# CUSTOMER LIFETIME VALUE PREDICTION

## (AUTO INSURANCE INDUSTRY)

BY

TEAM ALPHA SILOS

# CONTENTS

# 1. PROBLEM STATEMENT

To predict the customer lifetime value for a given customer for an automobile insurance industry, given a set of quantitative and qualitative features. The task would be to develop an analytical and statistically sound framework to predict the values of the target variable.

To identify the key characteristics of the customers with good customer lifetime value.

## 1.1 What is Customer Lifetime Value?

Customer Lifetime value is the overall present value of a customer from the perspective of a company, considering the revenue generated from the customer and the expenditure done by the company for that customer. It is a parameter which help us to judge the value of customer profitability.

## 1.2 Introduction to the Auto Insurance Industry.

Automobile insurance is a policy purchased by vehicle owners to mitigate costs associated with getting into an auto accident. Instead of paying out of pocket for auto accidents, people pay annual premiums to an auto insurance company. The company then pays all or most of the costs associated with an auto accident or other vehicle damage, depending on the contract which both company and vehicle owner have agreed upon.

Auto insurance premiums vary depending on age, gender, years of driving experience, accident and moving violation history, and other factors. Most states mandate that all vehicle owners purchase a minimum amount of auto insurance.

In exchange for paying a premium, the insurance company agrees to pay your losses as outlined in your policy. Coverage includes:

**Property**: Damage and theft of your car.
**Liability**: Legal Responsibility to others for bodily injury or property damage.
**Medical**: Costs of treating injuries, rehabilitation and sometimes lost wages and funeral expenses.

Policy terms are usually six- or 12 month time frames and are renewable. An insurer will notify a customer when it’s time to renew the policy and pay another premium.

### 1.3 Relevance of Customer Lifetime Value in Business perspective

Customer Lifetime Value is one of the very important parameters which determine a profitability from a given customer. Knowing the same can help companies to:

- Develop strategies to acquire more customers and retain the existing one.
- Design the policies for particular customers based on their customer lifetime value.
- Decide which key characteristics of good customers to identify more profitable customers with respect to the company.
- Identify the optimal monthly premium for specific customers according to his/her characteristics.
- Determination of the amount to be covered, for a specific person while maximizing the profit margins for the company.

# 2. UNDERSTANDING OF THE DATASET:

## 2.1 Dependent Variable:

**Customer Lifetime Value (in Dollars):** It is a continuous variable, which refers to the overall value of a customer to a business.

## 2.2 Independent Variable:

**Qualitative features**

- **State**: The state feature may affect the customer lifetime value, as some states may have more accident rates, more lenient traffic rules which may lead to larger amount of claims hence less Customer Lifetime Value.
- **Coverage:** {Basic, Extended, Premium}: Class of coverage amount to be covered under the current policy with basic referring to the class with minimum coverage and premium denoting the maximum coverage.
- **Education:** Education of the driver of a particular automobile. It has a total of 5 classes, ranging from high school to PhD.
- **Gender:** The gender of the driver, may statistically affect the customer lifetime value.
- **Employment status:** Employment Status may help in judging the reliability of a customer.
- **Marital status:** [Single, Married, Divorced] Person behaviour judgement parameter, which may or may not affect the Customer Lifetime Value.
- **Policy Type:** [Corporate Auto, Personal Auto]. The type of policy largely decides the amount of premium to be paid, and the claim rates which we can expect.

- **Policy:** L1, L2, L3
- **Renew Offer Type:** Offer 1, 2, 3
- **Sales Channel:**
- **Vehicle Class:** Policy type, premium,
- **Vehicle Size:** Small, Med, Large

**Quantitative Features**

- **Income:** Annual Income of a customer. (in dollars)
- **Monthly Premium Auto:** Monthly Premium to be paid by the customer, in USD ($)
- **Total Claim Amount:** Total amount claimed by the cus
- **Effective Date :** The date to which the policy is valid.
- **Months Since Last Claim:** Total number of months since Last claim.
- **Months Since Policy Inception:** Number of months since the starting of the Policy.
- **No. of Open Complaints:** Number of complaints whose decision is still pending.
- **No. of Policies:** Number of policies taken by a particular customer

# 3. PREPROCESSING STEPS/ DATA CLEANING

Data cleaning and preprocessing is an essential step before building any hypotheses and testing different models.

- No missing values were found in the dataset.
- Since most of the features were categorical smoothing noisy data or recognizing outliers is not of much significance.
- Categorical string data cannot be directly fed into machine learning models and hence a one-hot conversion is used.
- The numerical features are normalized.
- Features like date/time etc are converted into numerical data like number of days.
- Columns like feature id, date were dropped as they don't contribute to the predictions.
- The dataset is split into training and validation set.

# 4. FEATURE SELECTION AND GENERATION

## 4.1 Intuition through Visualizations.

We plotted different graphs in order to get an intuition of factors which are important for predicting the values of Customer Lifetime Value. Some of the important graphs have been shown here, and for a more detailed understanding do refer Appendix.

4.1.1 HISTOGRAMS:

To get a know how of effects different categorical variables have on the overall customer lifetime value, we have plotted histograms for different instance of a particular feature wrt to Customer Lifetime Value. Some of the important histograms are as follows:

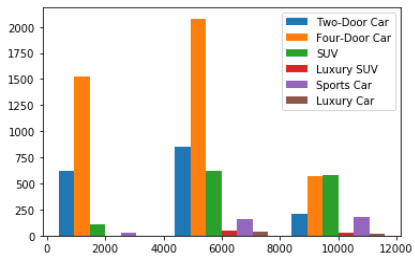



Fig 1

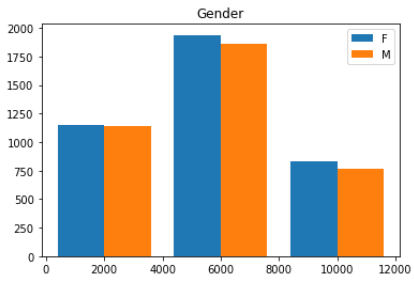



Fig 2

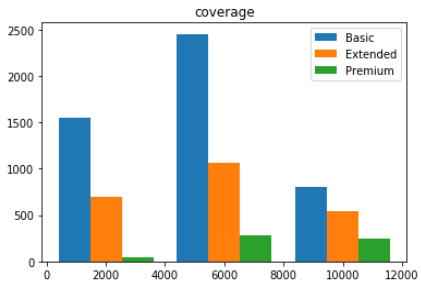



Fig 3

The graph shows the histograms for CLV, for specific types in a particular categorical variable.

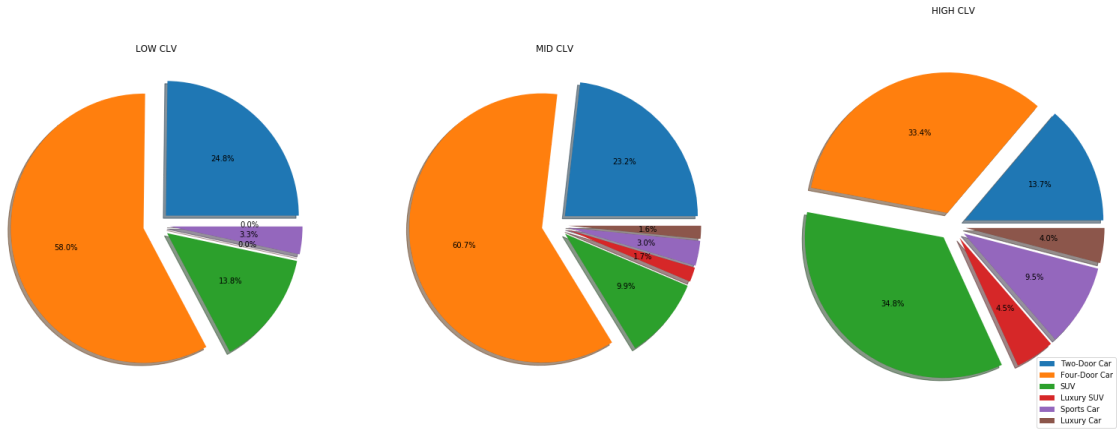
**Figure 1**: It is clear from the figure one, that the Luxury and sports car are more prevelant at the upper cap of CLV, then lower ones, hence is useful.

**Figure 2**: For the graphs like figure 2, it is clear that there is no effect of this particular variable on CLV.

**Figure 3:** The third type of graph clearly shows that there is a correlation, as the length of green histogram is increasing with increasing CLV.

4.1.2 PIE CHARTS

We also plotted some Pie Charts in three bins of CLV. i.e for CLV value less than a particular threshold, percentage of people belonging to a particular type of that categorical variable is one fragment of the pie chart. Some plots are as follows:



The whole dataset is divided into three bins, and for each bin we have plotted pie charts to check the distribution particular category type. For eg, the above fig shows that the percentage of  Sports Car, as well as SUVs are higher on higher CLV pie chart than in lower CLV pie chart. Similarly after the analysis of the graphs produced as shown in the appendix A.2, we got an intuitive idea that certain factors are better than others.

## 4.2 Feature Generation:

We have generated two new features, from the existing features, to improve its performance.

### 4.2.1 Value of Customer till date

The feature is represented as ‘present_value’ in the source code.

$$present\ value\ =\ [P * m\ -\ T]$$

Where,
P: Monthly premium in Dollars | m: Months Since Policy Inception | T: Total Claim Amount
The amount signifies, the value of the customer to the company at present intuitively.

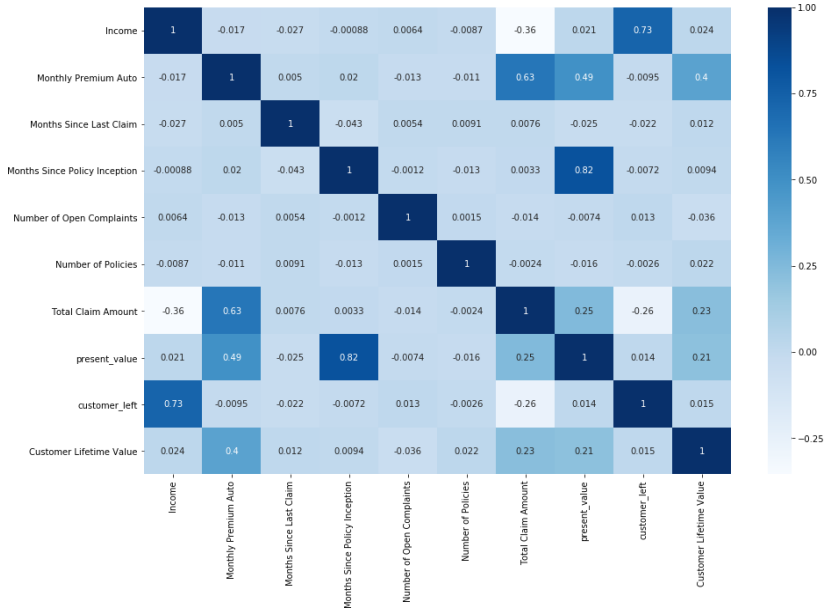
### 4.2.2 Time left for the policy to expire

The feature is represented as ‘customer_left’ in the source_code.

$$customer\ left\ =\ Income\ -\ (Monthly\ Premium\ Auto) * (Number\ of\ months\ leftt)$$

The feature signifies, the savings of the customer, after paying the premium every month, which gives the sense of how stable a customer is financially.

## 4.3 Feature selection using Pearson Correlation Matrix:



The pearson correlation matrix shows that the features generated have high correlation with the customer lifetime value, which supports the addition of generated features in the existing feature set.

According to our analysis of the above plots, we came to the conclusion that the following features are important:

A) Monthly Premium Auto
B) No. of Policies.
C) Present Value
D) Total Claim Amount
E) Vehicle Class
F) Income.

# 5. MODEL SELECTION AND HYPERPARAMETER TUNING

## 5.1 Customer Lifetime Value Prediction:

**Linear Regression**

- To start with, it is always advisable to use a simple model and thus a linear regression model is first used to make predictions.
- In statistics, **linear regression** is a **linear** approach to modeling the relationship between a scalar response (or dependent variable), **customer lifetime value** in our case, and one or more explanatory variables (or independent variables).
- High error on training as well as validation set are observed which implies the model is underfitting the data and the complexities of the data can't be captured by a linear regression model. **$R^2$ score obtained = 0.2**
- Therefore there is a need to increase the model complexity so polynomial terms are added and a **polynomial regression** model is trained. Even then the results obtained are poor.
- The number of variables increases exponentially in a polynomial regression model and the problem of overfitting is always a concern in such a case
- To train a model with a very high degree the data is insufficient and hence polynomial regression also performs poorly on both training and validation set.

**Bayesian Linear Regression**

- In the Bayesian viewpoint, we formulate linear regression using probability distributions rather than point estimates. The response, y, is not estimated as a single value, but is assumed to be drawn from a probability distribution.
- The output, y is generated from a normal (Gaussian) Distribution characterized by a mean and variance.
- The aim of Bayesian Linear Regression is not to find the single “best” value of the model parameters, but rather to determine the posterior distribution for the model parameters.
- Not only is the response generated from a probability distribution, but the model parameters are assumed to come from a distribution as well.
- Since the point estimates in linear regression are themselves very poor, predicting a distribution for the target variable doesn’t improve the performance of the model substantially and the results are still very poor.

**Few observations from the models trained so far:**

- Models which are trying to predict the target variable using some linear or non-linear combinations of the predictor variables are performing very poorly.
- Since most of the features are categorical in nature, a hypothesis is developed that models which sequentially check the categorical variables might be able to provide good results and hence the next model we choose to test is **Decision Tree**.

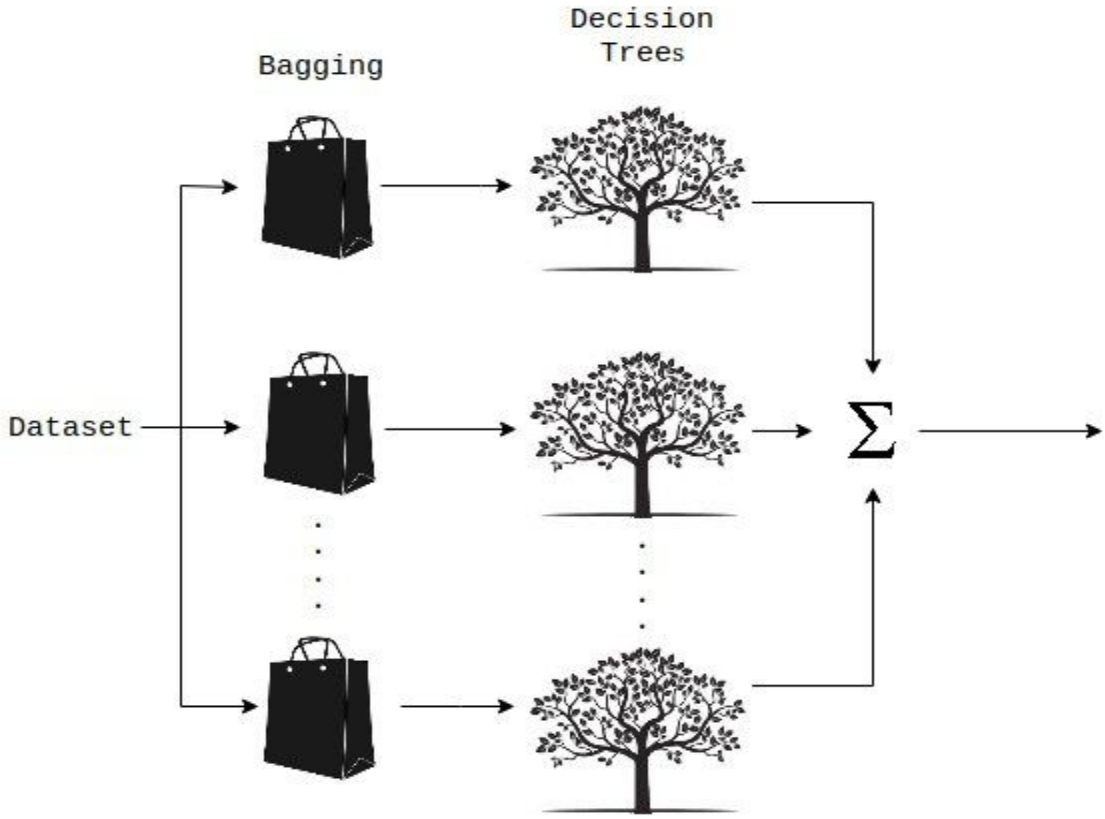
**Decision Trees**

- Trees answer sequential questions which send us down a certain route of the tree given the answer. The model behaves with “if this then that” conditions ultimately yielding a specific result.
- Tree depth is an important concept. This represents how many questions are asked before we reach our predicted classification.
- **Advantages to using decision trees**:
    - **Easy to interpret** and make for straightforward visualizations.
    - The internal workings are capable of being observed and thus make it possible to reproduce work.
    - **Can handle both numerical and categorical data.**
    - Perform well on large datasets
    - Are extremely fast
- **Disadvantages of decision trees**:
    - Building decision trees require algorithms capable of determining an optimal choice at each node. One popular algorithm is the Hunt’s algorithm. This is a greedy model, meaning it makes the most optimal decision at each step, but does not take into account the global optimum.

    - Decision trees are prone to overfitting, especially when a tree is particularly deep. This is due to the amount of specificity we look at leading to smaller sample of events that meet the previous assumptions.
    - This small sample could lead to unsound conclusions. Ideally, we would like to minimize both errors due to bias and error due to variance.
- As we expected decision trees performed very well on the training set and yielded $R^2$ score of more than 0.95.
- The score on validation set was still not satisfactory as the model was overfitting the training data.
- Regularization techniques like tree pruning, stricter splitting criterion were used but the validation scores did not improve much. Another approach to tackle this problem is to use ensemble learning and hence the next model we choose to test is Random Forest.

**Random Forest**

- A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called **Bootstrap Aggregation**, commonly known as **bagging**.
- Bagging, in the Random Forest method, involves training each decision tree on a different data sample where sampling is done with replacement.




- The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.
- The low correlation between models is the key. Just like how investments with low correlations (like stocks and bonds) come together to form a portfolio that is greater

than the sum of its parts, uncorrelated models can produce ensemble predictions that are more accurate than any of the individual predictions.

- The reason for this wonderful effect is that the trees protect each other from their individual errors (as long as they don't constantly all err in the same direction). While some trees may be wrong, many other trees will be right, so as a group the trees are able to move in the correct direction.
- **Feature Randomness —** In a normal decision tree, when it is time to split a node, we consider every possible feature and pick the one that produces the most separation between the observations in the left node vs. those in the right node.
- In contrast, each tree in a random forest can pick only from a random subset of features. This forces even more variation amongst the trees in the model and ultimately results in lower correlation across trees and more diversification.
- Requirements for Random Forests to make accurate predictions:
    - We need features that have at least some predictive power. After all, if we put garbage in then we will get garbage out.
    - The trees of the forest and more importantly their predictions need to be uncorrelated (or at least have low correlations with each other).
- Thus Random Forests are a strong modeling technique and much more robust than a single decision tree.
- **Results:**
    - On training set, the model achieved very high $R^2$ scores ranging from 0.95-0.99
    - On validation set, the $R^2$ score came out to be nearly 0.67 - 0.69 which after hyperparameter tuning increased to 0.71 - 0.72.

## 5.2 Categorizing Customers

The goal of this task is to find out the types of customers that would generally give more revenue to the company.

### Approach:

- We used **K-means Clustering** algorithm to cluster the dataset based upon CLV.
- The optimal number of clusters was found out to be 3.
- Now each cluster was assigned a rating. Cluster with highest CLV get labelled as Rating 1 customers and similarly other clusters are labelled.
- Now a **classification model** is trained using all the features to predict the class of each customer which gives an **accuracy of 95.5 percent.**
- The most important features required for this classification are extracted which form the basis of defining and segregating customers which will yield high revenue to the company.

### 5.3 Hyperparameter Tuning and Cross-Validation:

- Random-search is used to tune the hyperparameters and an improvement is obtained in the $R^2$ score.
- Hyperparameters tuned:
    - N_estimators
    - Max_features
    - Max_depth
    - Min_sample_split
    - Min_sample_leaf
    - Bootstrap
- **K-fold Cross-Validation** is used to check generalization power of the model.
- Both the tasks are done together to find out hyperparameters which yield good results across different validation sets.
- These are the best set of parameters obtained:
    - 'n_estimators': 311,
    - 'min_samples_split': 2,
    - 'min_samples_leaf': 2,
    - 'max_features': 'auto',
    - 'max_depth': 76,
    - 'bootstrap': True

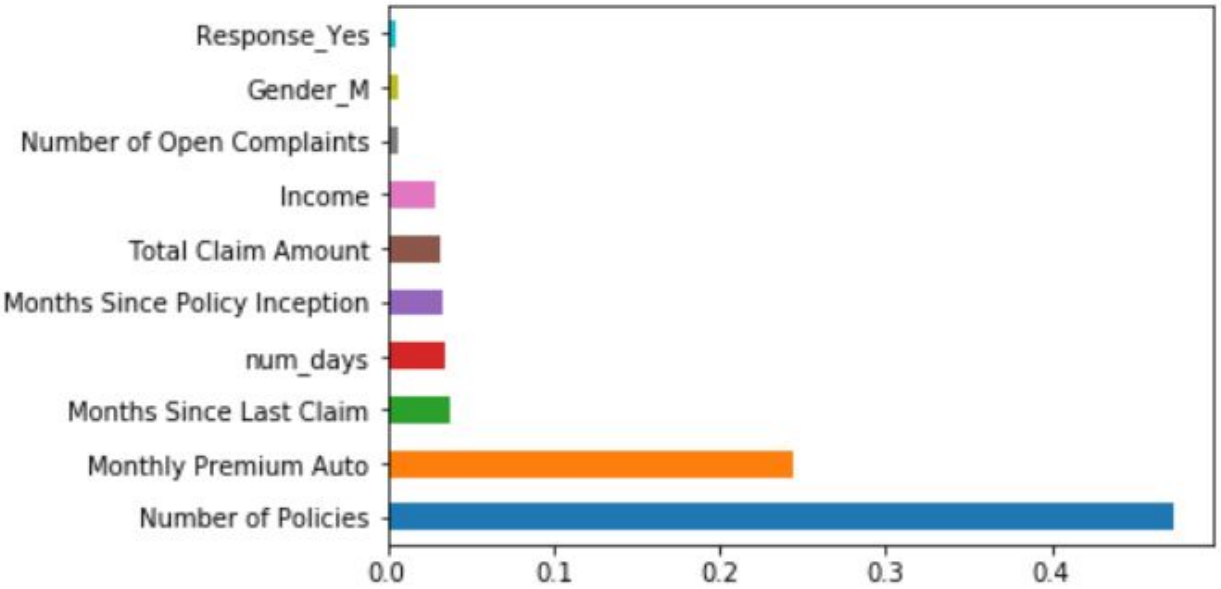
# 6. CONCLUSIONS AND RESULTS

## 6.1 Model Comparison

- Linear regression, polynomial regression and Bayesian Regression did not yield very good results.
- Decision trees captured the complexity of the data well but overfit the training data.
- Random Forests provided the best results.

## 6.2 Significant Features

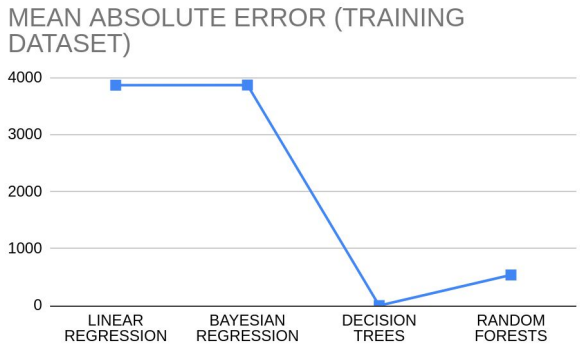
- Random Forests are selected as the final and best model
- After training the final model the most important features for regression are extracted and plotted here:

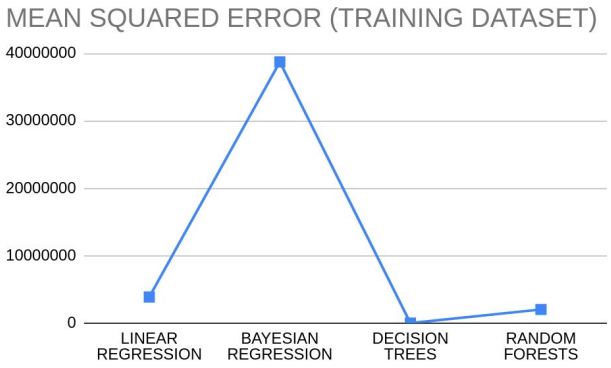


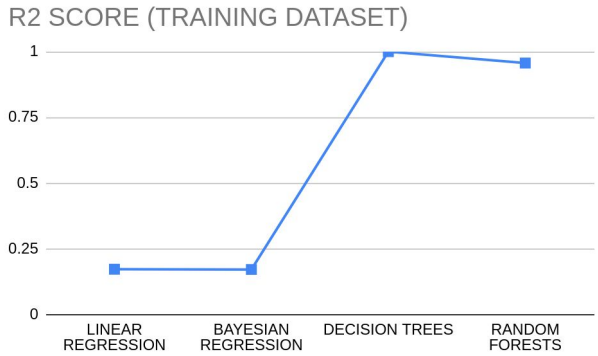

- The most important features come out to be Number of Policies and Monthly Premium which matches with our initial hypothesis.
    - The number of policies a particular customer has and the monthly premium he pays affects strongly the revenue generated by the company from that customer.
    - Customers with high premiums will also be the ones with potentially higher risks of car accidents than the other customers but on an average will give the highest revenue among all sections of customers.

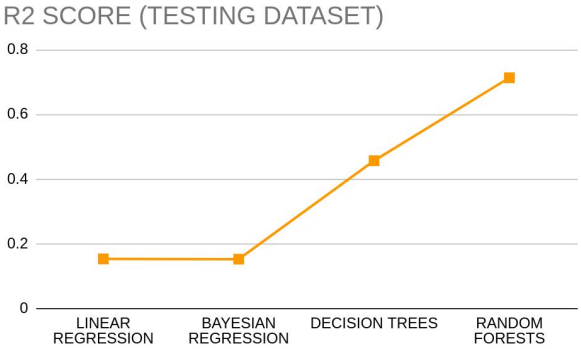
- Features like Months since Last Claim, Months Since Policy Inception, Total Claim Amount, and Income also play a significant role in determining Customer Lifetime Value.
    - Months since Last Claim give an idea about the frequency with which accidents occur and also the probability of recurrence.
    - Months Since Policy Inception tell about the amount of time the Customer is paying premiums hence an important feature.
    - Total Claim Amount is also important as it gives an idea of the severity of the accidents which occurred in the past and gives an indication of the future too.
    - Income definitely plays a significant role as the type of vehicle, type of policy, the amount of premium all somehow depend on the customer's income and hence is necessary for assessing the Customer Lifetime Value.

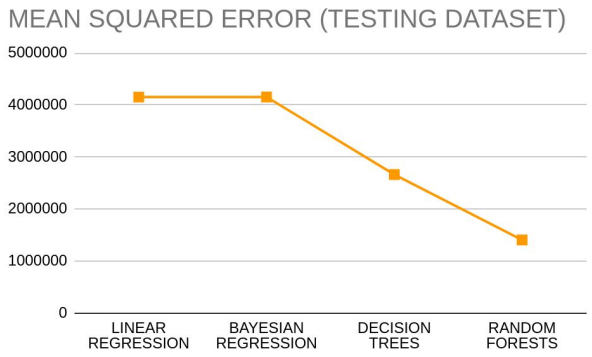
FINAL RESULT PLOTS

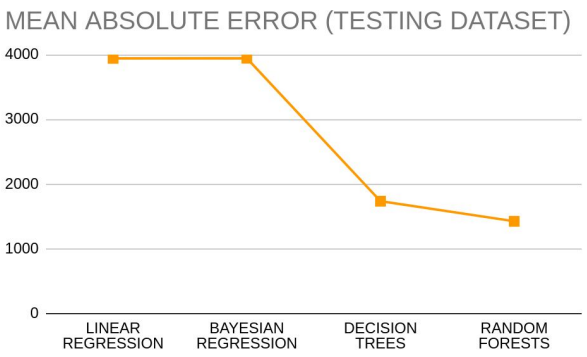



## 6.3 CUSTOMER TYPES

- Mean Values for Important features for Highest Rated Customers
  - Monthly Premium Auto: **112.5**
  - Number of Policies: **3.16**
  - Income: **39534**
  - Customer-left: **36853**
  - Months Since Last Claim: **14.9**

# 7. REFERENCES


- https://machinelearningmastery.com/feature-importance-and-feature-selection-with-xgboost-in-python/
- https://www.shanelynn.ie/summarising-aggregation-and-grouping-data-in-python-pandas/
- https://www.kaggle.com/emanueleamcappella/random-forest-hyperparameters-tuning
- https://scikit-learn.org/stable/modules/grid_search.html#multimetric-grid-search
- https://machinelearningmastery.com/develop-first-xgboost-model-python-scikit-learn/
- https://www.kaggle.com/rozester/xgboost-example-python
- https://towardsdatascience.com/better-heatmaps-and-correlation-matrix-plots-in-python-41445d0f2bec

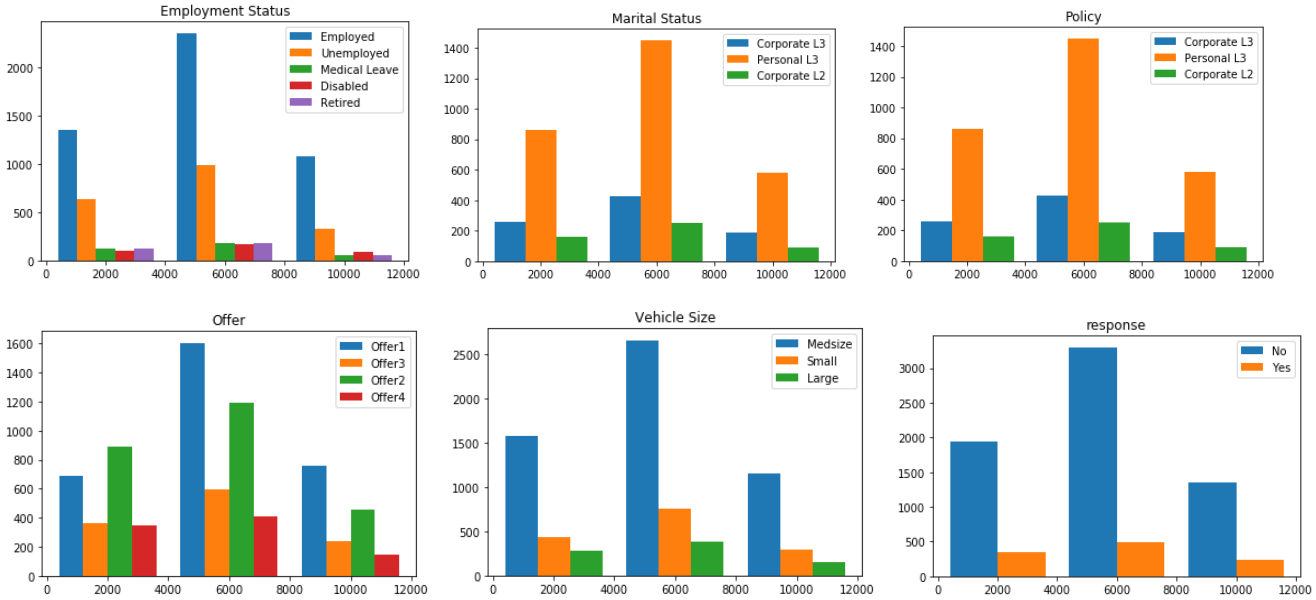
# APPENDIX - A.1

## Performance metrics for models

| Models | Mean Absolute Error | Mean Squared Error | $R^2$ Score |
|---|---|---|---|
| **Linear Regression** | **Training:**3867.0407<br>**Validation:**3941.67 | **Training:**38693850<br>**Validation:**4152088 | **Training:**0.1723609<br>**Validation:**0.15290 |
| **Bayesian Regression** | **Training:**3870.5741<br>**Validation:**3945.05 | **Training:**38734579<br>**Validation:**4154771 | **Training:**0.1714897<br>**Validation:**0.15235 |
| **Decision Trees** | **Training:**2.651e-14<br>**Validation:**1734.58 | **Training:**4.097e-26<br>**Validation:**2663461 | **Training:**1.0<br>**Validation:**0.45660 |
| **Random Forests** | **Training:**538.2125<br>**Validation:**1424.85 | **Training:**2014414.9<br>**Validation:**1405831 | **Training:**0.9569128<br>**Validation:**0.71318 |

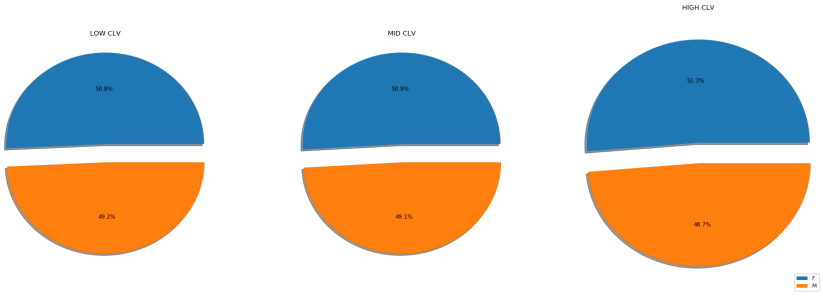
# APPENDIX - A.2

## HISTOGRAMS

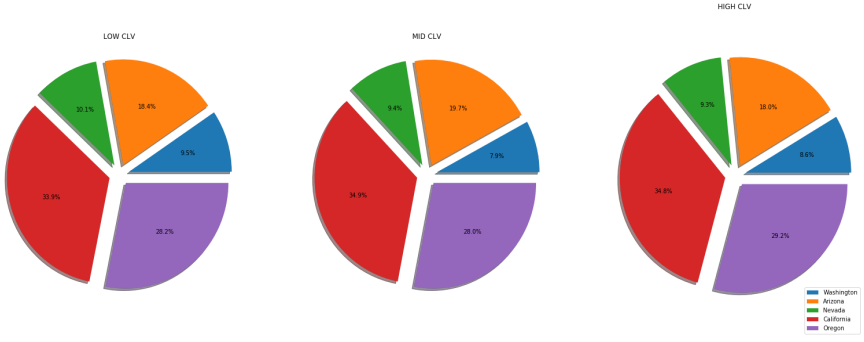



## PIE CHARTS

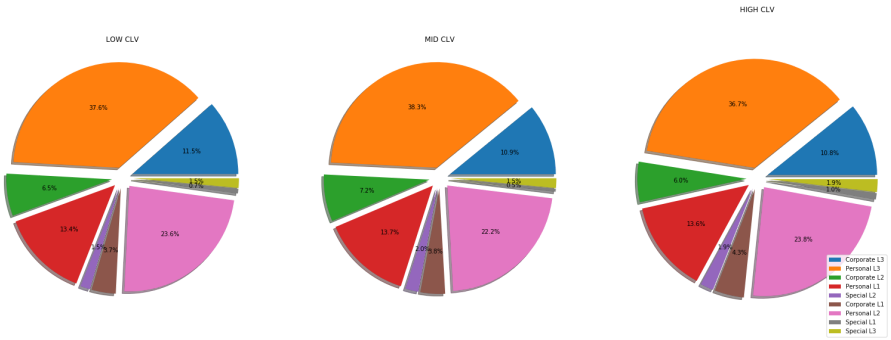
### a) GENDER



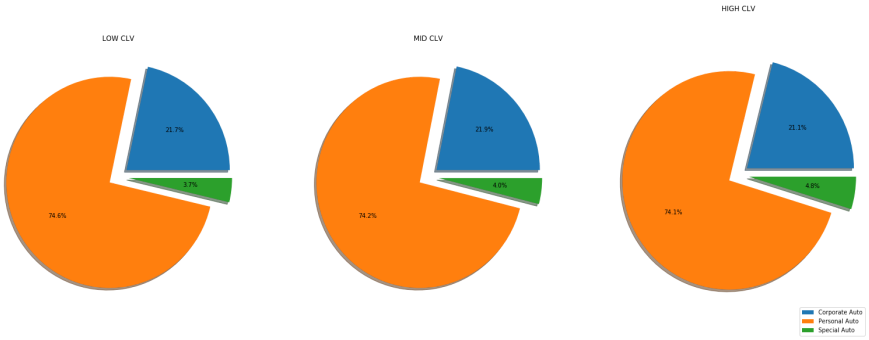

### b) STATE



c) POLICY




d) POLICY TYPE




PEARSON CORRELATION MATRIX OF ALL THE VARIABLES