Predicting
Customer Lifetime Value

CONTRIBUTORS

MUSKAAN SINGHANIA

MS88283

MUSKAN AGARWAL

MA64547

PRANAV CHERUKU

prc724

SUNIL KAMKAR SHESHAGIRI

SK56743

VIVIAN WANG

VW3852

Problem Statement

Determine Customer Lifetime Value (CLV) for Insurance Customers

Predicting Customer Lifetime Value is important for:

- Improving Customer Retention
- Determining customer segmentation
- Measuring the customer loyalty
- Determining efficacy of marketing strategies
- Increasing profitability overall

Understanding the Dataset

Overview

9134 Observations

22 Features

1 Target Variable

Customer Features

- State
- Education
- Employment Status
- Gender
- Income
- Marital Status
- Vehicle Class

Policy Features

- Customer
- Response
- Coverage
- Policy
- Policy Type
- Premium auto
- Sales Channel
- Total Claims

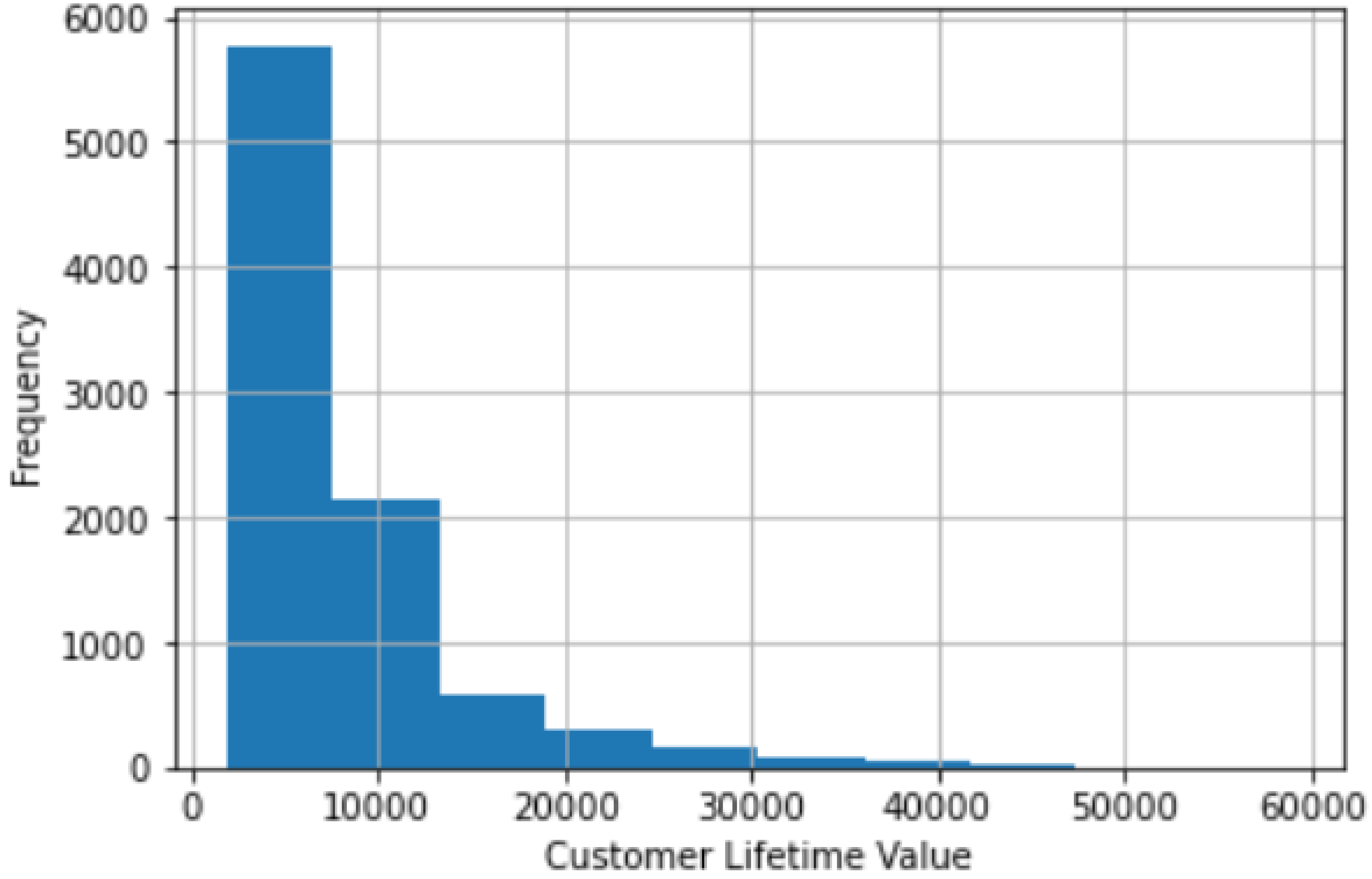
Target Variable

- Customer LifeTime Value

Approach

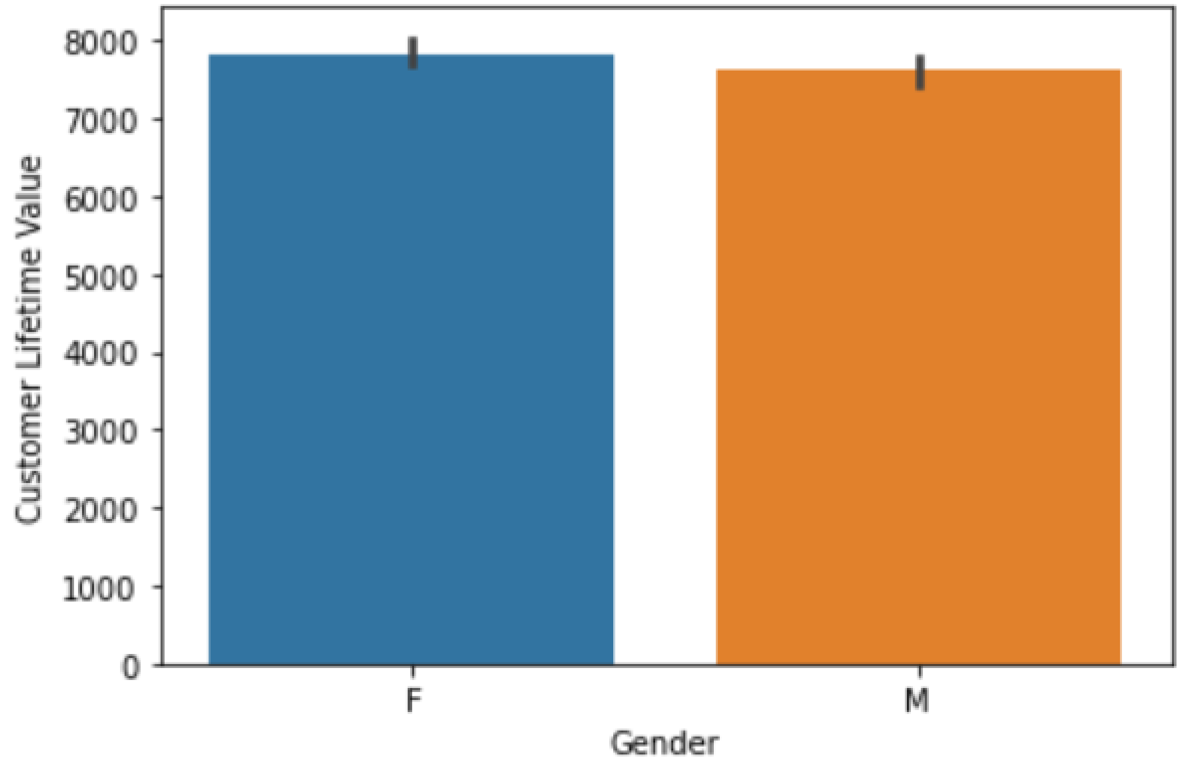
1. Bivariate Analysis
2. Multivariate Analysis
3. Feature Selection and Preprocessing
4. Modeling
5. Model Comparision
6. Conclusion

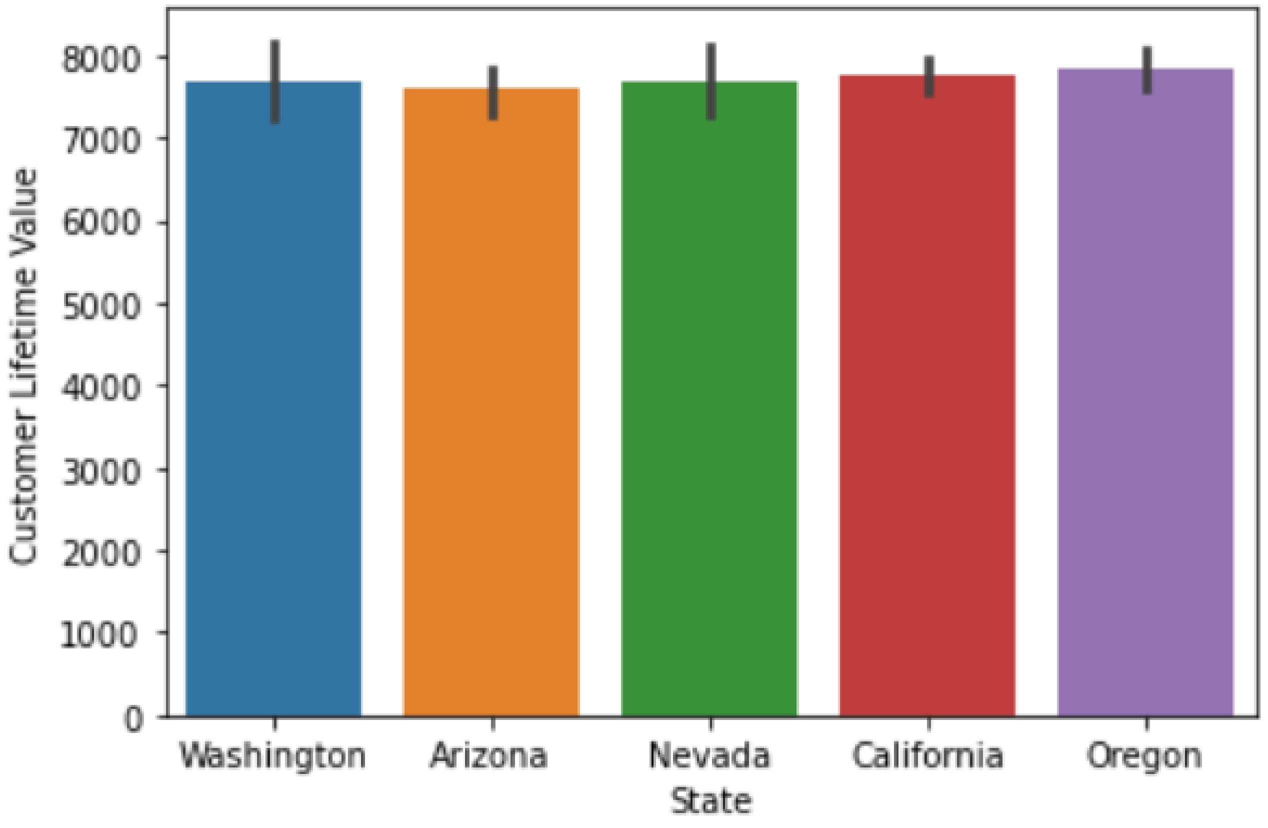
Customer Lifetime Value

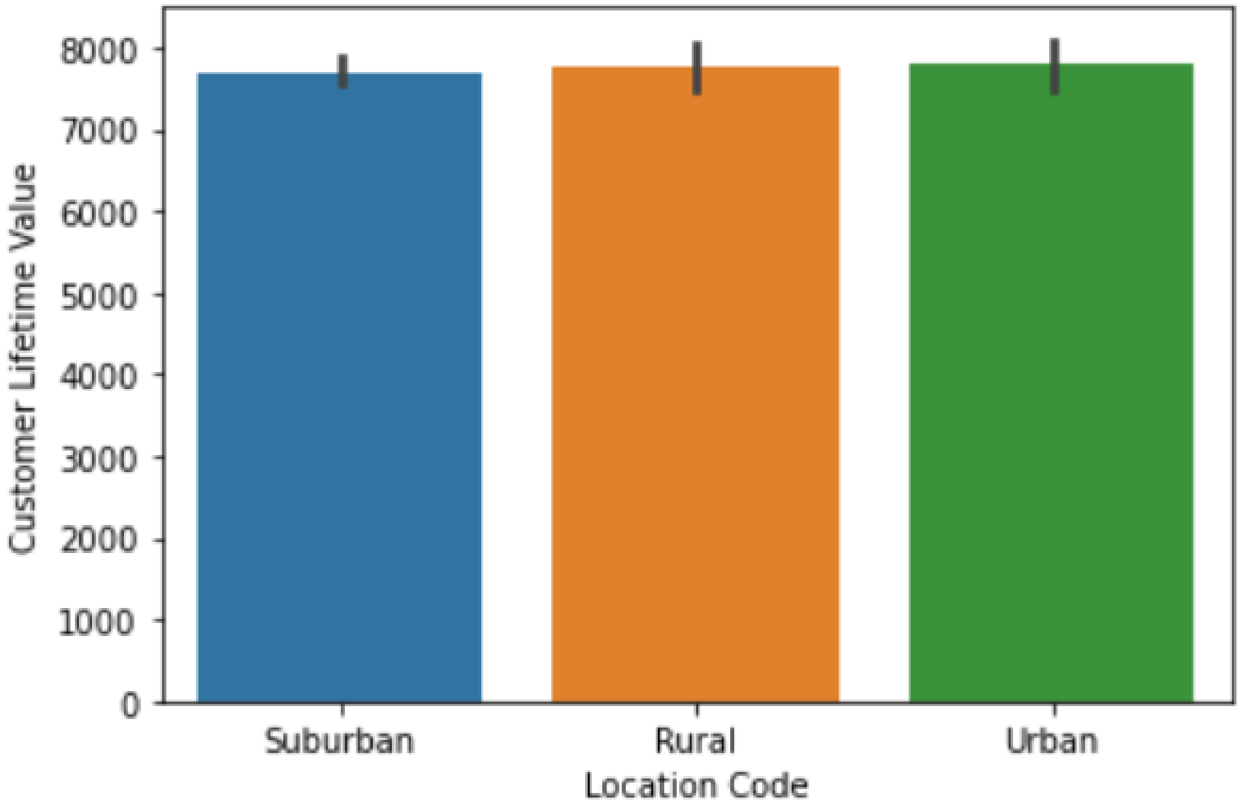


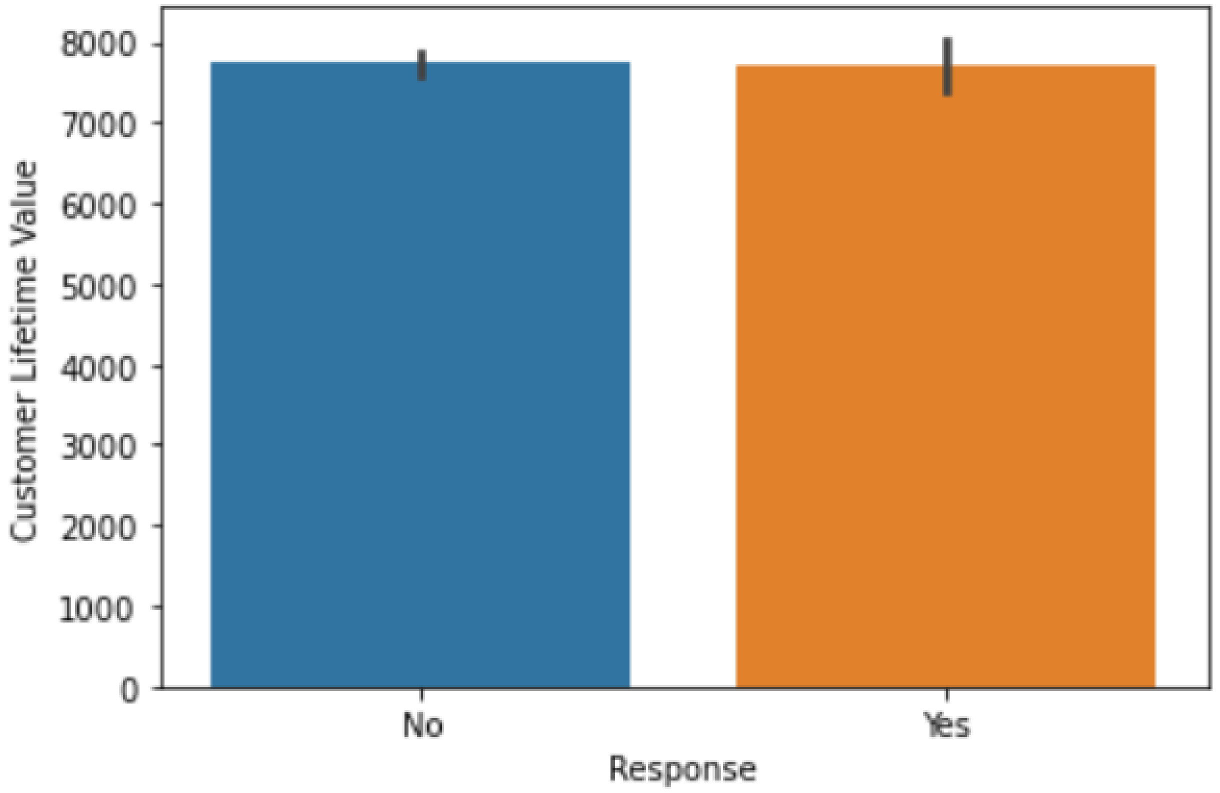
CLV is very right-skewed indicating imbalance of target variable

Bivariate Analysis



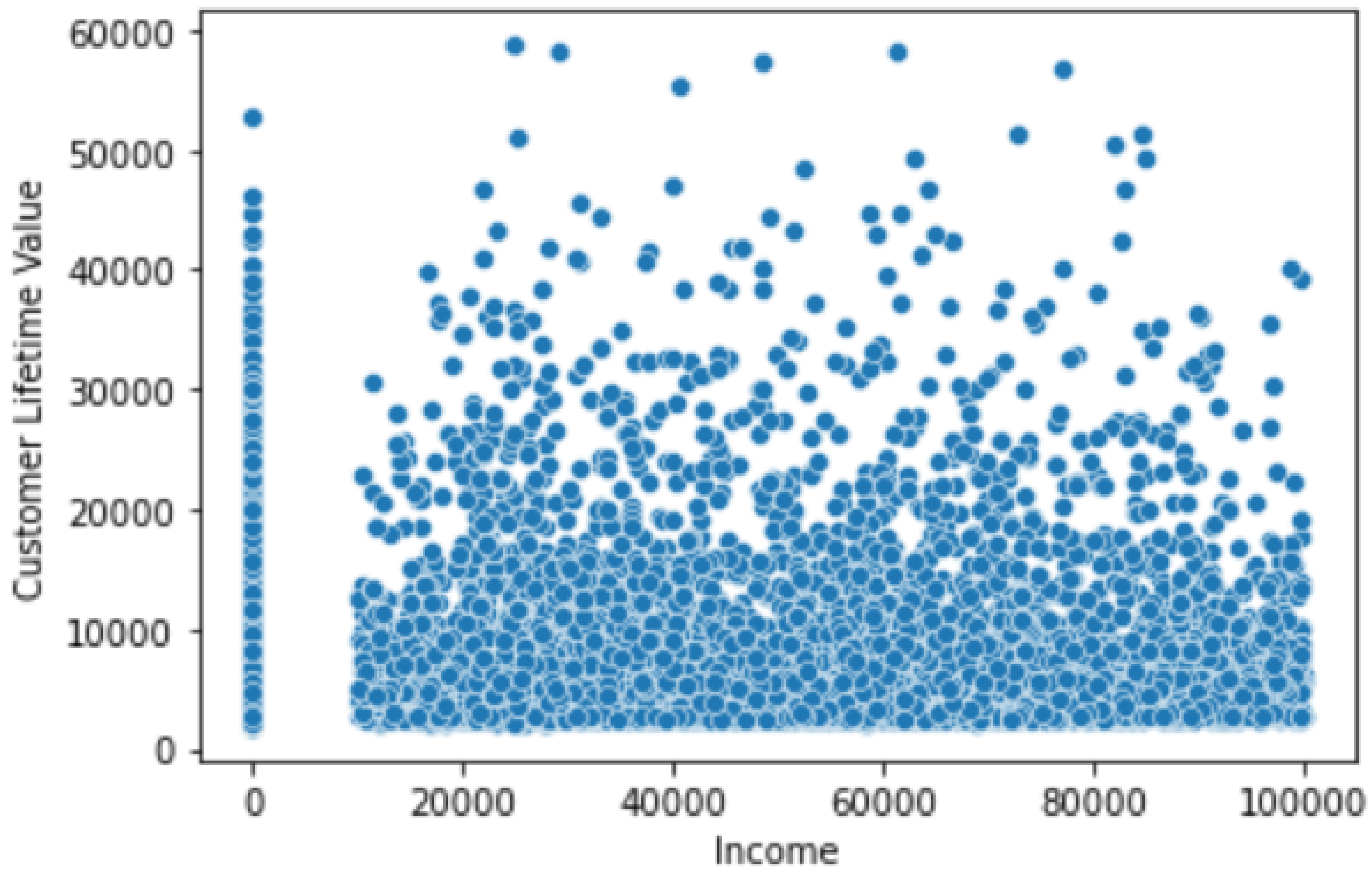




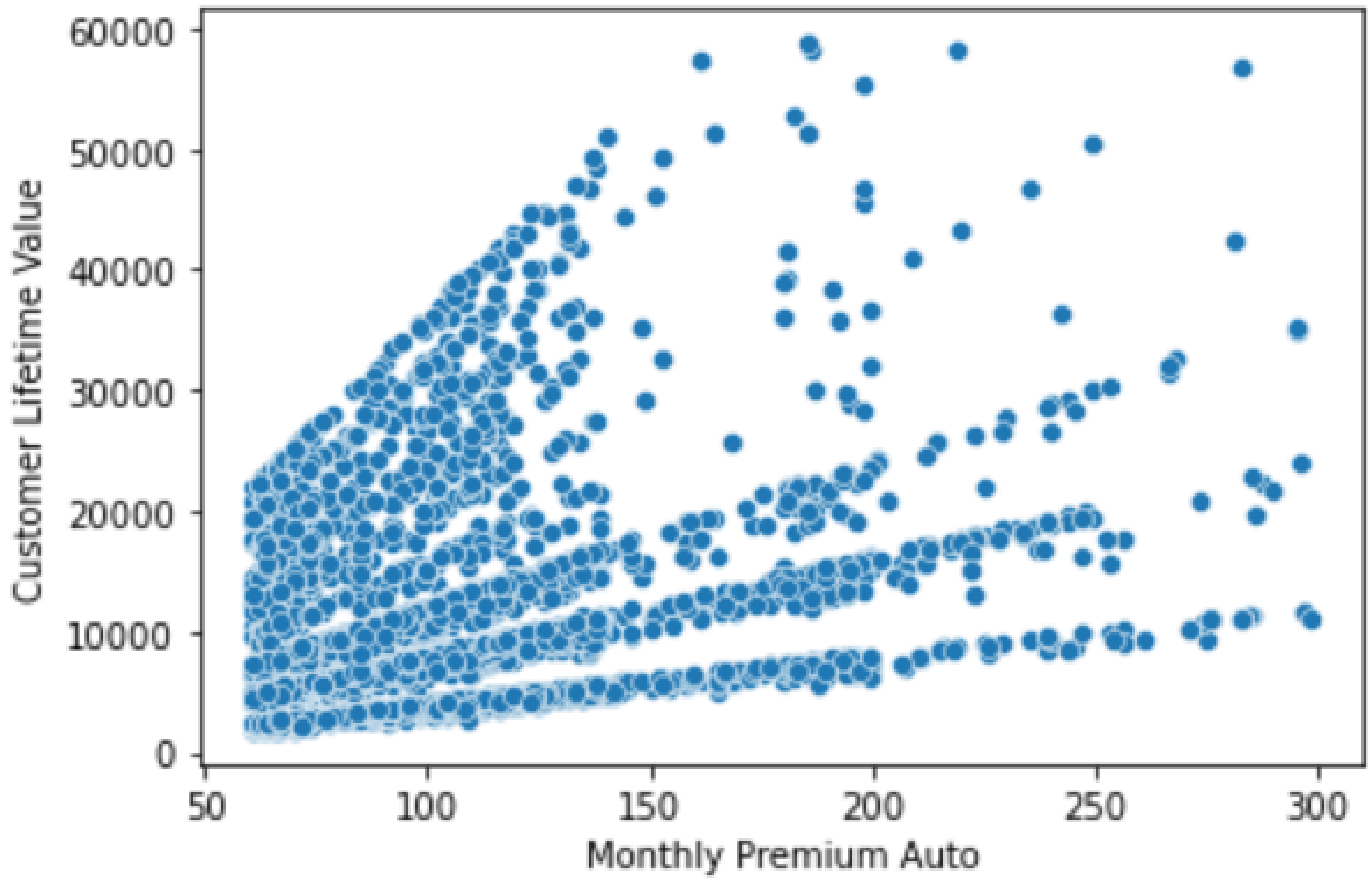


Variables that showed no association with CLV

Bivariate Analysis

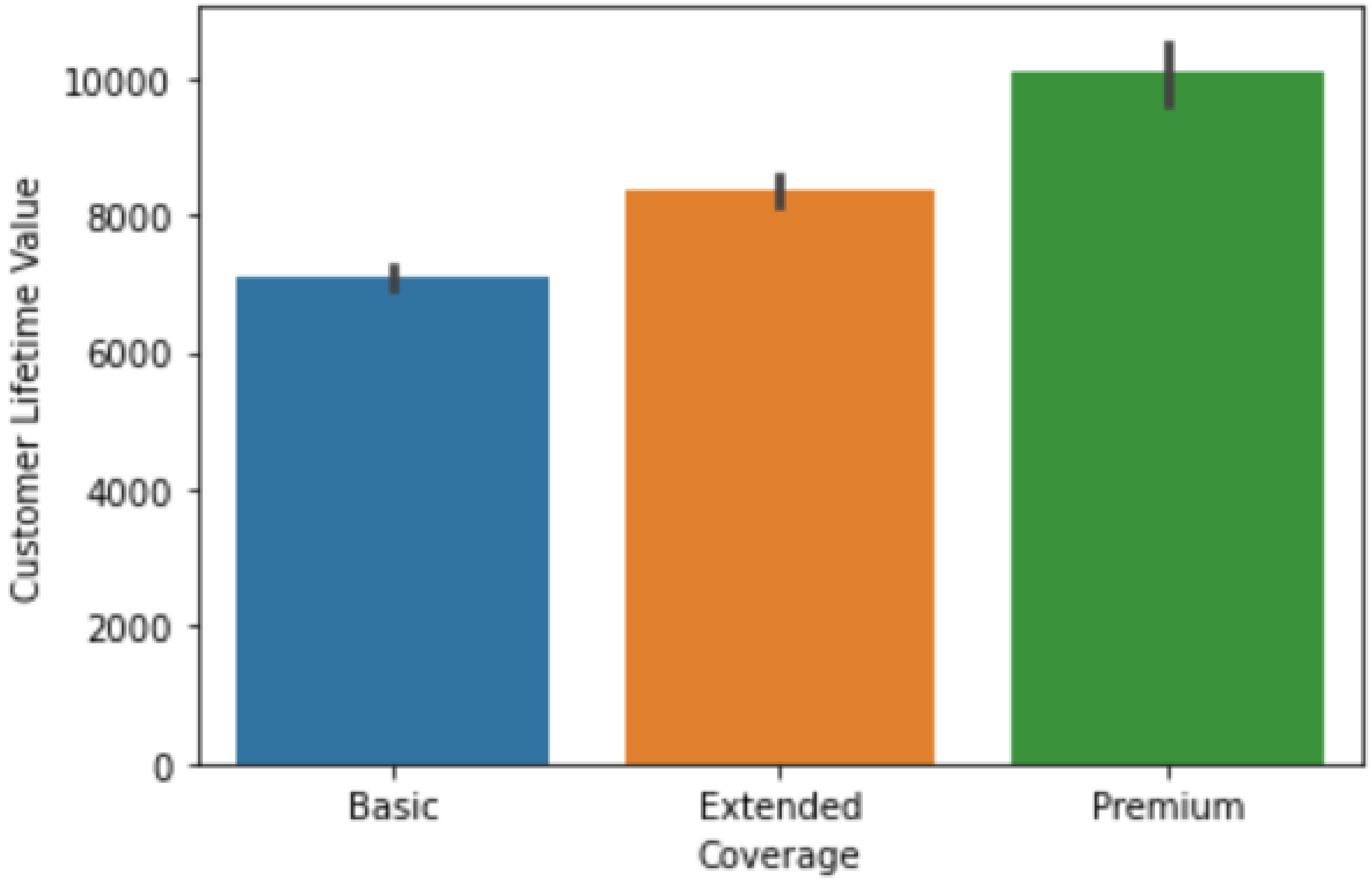


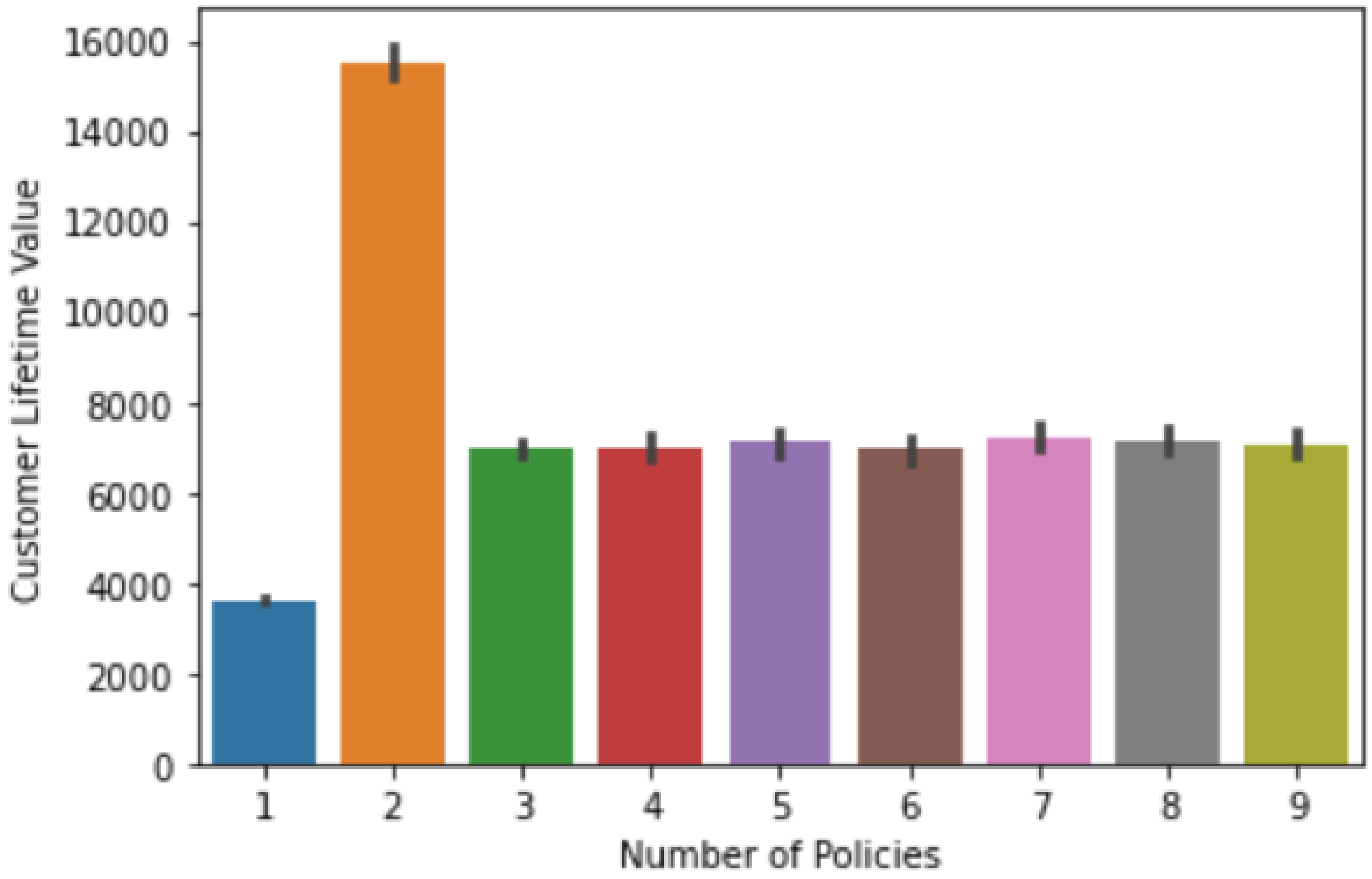
No association between Income and CLV



Weak positive correlation between Monthly Premium Auto and CLV

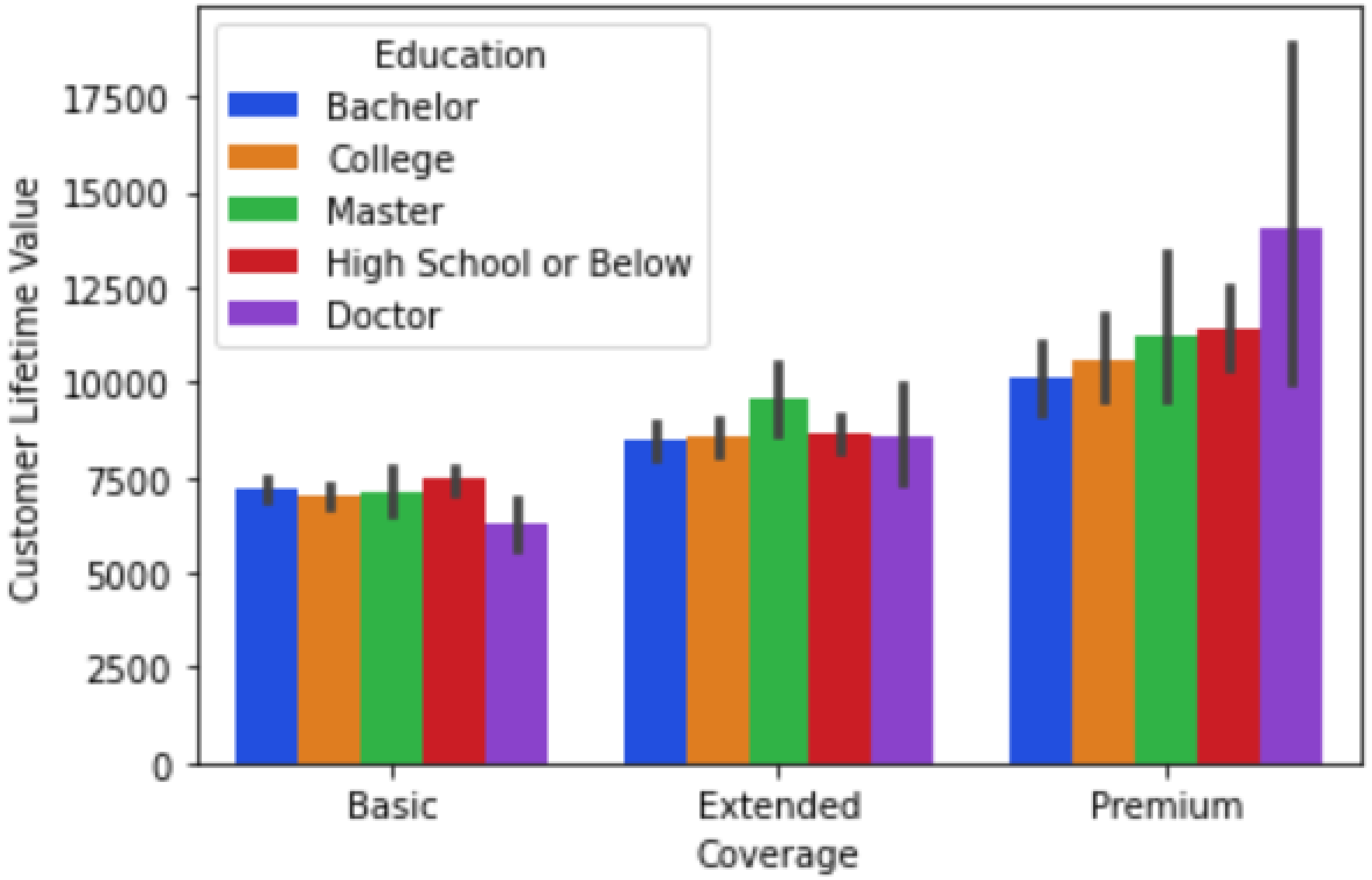
Bivariate Analysis

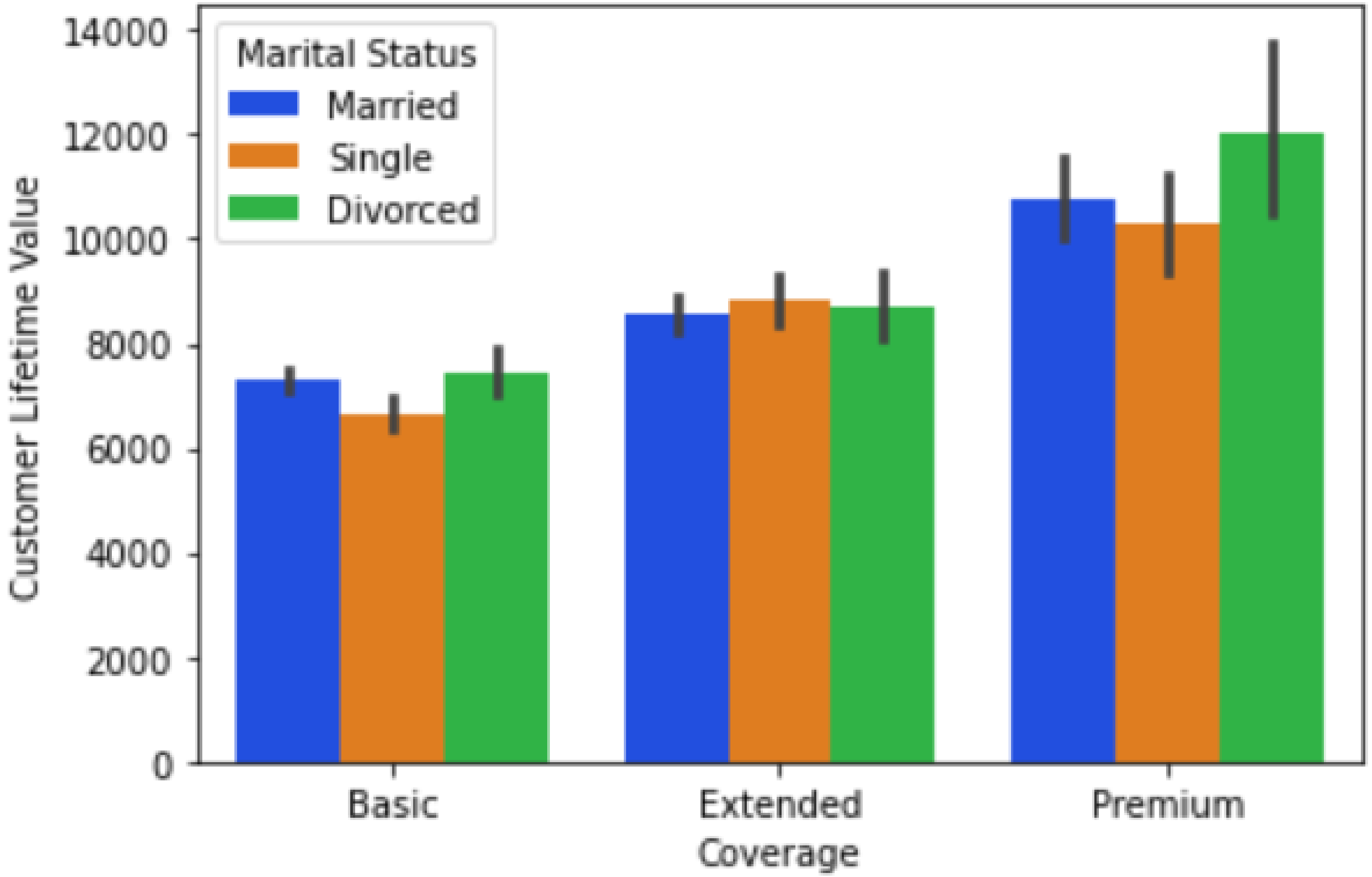




Relationships of CLV with Coverage and Number of Policies

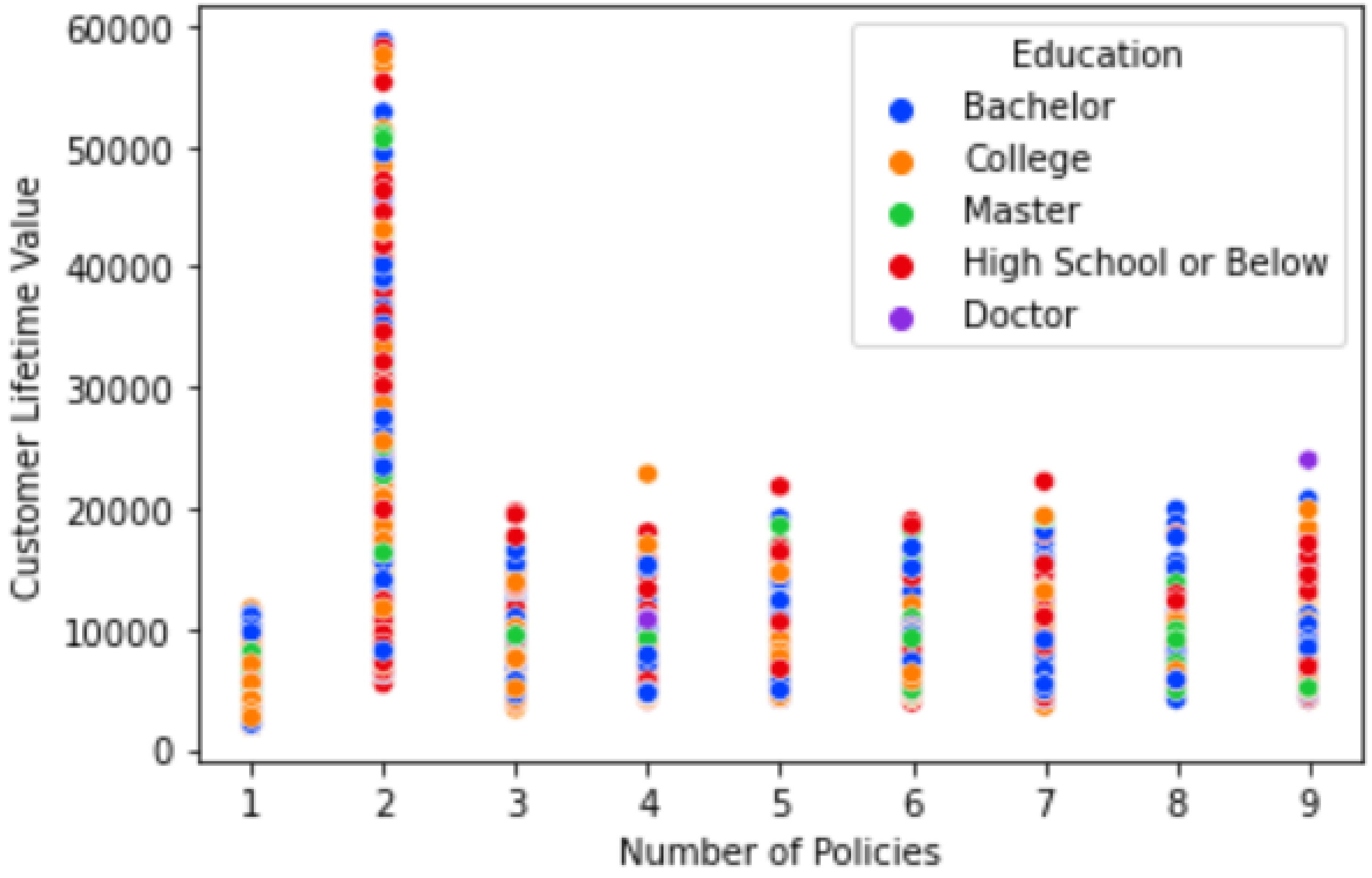
Multivariate Analysis

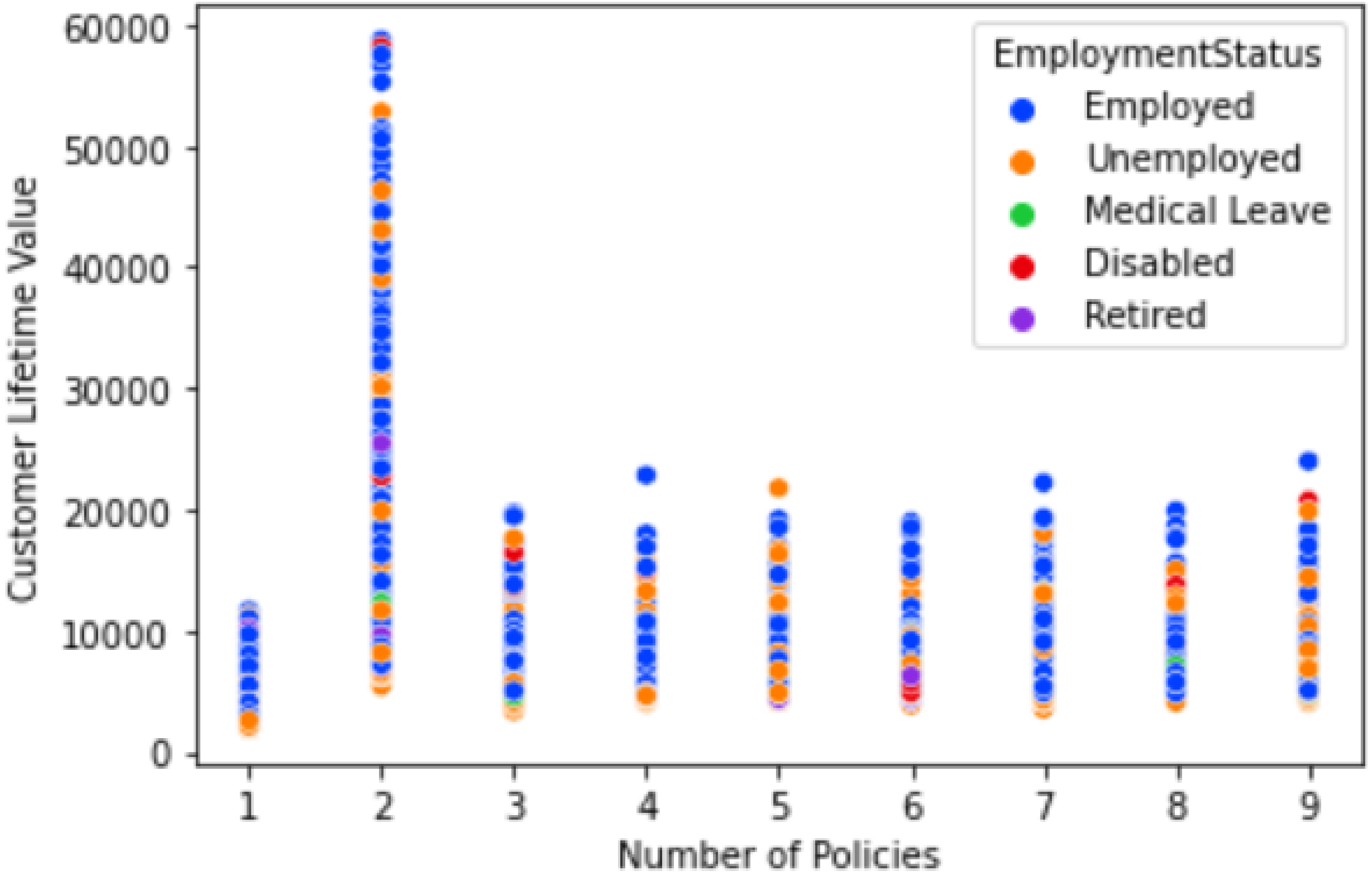




Relationship of CLV with Coverage with respect to different Marital Status and Education

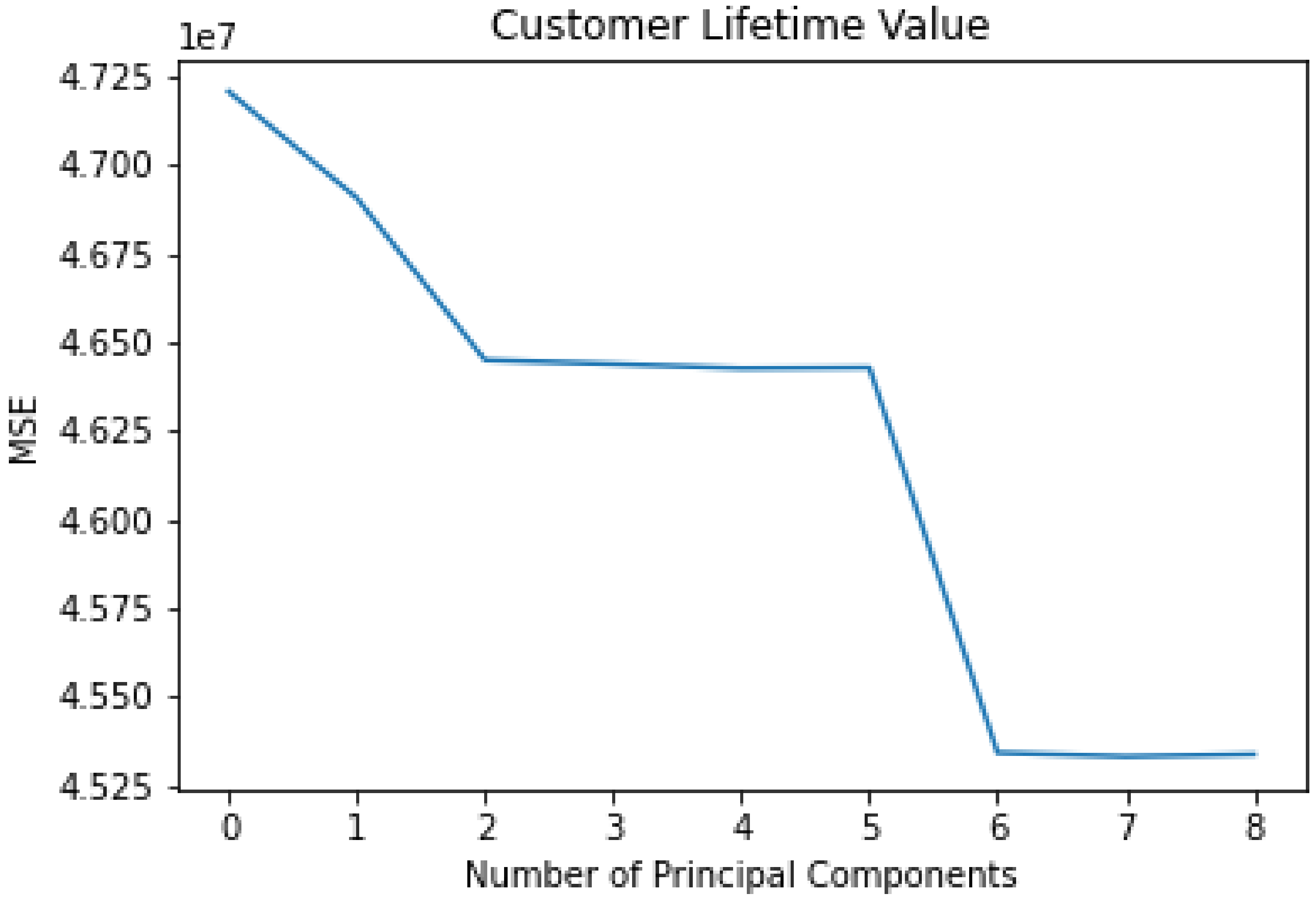
Multivariate Analysis





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

PCR and Feature Selection



- Find most important variables and reduce dimensionality
- Coverage, Education, Marital Status, and Policy Type

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

Feature Selection

Features	Correlation
Monthly Premium Auto	0.4
Total Claim Amount	0.23
No of Open Complaints	0.036
Income	0.024
Months Since Last Claim	0.012
Months Since Policy Inception	0.036

Correlation between numerical variables and target in the dataset

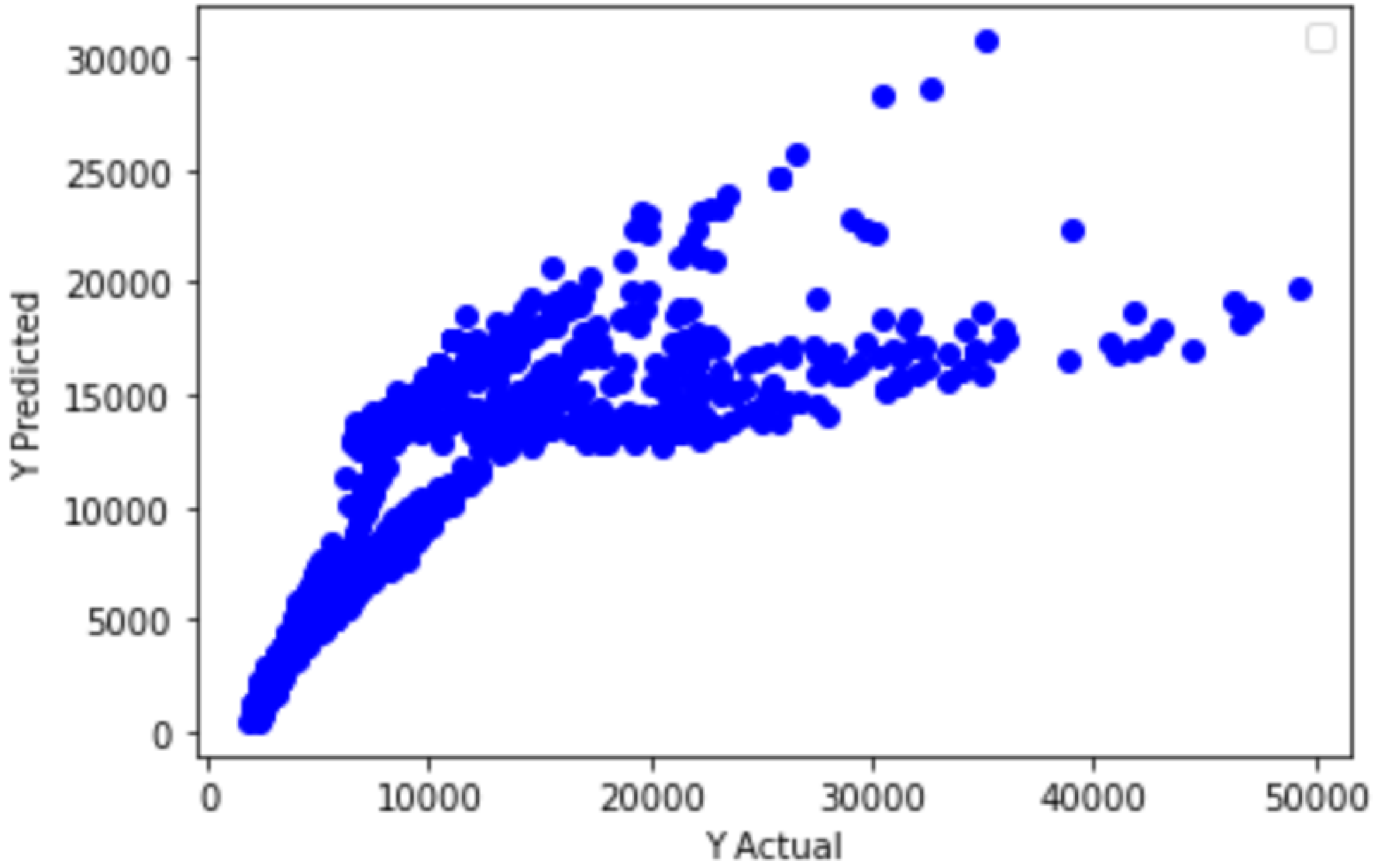
Based on EDA, PCR and correlation, we narrowed down to 7 features from 22 features

Data Preprocessing

- Removal of Outliers from Target Variable
- Bucketed No of Policies
- Encoded the categorical variables
- Dropped the irrelevant columns

Linear Regression

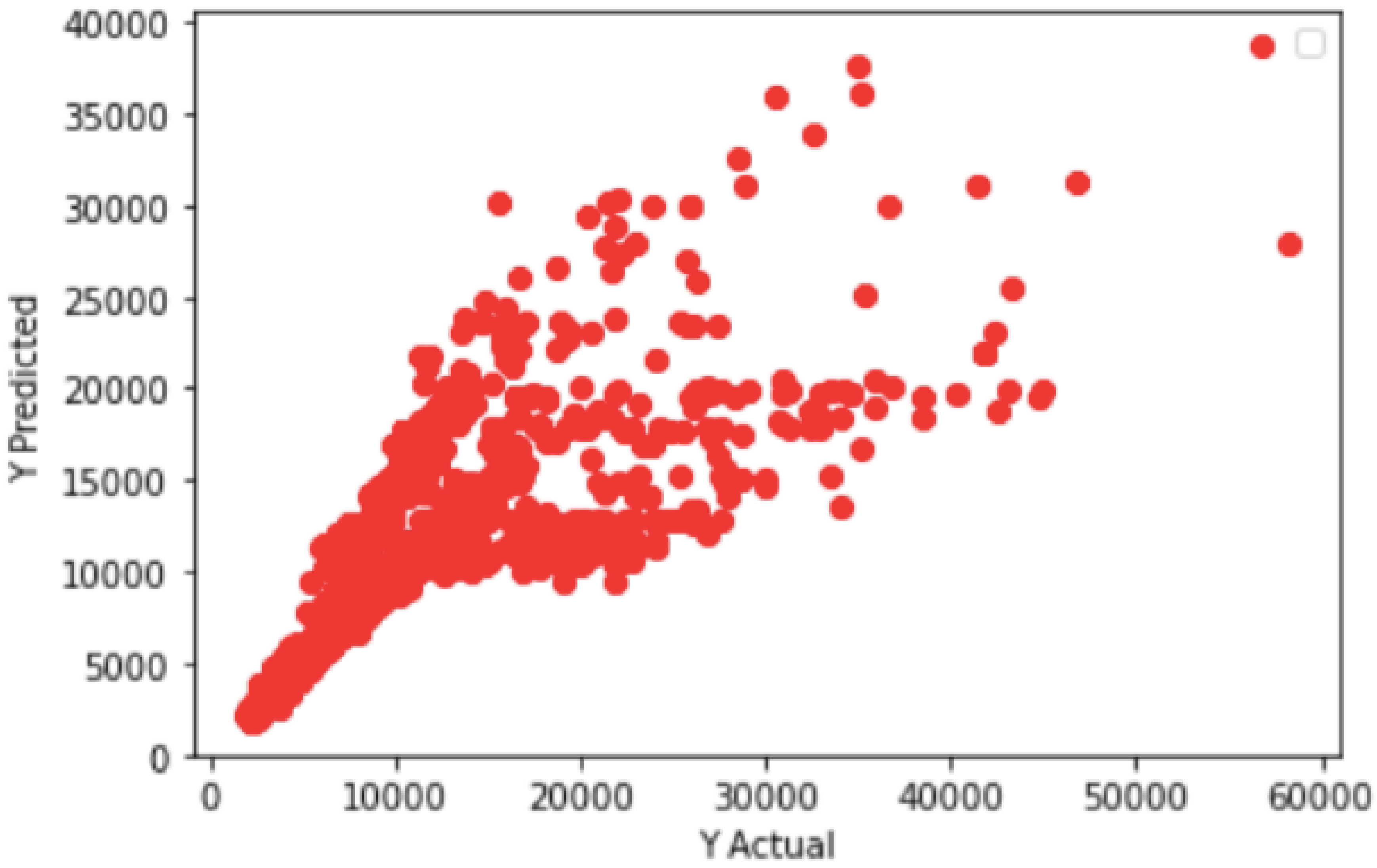
Metric	Value
R2 Score	66%
MAE	1884.5
MAPE	22%
Test RMSE	3701.4



The linear model is able to predict correct CLV for low value customers but unable to explain variation in the high value customers.

Gradient Boosting

Metric	Value
R2 Score	70.01%
MAE	1608.787
MAPE	13.49%
Test RMSE	3497.48



Improved R2 and decrease in Test RMSE. The chart illustration shows that this model fits the data better

CLASSIFICATION

SEGMENTATION OF CUSTOMERS

Split Criteria:

- CLV > Median(CLV) is High else Low

High Value

Low Value

Ada Boost

Segments	Precision	Recall
High	0.71	0.46
Low	0.60	0.81

ACCURACY: 63.9%
F1 Score for High: 0.56
F1 Score for Low: 0.69

Support Vector

Segments	Precision	Recall
High	0.91	0.96
Low	0.95	0.91

ACCURACY: 93.2% | Error Rate: 6.8%
F1 Score for High: 0.93
F1 Score for Low: 0.93

Conclusion

Model	Accuracy	Error
Linear Regression	66%	22%
Gradient Boosting	70%	14%
Ada Boost	64%	36%
Support Vector Classifier	93%	6.8%

Recommendations

- Targeted Product Pitching
- Customer Segmentation
- Product Development

Scope of Improvement

Coming up with a more elegant method of picking threshold for converting the Regression into a classification problem