Sunil Kamkar Sheshagiri
Pranav Cheruku
Vivian Wang
Muskan Agarwal
Muskaan Singhania
Date: 8/8/2022

Analysis and Prediction of Customer Lifetime Value

Description: Using the customer value dataset from IBM, our goal is to explore the different variables and discern any patterns relating to a customer's lifetime value. The dataset contains information about 9000 insurance customers. The dataset gives us insight into the customer features such as their education, income level etc. and features of the policy they are on such as Policy type, Coverage.

Importance of Problem: Obtaining more information about the potential lifetime value of a customer is of interest to all businesses. We want to build this prediction model to achieve this task because this is a problem relevant to all industries in the real world. Having a metric to predict customer lifetime value will be useful for a company in defining future objectives, directions, and budget allocations.

Exploratory Analysis

We discerned the following patterns after taking a closer look at the data:

- Our target variable, Customer Lifetime Value, is very skewed to the right, meaning that very few customers have an exceptionally high CLV. (Fig 1)
- Some features that showed an association with CLV are Coverage, Renew Offer Type, Number of Policies, and Monthly Premium Auto. (Fig 2)
- Surprisingly, some features such as Gender, State, Response, and Location Code didn't seem to show any association with CLV (Fig 3)
- Insights from the bivariate and multivariate analysis
  - Customers with Premium Coverage have high CLV
  - Customers with Luxury cars tend to have high CLV
  - Customers with two policies have high CLV which is not surprising but what is surprising is that customers with anything above two policies have a similar distribution of Customer Lifetime Value
  - Customers living in rural and urban areas seem to have similar

distribution of Customer Lifetime Value
- There is a weak positive relationship between monthly premium auto and customer lifetime value.
- We don't see any correlation between income and customer lifetime value

Feature Selection

PCA Analysis: PCA Analysis was done to reduce the dimensionality of the data, and isolate what is the most impactful in predicting a customer's CLV. The following data preprocessing were done for PCA analysis:

- Creation of Y Variable, dropped Index Column i.e Customer Label, and encoded the categorical columns into numbers

Outcome:

- Using cross-validation, we found that the "elbow" flattened out around 4 Principal Components, which indicates that an efficient and predictive model would have around 4 variables
- Charted the variables in order of importance from highest to lowest.
- The MSE leveled out around 4.625%
- This analysis is imperfect since this uses categorical variables, and ignores continuous variables, but gives us a sense of the relative importance of the categorical variables

Data Preprocessing

- To start off, we checked for the missing values in the dataset and found none
- From the EDA we know, that the customer lifetime value is heavily right skewed, hence we disregarded the values above 60000 to ensure that it wasn't affecting our model's predictive capability
- From the above conducted PCA analysis, we selected the top 4 categorical variables and dropped the rest of the categorical columns i.e Policy_Type, Policy, Renew Offer Type, Sales Channel, Vehicle Class
- Using the insights on Number of Policies from EDA, we figured that bucketing 3 and above number of policies together since they have the same average CLV makes sense and made it into a categorical column
- Using get_dummies function, the selected categorical variables were encoded

- Coming the continuous variables, we dropped Income from the feature set as there was no correlation between CLV and Income as found from the above graph and from the correlation matrix, we notice that the association between the target and Month Since Last Claim and Month Since Policy Inception to be less < 0.04, hence we dropped these columns from our feature set as well

Modeling

1. Linear Model

We initially tried Linear regression with all the predictor variables but the R2 and accuracy were very low. Next, using the result of PCA and inferences from EDA, we included the most relevant variables in the model.

Below are the model results:
The R2 is ~65%, so by using the Coverage, Education, Marital Status, Number of Policies, Monthly Premium Auto, Number of Open Complaints and Total Claim Amount, we can explain 65% of the variation in CLV.

The plot of Y actual vs Y Pred (Fig.) gives a near 45-degree line low CLV customers. This shows that the linear model is able to predict correct CLV for low value customers but unable to explain variation in the high value customers.

We have used the regression model as a baseline model and tried more models to predict more accurately.

2. Gradient Boosting Regression

To improve upon our baseline model, i.e, Linear Regression. We ran the Gradient Boosting Regression on the dataset. The accuracy did improve (Please see Table no: 2 Modelling Appendix), Gradient Boosting Regression was able to explain more of the variation and we also noticed a decline in the test MSE when compared to the Linear Regression Model

Further on, we tried to tune the hyperparameters of the model such as the loss, learning rate and n-estimators by running a grid search to pick the most optimum values. The result we got from hyperparameter tuning suggested that the default values for the hyperparameters were optimal.

As we can see, there has been considerable improvement from our baseline model. When we compare the Linear Regression and Gradient Boosting Regression Actuals vs Predicted scatter plot(Fig 1 and Fig 2 of Modelling Appendix) , the points are closer to diagonal in the latter than in the former which basically means that the fit has improved.

3. **SVM Classifier**

   As an extension to the work already done, we wanted to try out classification Model for a broader use case. We created two classes i.e., High value and Low Value Customers and split it based on median Customer Lifetime Value

   We ran the Ada Boost model with the selected features which gave us 64% accuracy. Further on, we ran a support vector classifier on our data to reduce the miss-classification rate. This model gave us 92% accuracy. Since this model gave us great precision and recall, we decided to stick with it.

Conclusion and Recommendations :

- Gradient Boosting is 70% accurate while predicting CLV as a continuous variable.
- Support Vector Classifier is 93% accurate while predicting the group customer belongs to - High or Low CLV group
- Recommendations to use CLV by the Insurance company:
  - Targeted Product Pitching: High CLV customers prefer Policy Renew Offer 1 so they should be targeted for that.
  - Customer Segmentation: Using CLV, customers can be segmented, and more budget can be allocated on marketing campaigns for High CLV customers.
  - Customized Product Development: According to CLV, product pricing, terms and conditions can be customized.

Appendix:

PCA Analysis Figures:

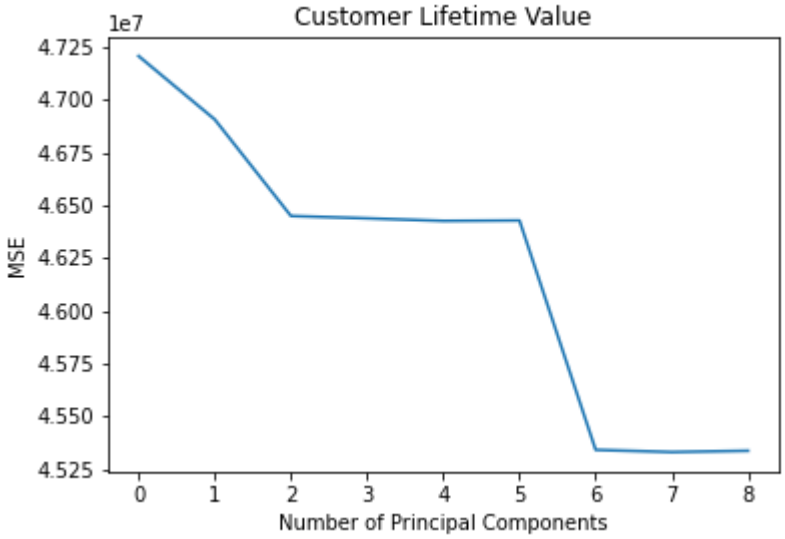


Figure 1: This is our PCA graph, measuring the impact on MSE for each number of Principal Components. Note that the graph levels around 9 principal components.

Predictors	Importance
Coverage	20.89
Education	12.98
Marital Status	11.42
Policy Type	11.37
Policy	11.28
Renew Offer Type	10.75
Sales Channel	10.65
Vehicle Class	9.32
Vehicle Size	1.33

Figure 2: This is our variable importance table, in order of most important to least important. The importance. column shows the percent of variance explained by the predictor.

Exploratory Analysis Figures:

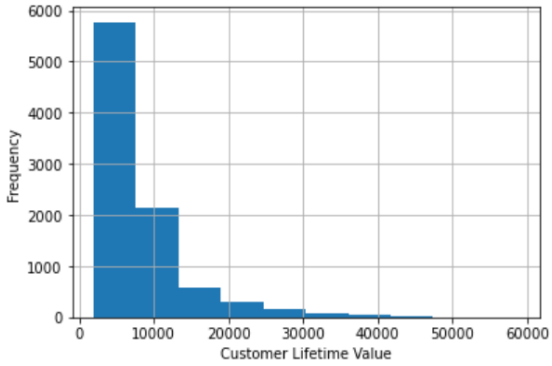


Figure 1: A histogram showing the skewness of Customer Lifetime Value

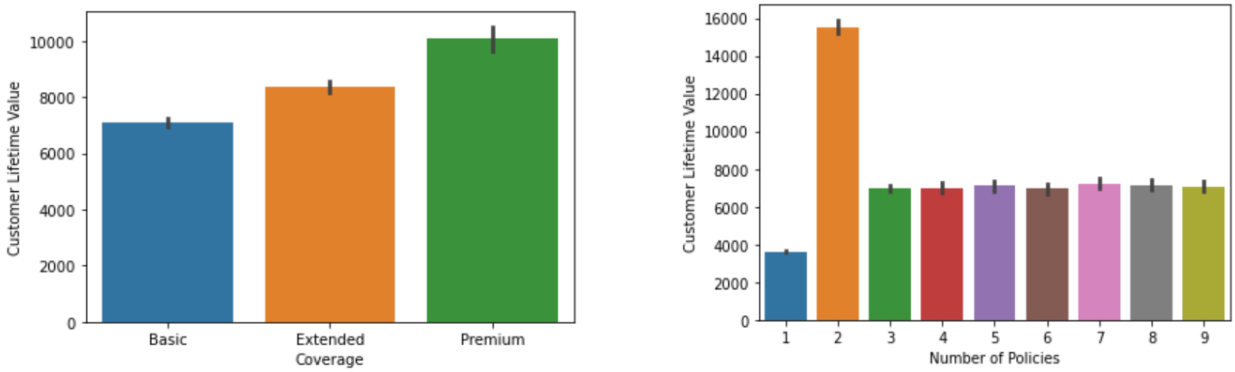


Figure 2: Relationships of CLV with Coverage and Number of Policies

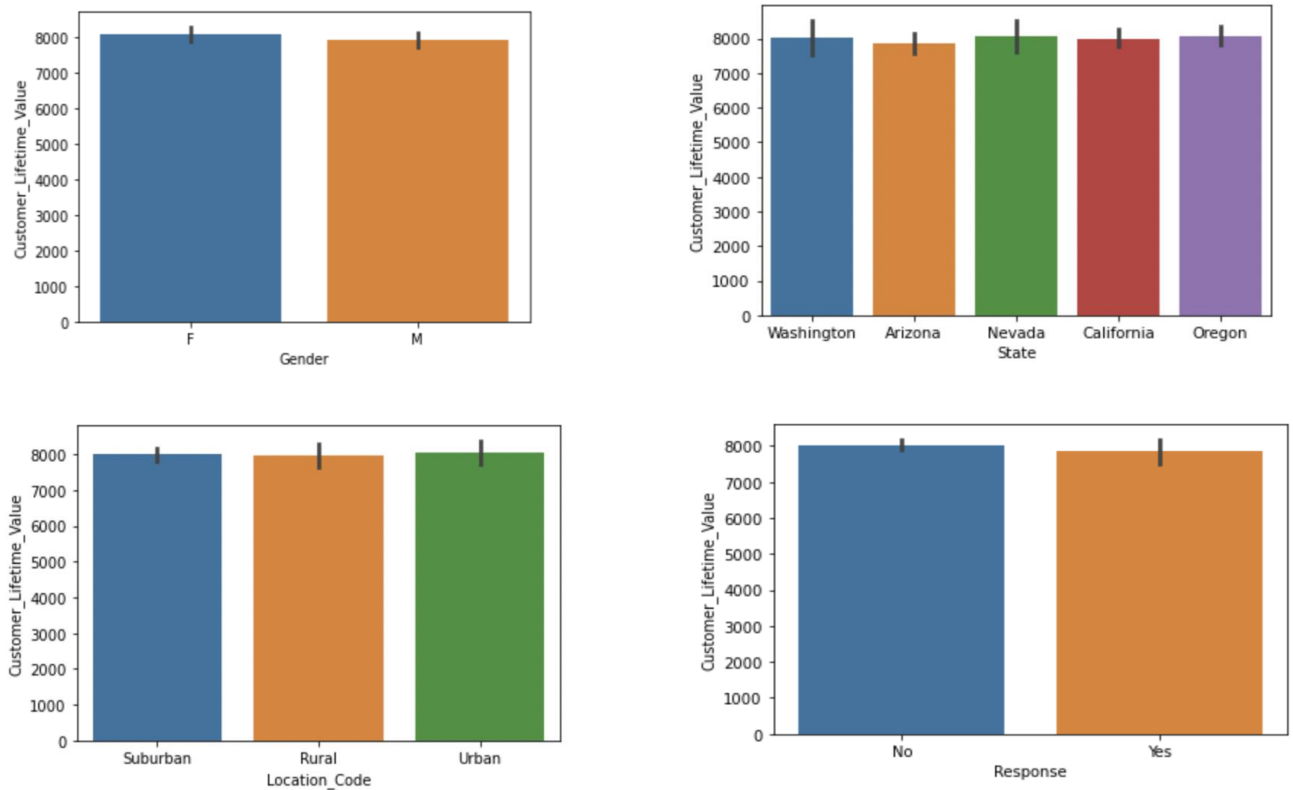


Figure 3: No relation of CLV with Gender, Location Code, State and Response

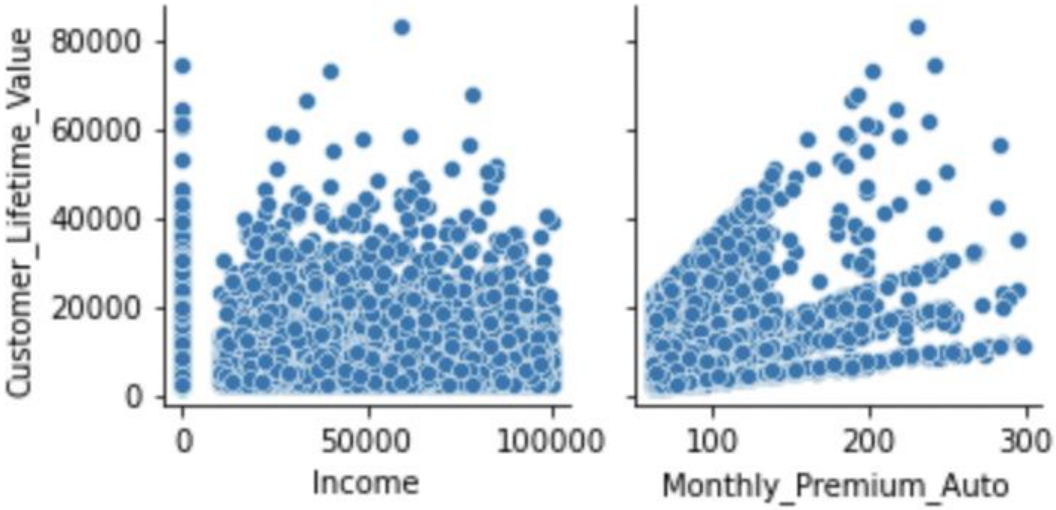


Figure 4: Relationship of Income vs Customer Lifetime Value and Monthly Premium Auto vs Customer Lifetime Value

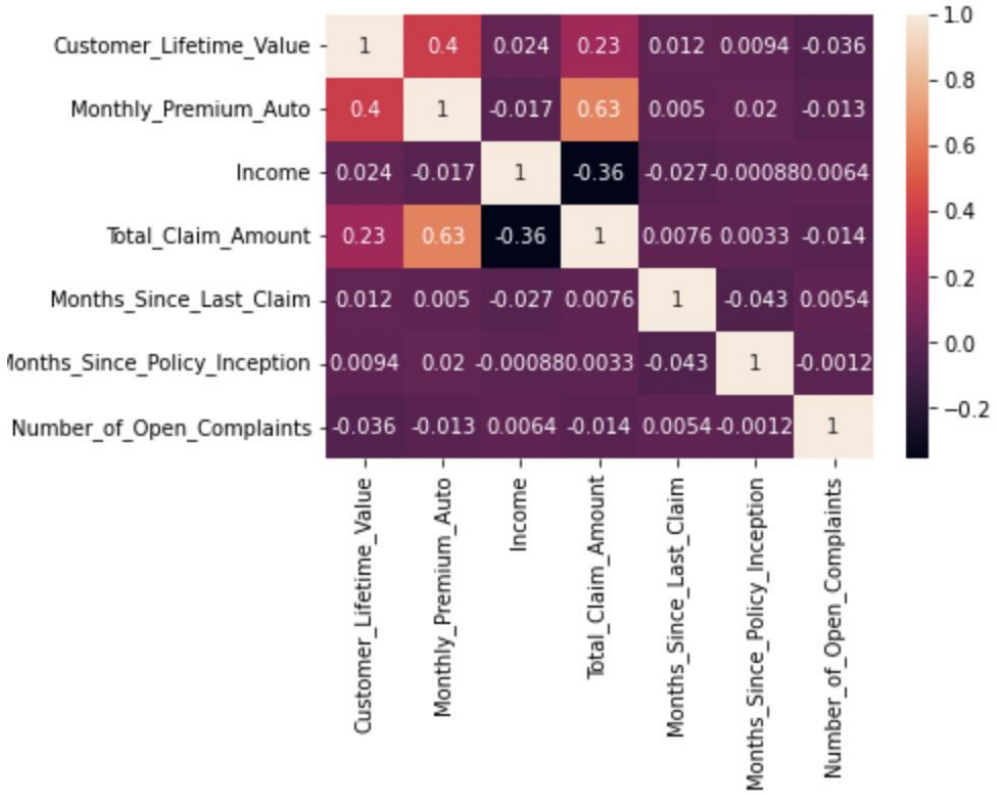
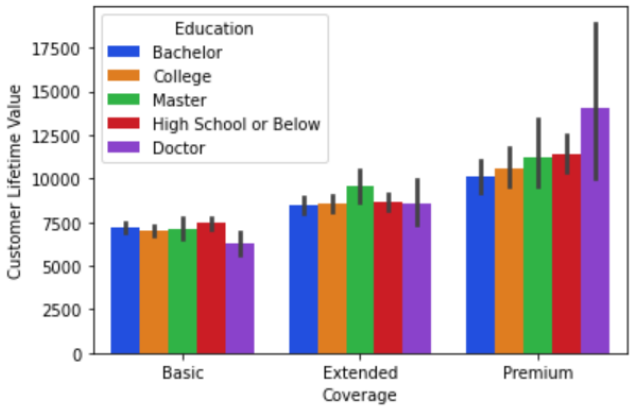


Figure 5: Heat Map of Numerical Variables in Dataset



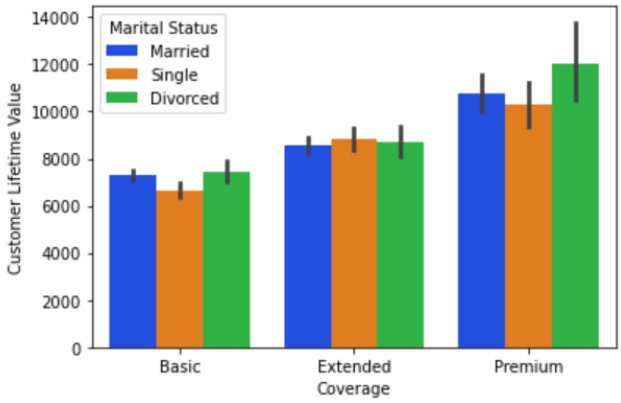
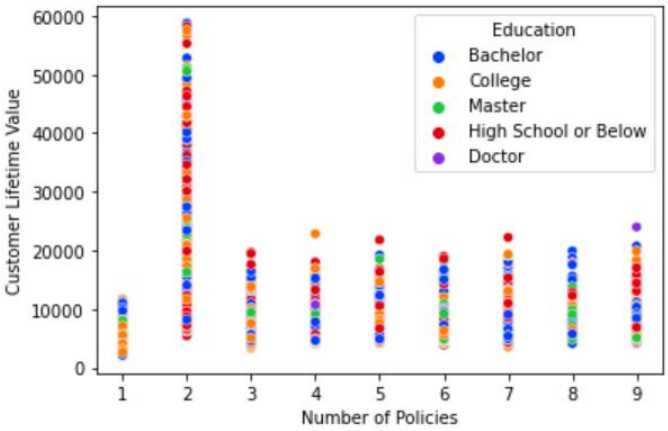


Figure 6: Relationship of CLV with Coverage with respect to different Marital Status and Education



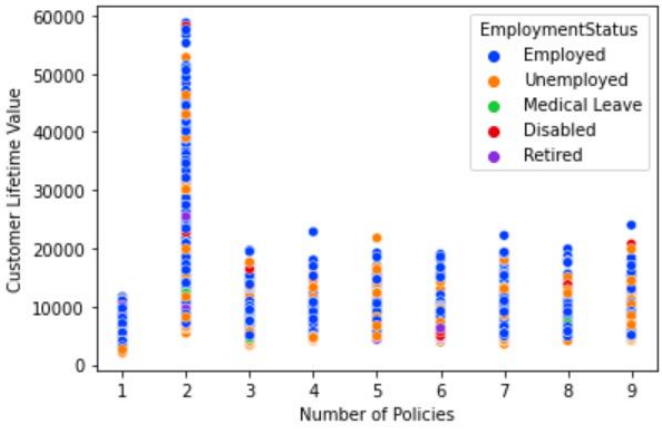


Figure 7: Relationship of CLV with Number of Policies with respect to different Education and Employment Status

Modelling Table and Figures

Metric	Value
R2 Score	70.01%
MAE	1608.787
MAPE	13.49%
Test MSE	12232373.3

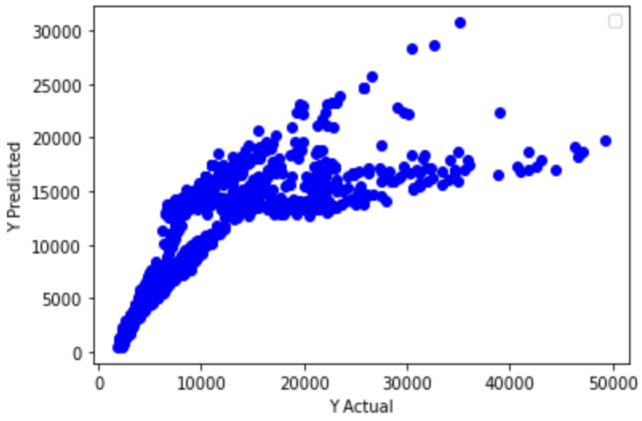


Table and Figure 1: Linear Regression Results

Metric	Value
R2 Score	66%
MAE	1884.5
MAPE	22%
Test RMSE	3701.4

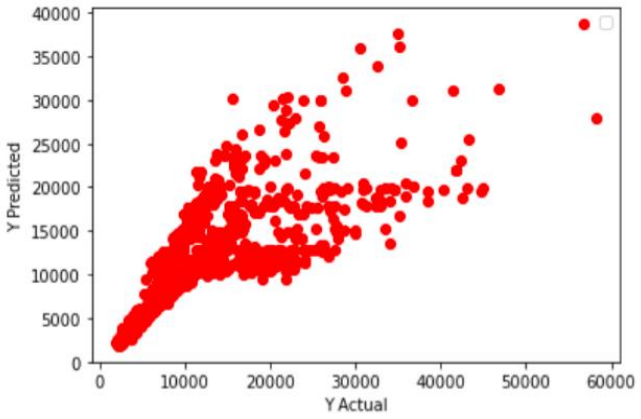


Table and Figure 2: Gradient Boosting

ADA Boost Results:

Segments	Precision	Recall
High	0.71	0.46
Low	0.60	0.81

Table 3

Support Vector Results:

Segments	Precision	Recall
High	0.91	0.96
Low	0.95	0.91

Table 4