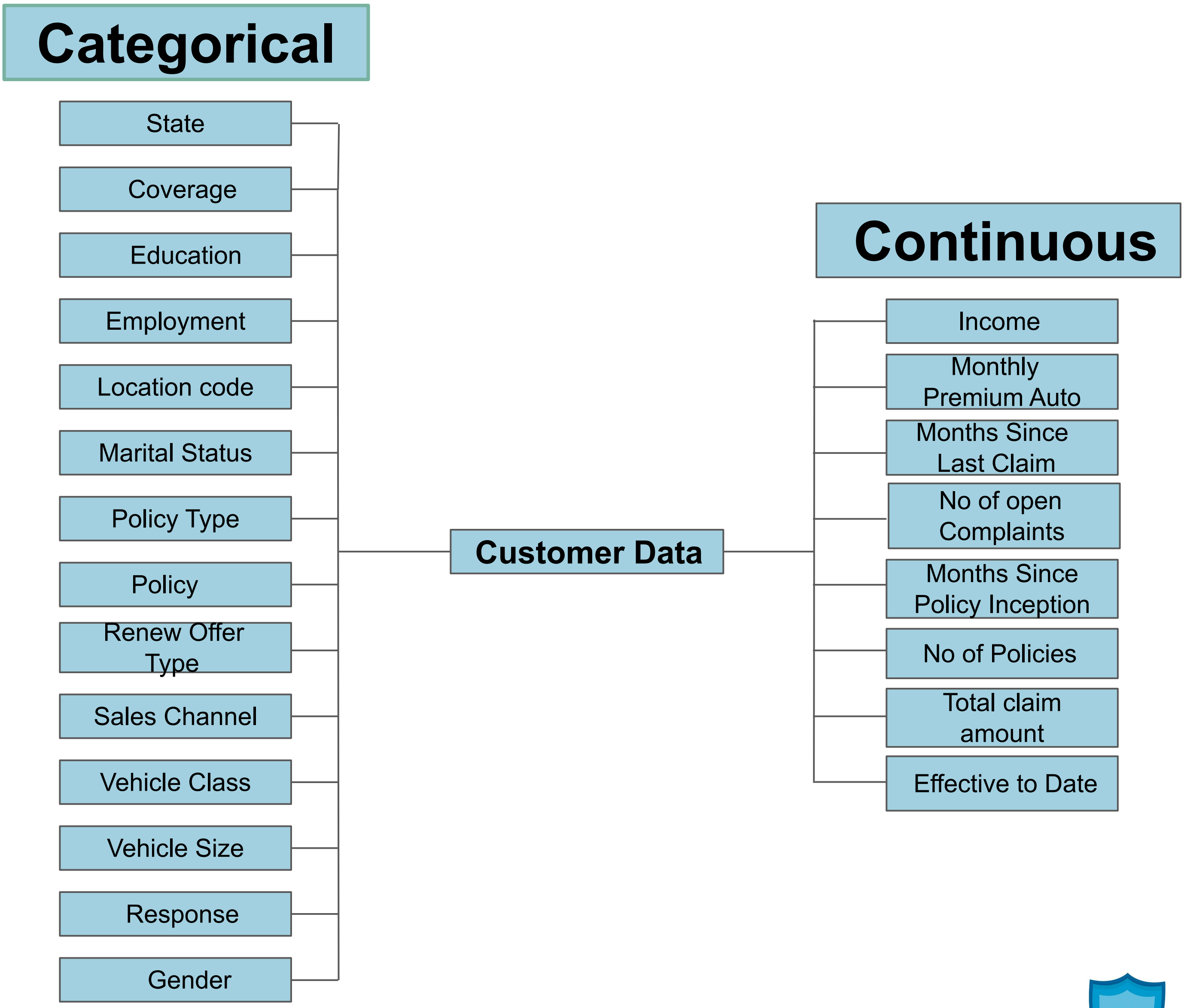
DATA DESCRIPTION

The subjective analysis of the variables present in the automobile insurance company statistics gives the entire relationship between customer and the insurance company . The data is provided for predicting the CLV(customer lifetime value). CLV is the total revenue the client will derive from their relationship with a customer.

The dataset provided here is a subset of original dataset . This dataset has sample data from the months January and February 2011 . In order to maintain discrepancy and avoid leakage of proprietary information some fields in the current dataset are anonymized; while the usage data has also been anonymously scaled or randomized by single or constant factor.

The training data is a pretty balanced dataset containing approximately 10,000 entries for each customer. The user data is perfectly categorized into different columns and covers almost all types of information about customer required for the company.

- Many machine learning algorithms cannot operate on label data directly. They require all input variables and output variables to be numeric.So, all categorical attributes were One-Hot-Encoded.
- Initially there are 22 attributes, after encoding labels total number of attributes were increased to 65.
- Since, most of the learning algorithms use Euclidean distance between two data points in their computations, the continuous attributes were normalized, so that model can gain more insights from the dataset.



Categorical

State

Coverage

Education

Employment

Location code

Marital Status

Policy Type

Policy

Renew Offer Type

Sales Channel

Vehicle Class

Vehicle Size

Response

Gender

Customer Data

Continuous

Income

Monthly Premium Auto

Months Since Last Claim

No of open Complaints

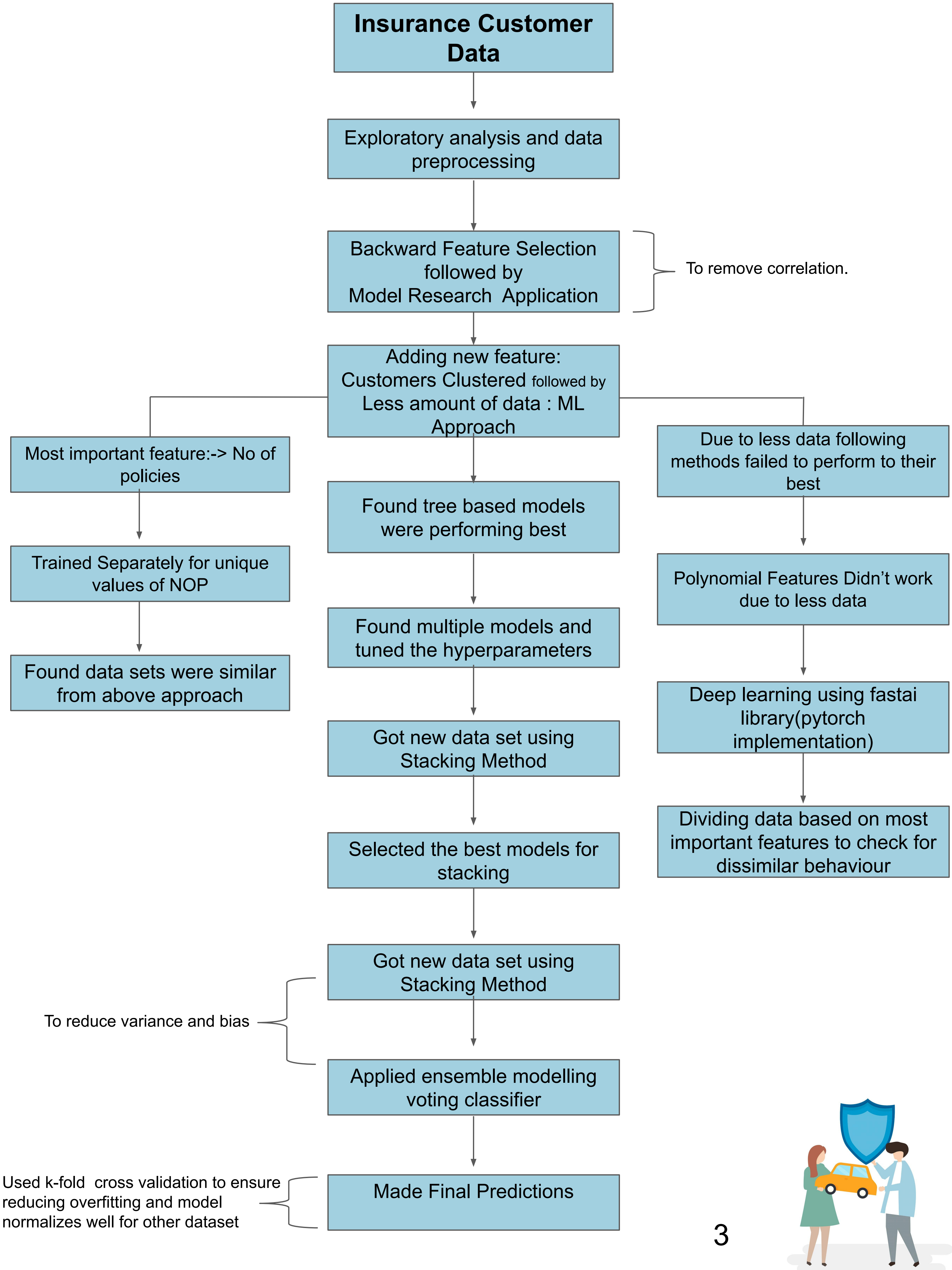
Months Since Policy Inception

No of Policies

Total claim amount

Effective to Date

MODEL APPROACH



Insurance Customer Data

Exploratory analysis and data preprocessing

Backward Feature Selection followed by Model Research Application

To remove correlation.

Adding new feature: Customers Clustered followed by Less amount of data : ML Approach

Most important feature:-> No of policies

Trained Separately for unique values of NOP

Found data sets were similar from above approach

Found tree based models were performing best

Found multiple models and tuned the hyperparameters

Got new data set using Stacking Method

Selected the best models for stacking

Got new data set using Stacking Method

Applied ensemble modelling voting classifier

To reduce variance and bias

Made Final Predictions

Used k-fold cross validation to ensure reducing overfitting and model normalizes well for other dataset

Due to less data following methods failed to perform to their best

Polynomial Features Didn’t work due to less data

Deep learning using fastai library(pytorch implementation)

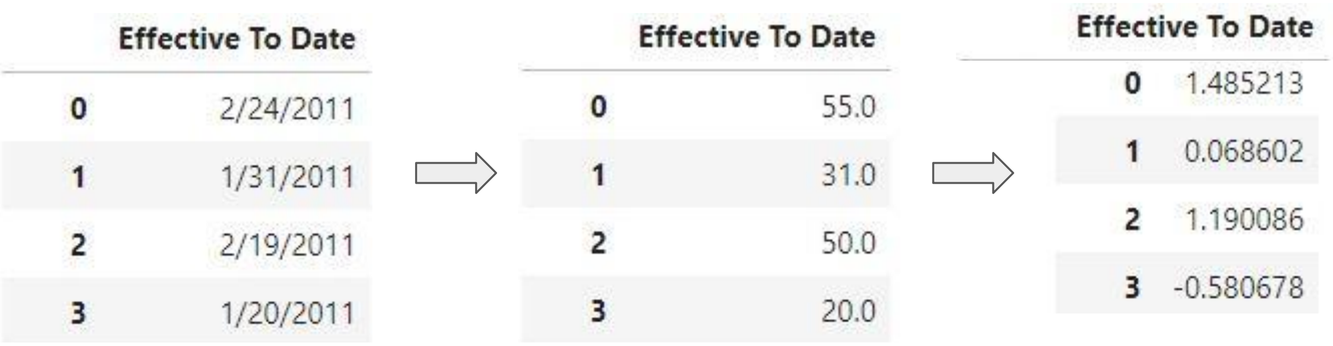
Dividing data based on most important features to check for dissimilar behaviour

DATA CLEANING

Incorrect or inconsistent data may lead to false predictions and conclusions. That is why before investigating the data, data cleansing is done. Data cleansing improves the data quality and in doing so, increases overall productivity.
We checked for the errors or corruptions or unreal values in the data.
But there are no such values in the given data.

FEATURE ENGINEERING:

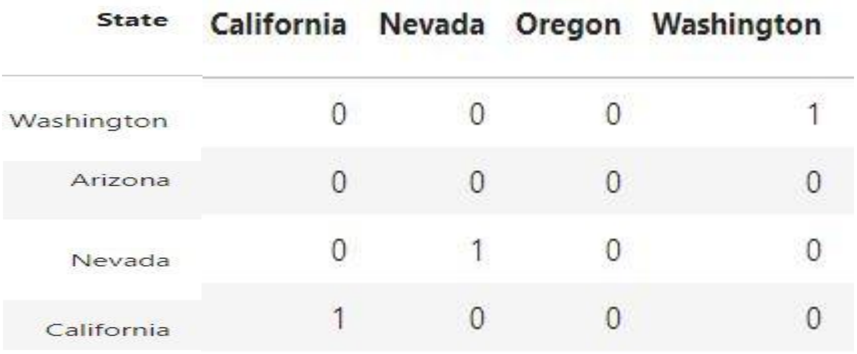
We altered the column to the total number of effective days for each customer. As **effective to date** is a relative term among customers, we took Jan 1, 2011, as a starting date and calculated the total number of effective days for each customer. **One-hot encoding** is one of the most common encoding methods in machine learning. This method spreads the values in a column to multiple flag columns and assigns **0** or **1** to them. These binary values express the relationship between grouped and encoded columns. Normalization (or **min-max normalization**) scale all values in a fixed range between **-1** and **1**. This transformation does not change the distribution of the feature and due to the decreased standard deviations, the effects of the **outliers** increases. Therefore, before normalization, it is recommended to handle the outliers.



	Effective To Date
0	2/24/2011
1	1/31/2011
2	2/19/2011
3	1/20/2011

	Effective To Date
0	55.0
1	31.0
2	50.0
3	20.0

	Effective To Date
0	1.485213
1	0.068602
2	1.190086
3	-0.580678



State	California	Nevada	Oregon	Washington
Washington	0	0	0	1
Arizona	0	0	0	0
Nevada	0	1	0	0
California	1	0	0	0

K means clustering:

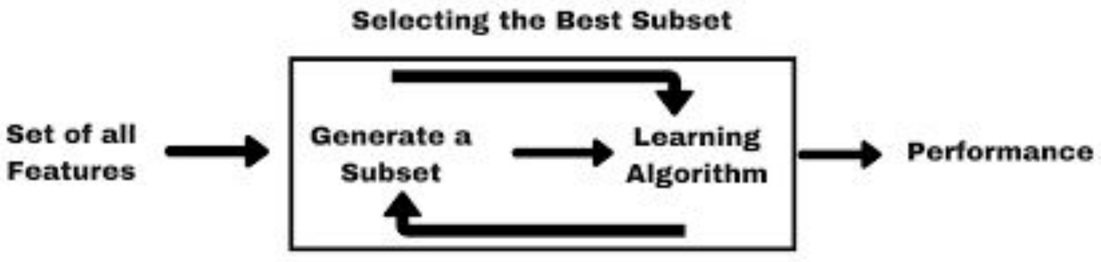
Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups.

Using the K-means clustering the optimal K value is found. Using this K value new feature is generated. Since the Elbow method gave a very soft curve ,this is why here the R2 score and MAPE is taken for that.

FEATURE SELECTION

Backward Feature Elimination

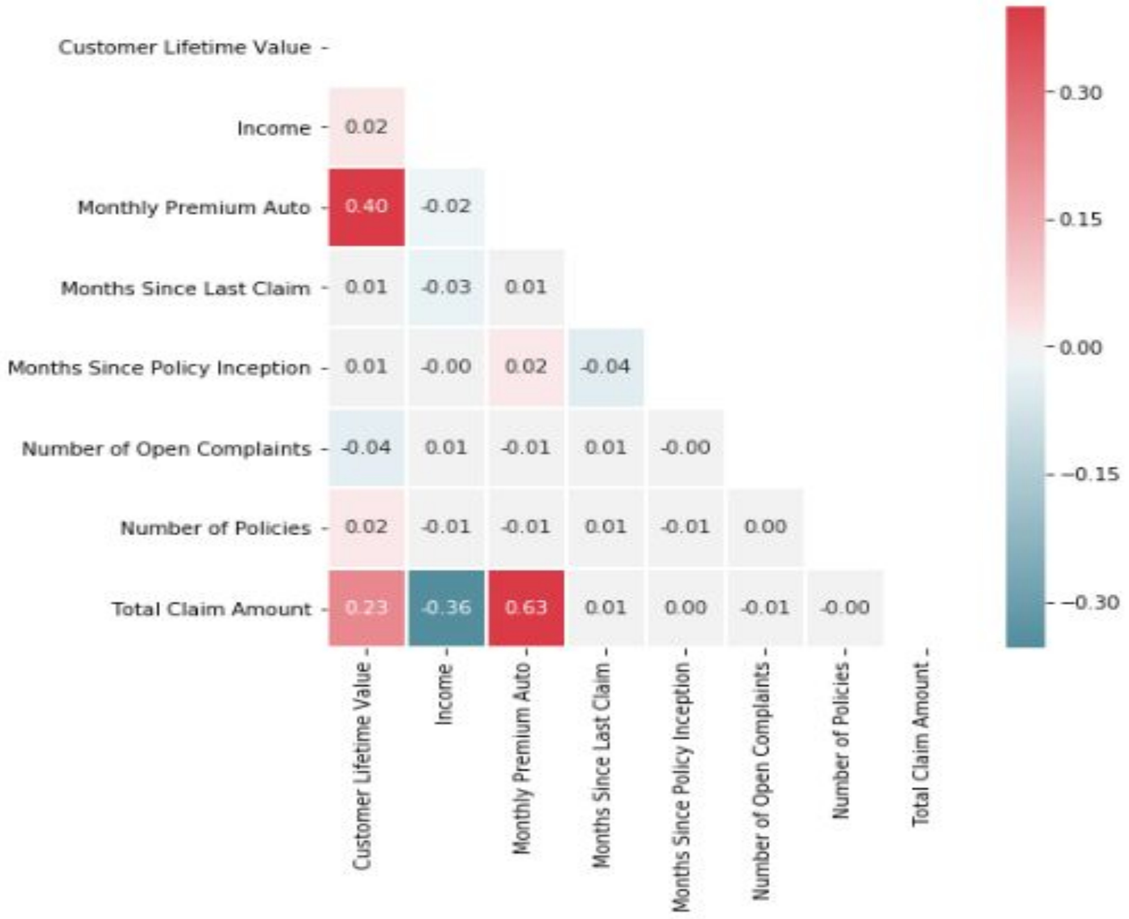
First,the model is trained with all features using training dataset,accuracy is noted as reference.Then features are dropped one at a time,model is trained and corresponding accuracy is noted.Now we examine the accuracy rates for each feature dropped and drop the feature which produce maximum increase in accuracy.The same procedure is repeated on the modified dataset until no increase in accuracy is observed.
Ideal implementation of feature selection runs in $O(2^n)$ but using BFE we can optimise it to $O(n^2)$.
For the given dataset,there was a large decrease in accuracy when feature 'no of policies was dropped' indicating that this is one of the most significant feature.Dropping all other features did not increase accuracy compared to referenced accuracy implying they are of equal relative importance.
We engineered many features and many of them got eliminated by “**Backward Feature Elimination**”



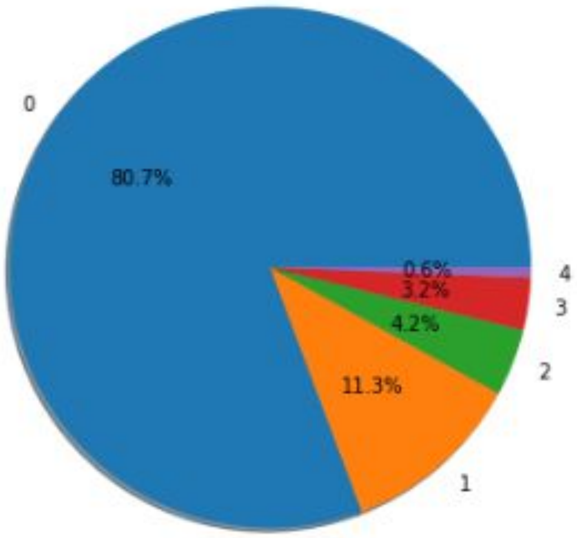
DATA VISUALISATION

As the age of Big Data kicks into high-gear, visualization is an increasingly key tool to make sense of the trillions of rows of data generated every day. It makes the massive datasets 'understandable' ,justifying its extensive usage in exploratory data analysis.
For this Report,Visualization had been done with 'Seaborn' and 'Matplotlib '-both python data visualization libraries of appealing aesthetics.Data visualization was used on the given dataset to get an idea of dependance of various features over 'Customer lifetime value'.This exposes unforeseen connections,relation between features and helps in modelling.

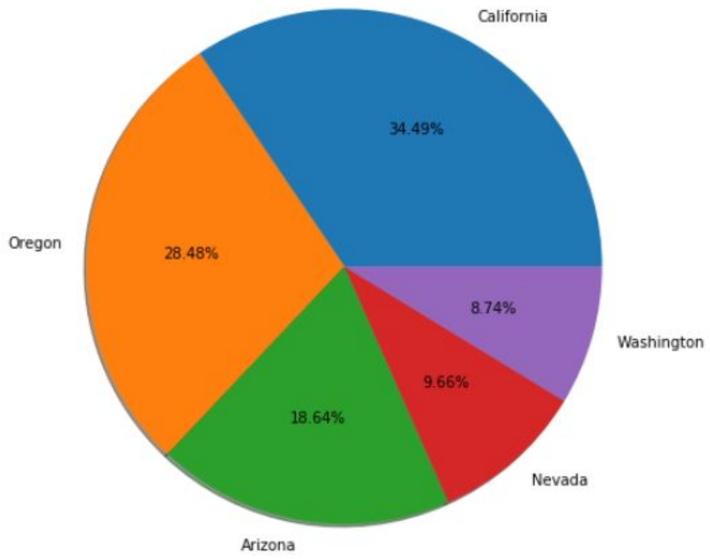
HEATMAP for continuous variables



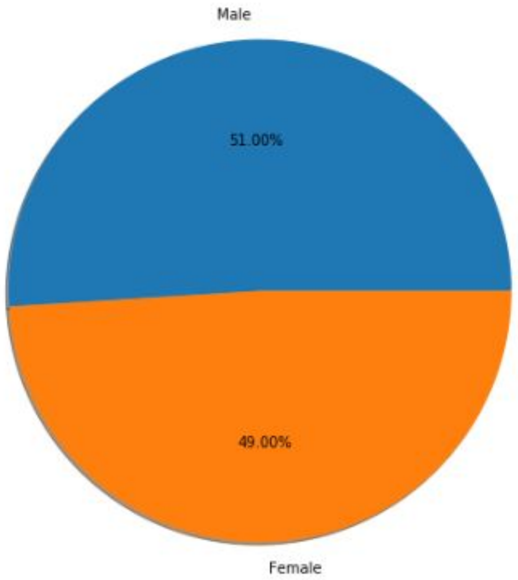
DATA DISTRIBUTION



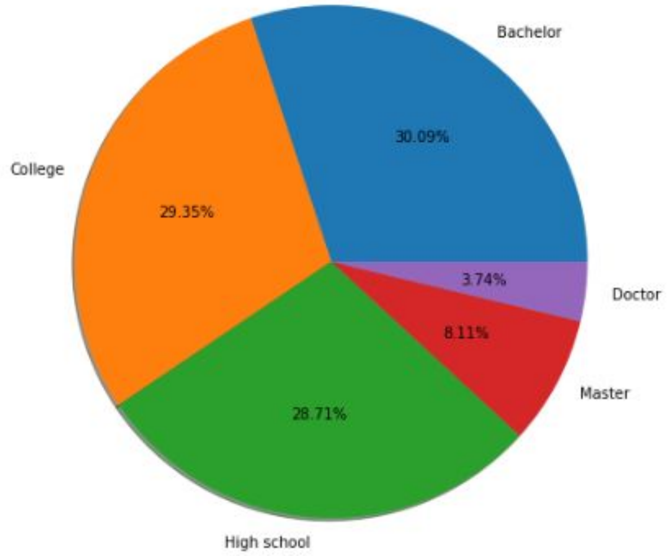
Number of policies



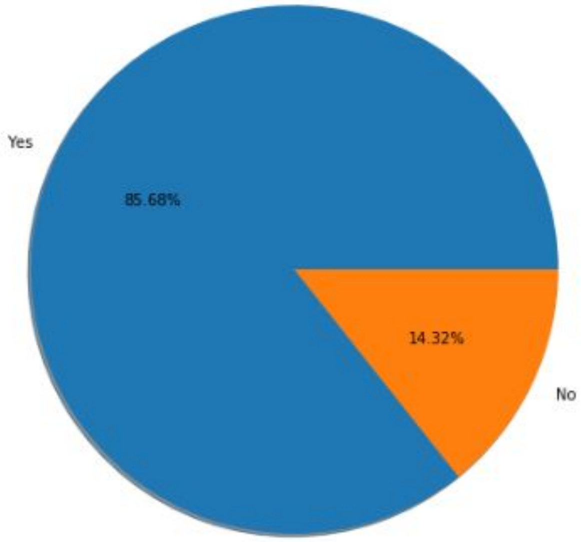
States



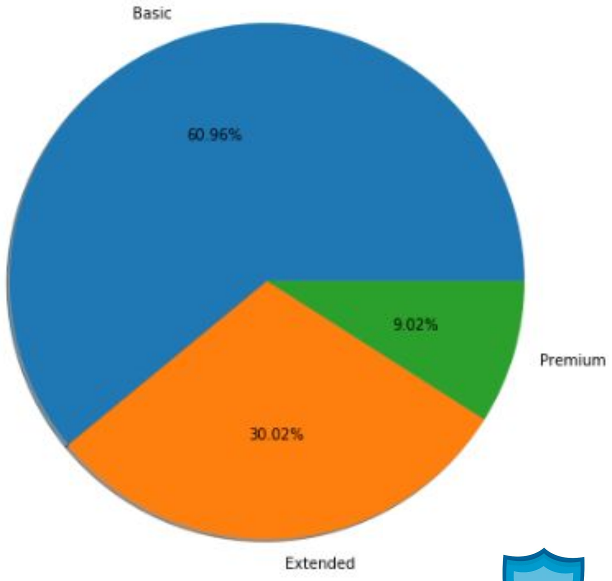
Gender



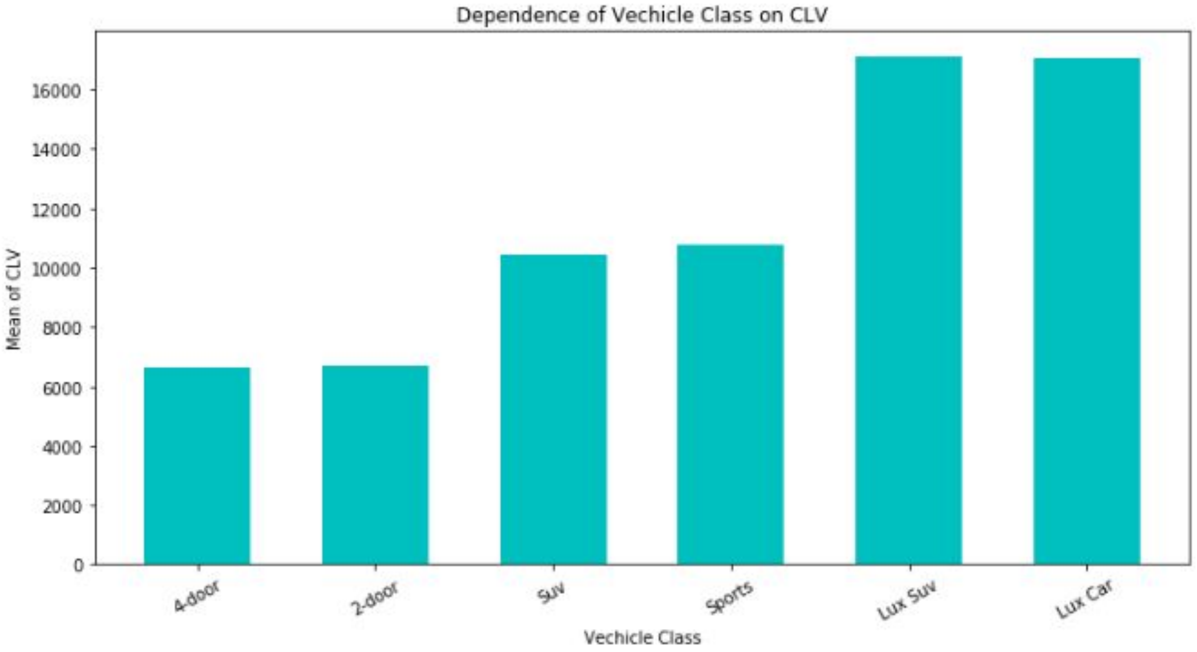
Education



Response

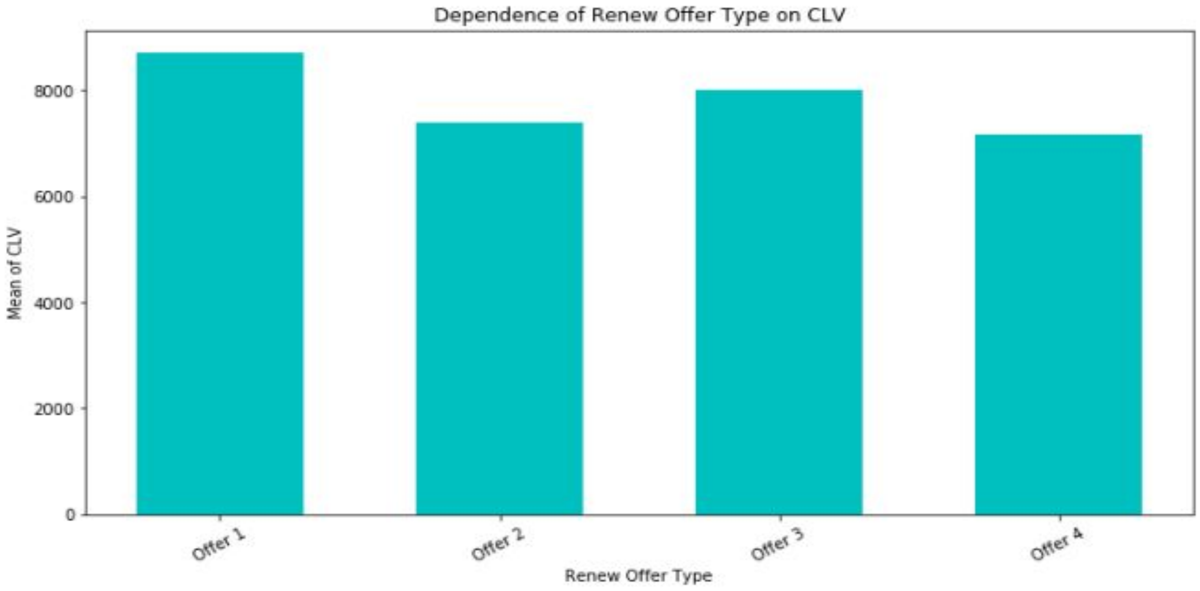


Coverage

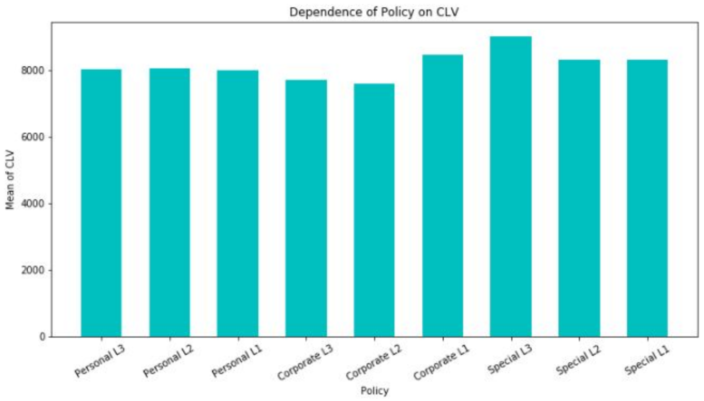


Vehicle Class	Mean Value
4- door	6631.72
2-door	6671.03
Suv	10443.51
Sports	10750.98
Lux Suv	17122.99
Lux Car	17053.34

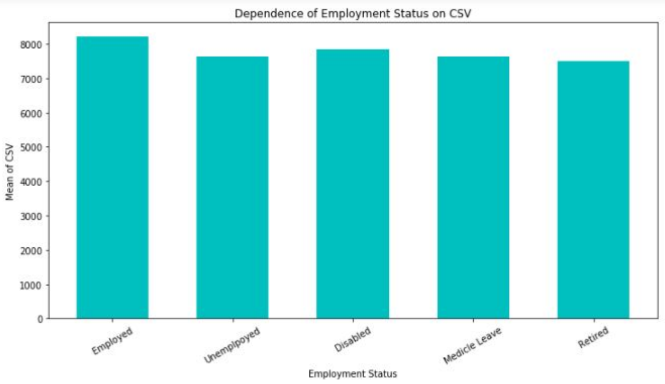
From the above graph it is obvious that, Customer with Luxury Suv or Luxury Car will generate more revenue to the Company.



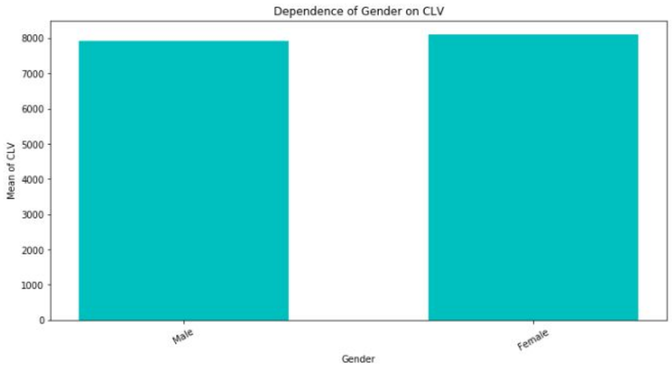
Customer who are availing Offer 1 & Offer 3 will generate more revenue to the company as compare to Offer 2 and Offer 4



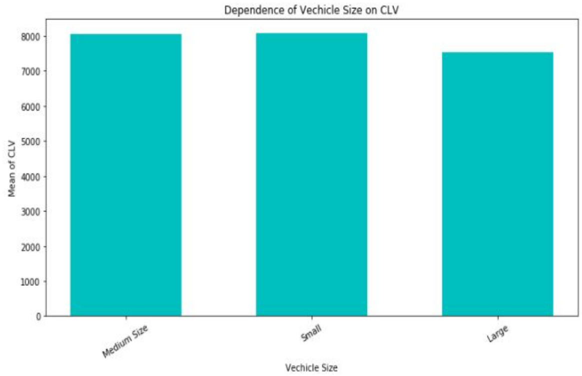
Customer with Special L1, Special L2, Special L3 and Corporate L1 Policy has more average Customer Lifetime value compare to rest policies



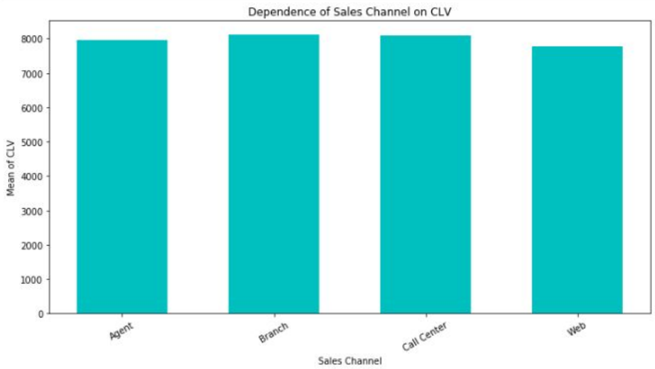
Graphs shows that Employed customers are generating more revenue to the company, and the next highest generators are the disabled ones



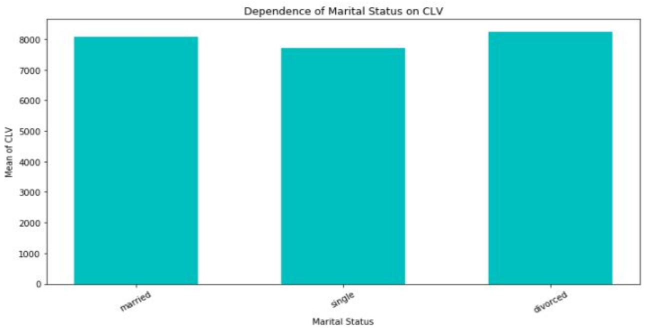
Female customers are generating a bit more revenue to the company than male customers



Customers with small and medium sized cars are generating equal and high revenue compared to the customers with large sized cars



Sales Channel is not making much difference in the Customer Lifetime Value, but the customers who approached the company through web generated less revenue than others

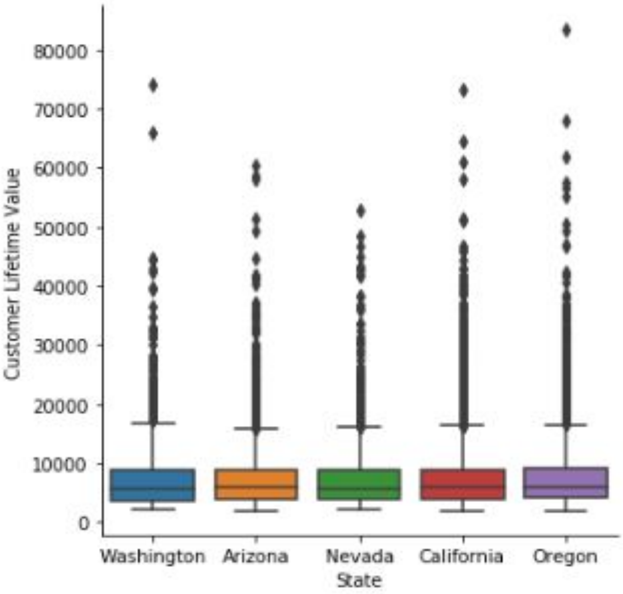


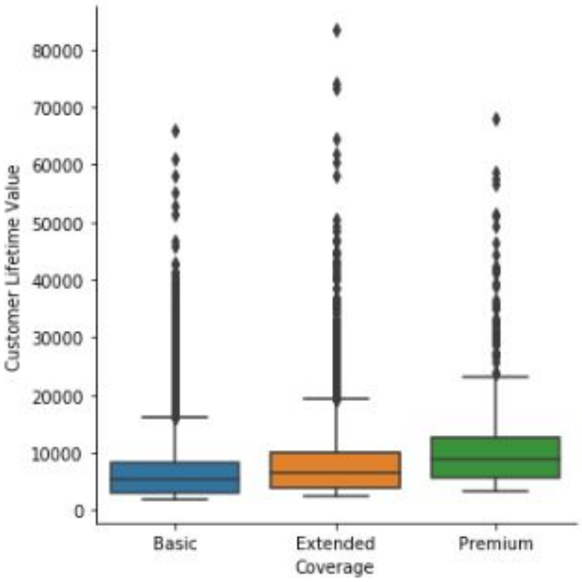
Divorced customers are spending more amount on the automobile insurance, company can get more revenue from the divorced customers,

OUTLIERS

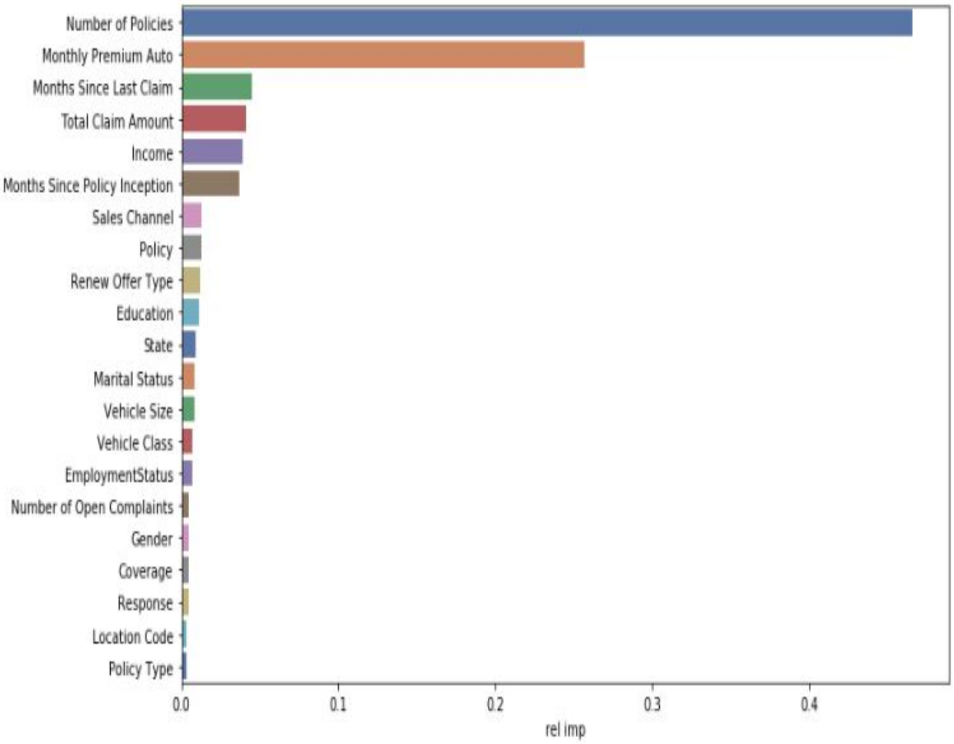
Detection of outliers can be solved by supervised learning algorithms if we have information on anomalous behavior before modeling, but initially without feedback its difficult to identify such points. So we model this as an **unsupervised** problem using algorithms like **Isolation Forest**, One class SVM and LSTM. Here we are identifying anomalies using **Isolation Forest**.

We divided the dataset into **training** and k-fold **cross validation** sets, and applied **Isolation Forest** on the training set and removed the outliers. We then trained our model using treenbased regressors but the error was similar to the original data (possibly since removing outliers also reduces the amount of data for training, thus reducing **accuracy**). We varied different parameters but no significant progress was made. So we decided NOT to remove outliers from our data.





FEATURE IMPORTANCE

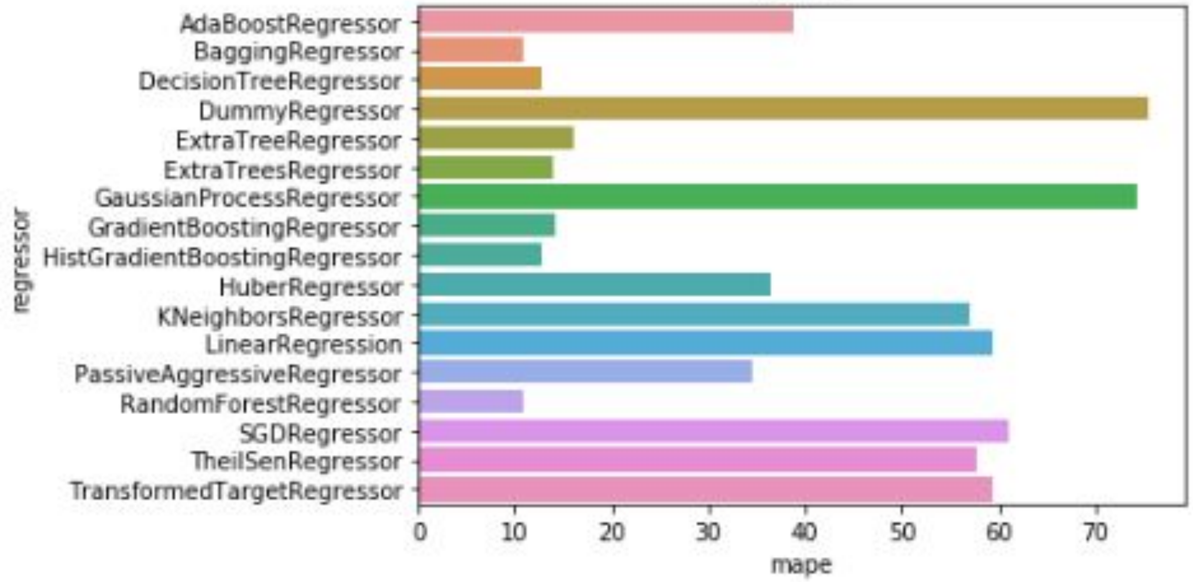


From the feature importance plot, we found out that the **number of policies** is the most important feature. According to the values in the number of policies column, the data set is splitted. And the regressor is trained and tested on each data set, but the error comes out to be nearly similar for all the data sets. In regression and multivariate analysis in which the relationships are of interest, we do normalization to reach a linear, more robust relationship. Commonly when the relationship between two datasets is non-linear we transform data to reach a linear relationship. So we normalized the dataset to get a more linear relationship.

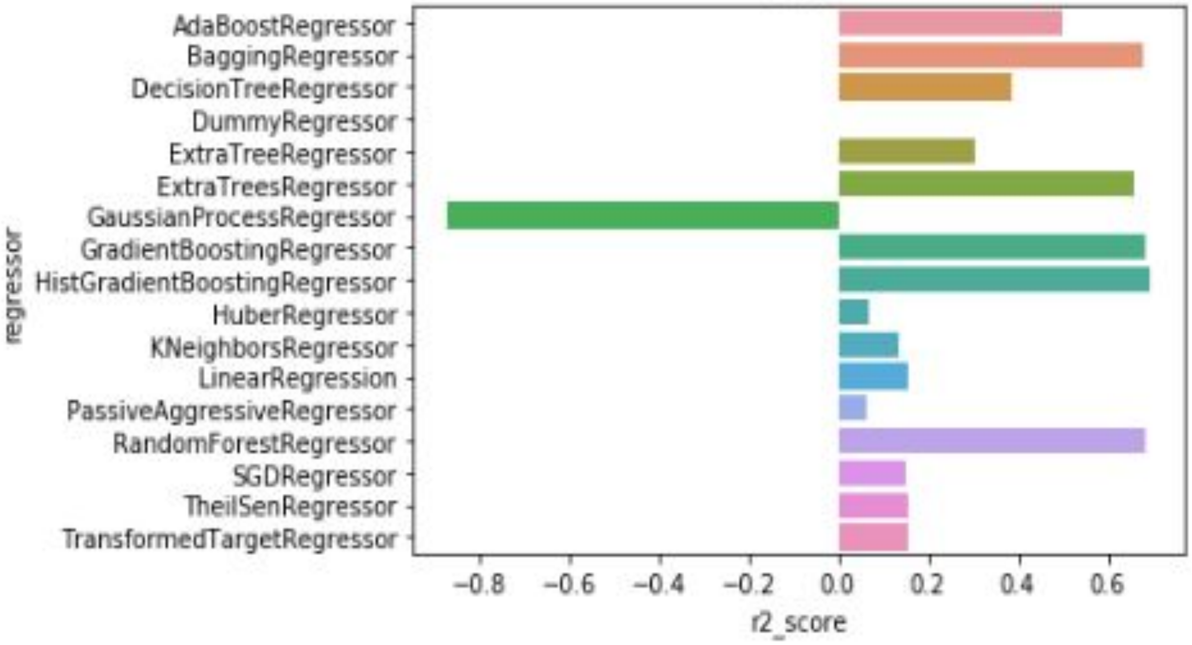
MODEL SELECTION

The scores are on the k-fold cross-validation set to ensure that the model doesn’t overfit and is efficient for new datasets also.

MAPE



R2 SCORE



NOTE:
Adjusted r2 score is nearly equal to r2 score because the number of data points are much higher than number of features.

MODEL DESIGNING

Stacking is an ensemble learning technique that uses predictions from multiple models (for example random forest, knn or svm) to build a new model. This model is used for making predictions on the test set. Below is a step-wise explanation for a simple stacked ensemble:

1) The train set is split into 10 parts.

2) A base model (suppose a RandomForestRegressor) is fitted on 9 parts and predictions are made for the 10th part. This is done for each part of the train set.

3)The base model (in this case, RandomForestRegressor) is then fitted on the whole train dataset.

4)Using this model, predictions are made on the test set.

5)Steps 2 to 4 are repeated for another base model (say knn) resulting in another set of predictions for the train set and test set.

6)The predictions from the train set are used as features to build a new model.

7)This model is used to make final predictions on the test prediction set.

Random Forest Regressor:

A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is always the same as the original input sample size but the samples are drawn with replacement if, bootstrap=True.
tuned parameters:

>Tuning in the first Randomforest:

bootstrap=True, criterion='mse', max_depth=None, max_features=0.7500000000000001, max_leaf_nodes=None,min_impurity_decrease=0.0, min_impurity_split=None,min_samples_leaf=11, min_samples_split=9,min_weight_fraction_leaf=0.0, n_estimators=100,n_jobs=-1, oob_score=False, random_state=None,n_jobs=-1, oob_score=False, random_state=None, verbose=0, warm_start=False

>Tuning in the second Randomforest:

bootstrap=True, max_features=0.8, min_samples_leaf=13, min_samples_split=19, n_estimators=100, n_jobs=-1

>Tuning in the third Randomforest

:bootstrap=True, max_features=0.45, min_samples_leaf=19, min_samples_split=16, n_estimators=100

Randomforest tuned 3 times,because one of them may be over fitted or one of them basically less fitted ,so we putting the data in the three different kind of tuned models.

Gradient boosting (Xgboost,Catboost,LGBMregressor):

All boosting models implemented and it may decrease the variance in the predictions.

>Tuning parameters: Default.

Extra-trees regressor:

This class implements a meta estimator that fits a number of randomized decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

>Tuning in the first Extra-trees regressor:
bootstrap=False, max_features=0.7500000000000001, min_samples_leaf=17, min_samples_split=17, n_estimators=1007

>Tuning in the second Extra-trees regressor:
bootstrap=False, max_features0.3 min_samples_leaf=3, min_samples_split=17, n_estimators=100

Here Extra Trees Regressor tuned two times for the same reason as defined in Random Forest Regressor

These models are implemented and now we going to use Votingressor .

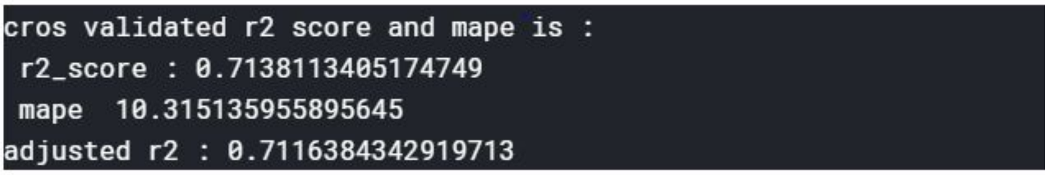
Voting regressor:

A voting regressor is an ensemble meta-estimator that fits base regressors each on the whole dataset. It, then, averages the individual predictions to form a final prediction.
base regressors :

The base regressors we are using here are-----

RandomForestRegressor(tuning three times) -> Noted as lg, lg5,lg7
ExtraTreesRegressor(tuning two times) -> Noted as lg4, lg6,
the averages will give us the accuracy.

RESULTS:

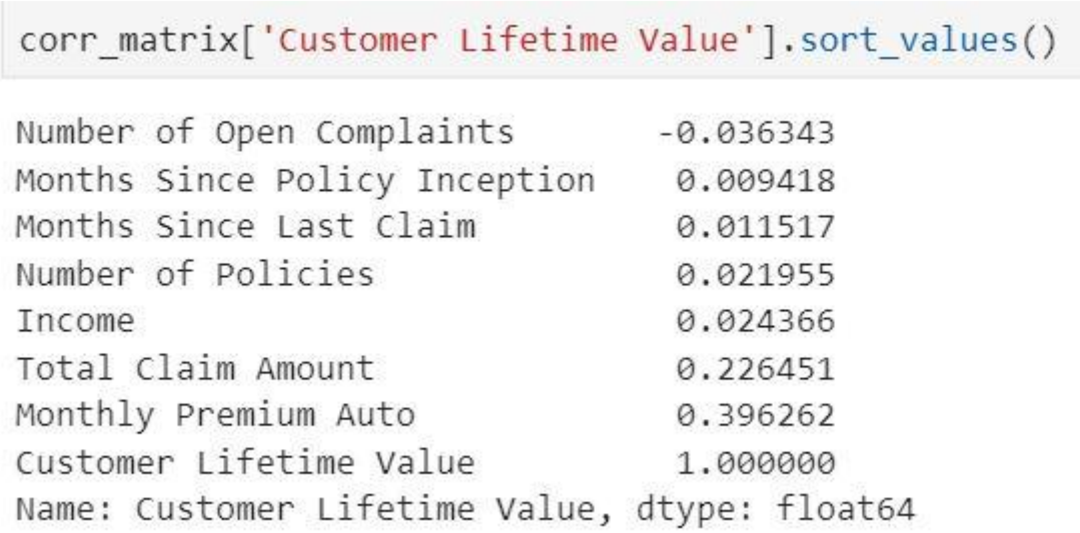


```
cros validated r2 score and mape is :
 r2_score  : 0.7138113405174749
 mape   10.315135955895645
adjusted r2  : 0.7116384342919713
```

Our K-fold cross validation score are :
MAPE : ~10% (< 11%)
R2 score : > 0.71
Adjusted R2 score : > 0.71

BUSINESS STRATEGY

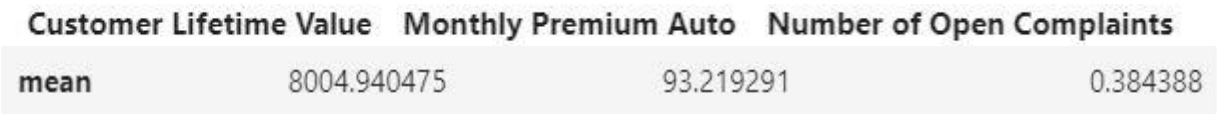
Best Attributes of the customer to get highest revenue for the Insurance firm. For sorting out this problem we used correlation matrix in our trained model.



```
corr_matrix['Customer Lifetime Value'].sort_values()
```

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

The customers with less Number of Open Complaints and more Monthly Premium Auto will generate highest revenue to the Auto Insurance company.
“Months Since Policy Inception” , “Months Since Last Claim” are not really much correlated to Customer Lifetime Value compare to the other attributes.



	Customer Lifetime Value	Monthly Premium Auto	Number of Open Complaints
mean	8004.940475	93.219291	0.384388

On an average, customer with $100 of **Monthly Premium** Auto will generate revenue upto $8000 to the company and “Monthly Premium Auto” is directly proportional to the “Customer Lifetime Value” with positive constant of 0.3962 and “Number of Open Complaints” with negative constant of -0.036. So the both attributes are highly important for calculating the “Customer Lifetime Value”.

If a person owns a luxury car or suv, then company can generate more revenue. If not, person who is employed and availing offer1 and spacial L3 policy, client will get more revenue.
To see the graph go to page 8 and 9.(click here)