Quantitative Auto Insurance Pricing Strategy

A Group: Lavanda Wang, Weijian Shi, Shaoxiong Jia, Fenghe Liu, Wenyu Xu

Introduction

As machine learning becomes an increasingly effective modeling approach in the finance world, professional actuaries endeavor to employ machine learning tools into insurance pricing these days. In this project, our team will walk into actuaries' shoes and create a quantitative auto insurance pricing strategy employing the data science skills we learnt from this course.

Data

We downloaded the dataset from Kaggle and fortunately it was very clean and well processed. We merged it with another dataset containing information on car accidents in different states due to the reason that car accidents frequency and severity might have strong correlation with auto insurance. After merge, the data has 9134 entries and 26 columns in total. See Figure 1.

Exploratory Data Analysis

Distribution of features(See Figure 2)

We firstly explore the distributions of the features because we want to build up a good understanding of how our customers are distributed regarding different features. This also helps us recognize outliers and form hypotheses of feature correlations. We can see some features have an exponentially decreasing distribution such as the number of policies and employment status; some have approximately linear decreasing distributions such as the marital status and the months since last claim. Very interestingly, the total claim amount has a skewed normal distribution.

Violin plots of monthly auto premium under different features(See Figure 3)

We then explore the relationship between features and our final target variable - monthly auto premium. From the violin plots we can tell that the coverage plan and vehicle class have very strong relationships to the monthly premium.

Numerical features correlation table(See Figure 4)

From this correlation table, we can tell that the number of accidents in the state has a slight positive correlation to the monthly auto premium, although we expected this number to be much. Customer lifetime value and total claim amount have very strong positive correlation to the premium. As the total claim amount increases, the risk expected to involve the customer is higher, thus the insurance company charges more premium. And as the total claim amount goes extreme, the company will consider dropping the customer.

Total Claim Amount vs. Monthly Premium(See Figure 5)

We dive more into the relationships between total claim amount and the monthly premium, along with the coverage plan and vehicle class that are identified in different colors in the below charts. The different color distributions indicate that premium plans have higher premiums than the basic and extended plans most of the time. The premiums of basic and the extended plan are quite intertwined, but they still appear to have different levels of premiums. Similarly, the premiums of different vehicle classes are intertwined with each other, but luxury cars appear to have much higher premiums than the rest. Moreover, below around 500 of total claim amount, it was noted to have an obvious correlation with the monthly premiums, however, as we go above and beyond, there is an obvious positive linear relationship between the claim amount and the premiums.

Customer Lifetime Value vs. Monthly Premium(See Figure 6)

We also dive more into the relationships between customer lifetime value and the monthly premiums. It appears that the customer lifetime value has an almost linear relationship to the premiums. The parallel linear lines in the chart indicate that there is another strongly correlated factor hidden. Although we are still unsure about the exact factor causing this, we speculate that it might be the length of business relationship that the company expects to have with the customer. Under this assumption, we can well explain the distinct linear lines by tracing to the different length of insurance relationships between customers and the company.

Pricing Strategy - Coverage Classification

When a new customer comes to the insurance company and asks for a quote (most insurance providers use this method to help the potential customers learn about their insurance rate), it's necessary to generally classify what kind of insurance plan a customer would consider. For instance, there may be a premium plan, which covers more and has a higher amount of compensation. Also, there are basic plans which may only cover the compulsory insurance, and mostly purchased by students, and whose cars are not that expensive. In this part, instead of emphasizing on the classification models, I prefer to indicate a potential imbalancing problem which will typically affect the model accuracy, especially for a sub classification.

Such problems have been indicated during the EDA part from two aspects. Firstly when we visualize the target values, we can see unbalanced data. From the pie and bar chart we can see that people purchasing premium plans are limited–only around 9% people enrolled. Typically in machine learning model construction, we consider this as an extreme unbalanced data problem.(See Figure 7)

Another way we can visualize this unbalanced problem is outlier visualization. Typically we use boxplot to detect the outliers and drop them, and the outliers takes little proportion of all the data. But in the box plot visualization for a few key features in the dataset, the proportion of "outliers" is so high that it is unlikely to purely drop them. For instance, for the monthly premium auto

feature, the outliers take over 8%. And for the total claim amount, the outliers proportion exceeded 5%, which should not be treated as outliers at all. Instead, they should contain considerable information about the underrepresented class, like the premium class.

If we ignore such problems and directly construct ML models for classification, the classification accuracy rate for the underrepresented class will be extremely low. From the confusion matrix of logistic regression, which is the most fundamental model, we can see for the premium class, only 24.4% of premium users are correctly classified, which indicated the influence from unbalanced data. (See Figure 9)

The way we solve this problem is utilizing SMOTE_NC, which is the Synthetic Minority Over-sampling Technique for nominal and continuous features. This method generates similar data points with the record in the unrepresented class, and it is designed to deal with the categorical data. After applying this technique, we noticed that the accuracy especially for the premium class has been improved dramatically. Now around 75% of them are classified correctly, and the overall accuracy reached 86.6%.(See Figure 10)

Pricing Strategy - Customer Lifetime Value

In the process of Lifetime Value Modeling , we firstly explore the significant factors for predicting lifetime value through drawing barplot and boxplot.(See Figure 11) Then, we use some specific analysis to deal with some factors, we find that when the number of policies equals to 2, it is different with when it is larger than 3, and we find that income, Employment Status, Marital Status are important factors, this can be shown with the scatter plot. Thirdly, We build the OLS model to predict. The result turns out to be solid. Our R-squared shows that our OLS model could explain more than 75% of the customer life value, along with an MSE of 0.109. (See Figure 13)

We dive more into the coefficients of the variables. The first is Total Claim Amount, the result shows that it has a u-relationship with customer lifetime value, which means that , as the total amount claimed increases, the customer lifetime value first decreases and then increases.(See Figure 14) That may reveal the structure of the cost and profit of Auto insurance. It is possible that when the Total Claim Amount is less than a number, customers are seen as “low risk”. They only need to give a little money to get insurance. Thus, as the total amount claimed increases, the cost increases but the revenue keeps the same, and the customer lifetime value decreases. However, when Total Claim Amount is out of the range, they can be seen as “high risk”. Therefore, as the total amount claimed increases, the customer needs a great amount of money to get their insurance, and the customer lifetime value increases.Something more interesting is the relationship between Marital Status and Customer lifetime value. It shows that Married people have higher Customer lifetime value than single people, following the divorced people. (See Figure 15)

At last, we built a k-means clustering model to classify customers into different groups based on customer lifetime value. The incentive of building the model is to classify the customers and target the type of customers that generate the most customer lifetime value. Based on the OLS regression result, we picked the variables that are significant (P-value <0.05). Using these variables, we clustered the customers into two groups, and we could see below that cluster2 has a higher customer lifetime value.(See Figure 16)

Then, we tried to get the characteristics of these two different groups. As shown below, we delved into 3 aspects: insurance statistics (indicated in bar charts at the top left), personal features (bar charts at the bottom left), and car type (pie charts on the right). In general, customers in cluster2 pay a higher amount of monthly premium and have a lower complaints rate. They also have more stable marital status and higher education levels. Besies, they also tend to buy more expensive cars, all these characteristics could easily explain why they have a higher customer lifetime value.(See Figure 17)

Pricing Strategy - Monthly Premium

In this section, we will build a machine learning model predicting the monthly auto premium for the customers. We applied the scikit learn library in the modeling and took 70% of the data points as training data and 30% as the testing data. Here are the modeling results we get.

We obtained a decent predicting performance along with a fair R squared score. We suggest the company adjust the monthly premium according to the predictive results from our classification model and customer lifetime value model above. For example, if a customer has a very high customer lifetime value, we suggest the company shrink the monthly premium so as to maintain a long-term business relationship.(See Figure 18)

Conclusion and Future Perspectives

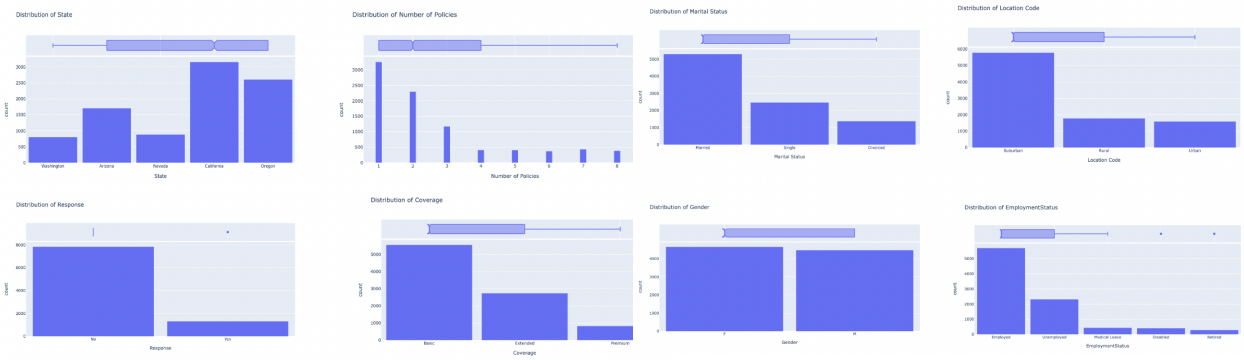
Based on the analysis we did in this project, we concluded our pricing strategy shortly in three steps: first, we classify customers into different coverage plans using our classification model; second, we predict a customer lifetime value based on our predictive model; finally we compute a predictive price for each customer. However, the final step will involve adjustments reflected by the results from the first two models.

From future perspectives, we would like to continue doing more model optimization using skills beyond this class and thus improve the performance of our predictions. Moreover, we look forward to involving corporate management in our strategy such as revenue management so that we can capture needs from the corporate angle. For example, if our company has a certain revenue goal, how should we adjust the model to guarantee hitting the goal while keeping the prices reasonable.

Appendix

Categorical Features	Numerical Features
State/State Abbreviations	Accident Severity in State
Response, Coverage	Number of Accidents in State
Education, Gender	Customer Lifetime Value
Employment Status	Income
Effective To Date	Monthly Premium Auto
Location Code	Months Since Last Claim
Marital Status	Months Since Policy Inception
Policy Type, Policy	Number of Open Complaints
Renew Offer Type, Sales Channel	Number of Policies
Vehicle Class, Vehicle Size	Total Claim Amount

Figure 1




Distribution of Renew Offer Type
Distribution of Sales Channel
Distribution of Months Since Last Claim
Distribution of Total Claim Amount
Distribution of Vehicle Class
Distribution of Policy Type


Distribution of Education
Bachelor
College
Master
High School or Below
Doctor
Education
count
Distribution of Vehicle Size
Medsize
Small
Large
Vehicle Size
count

Figure 2

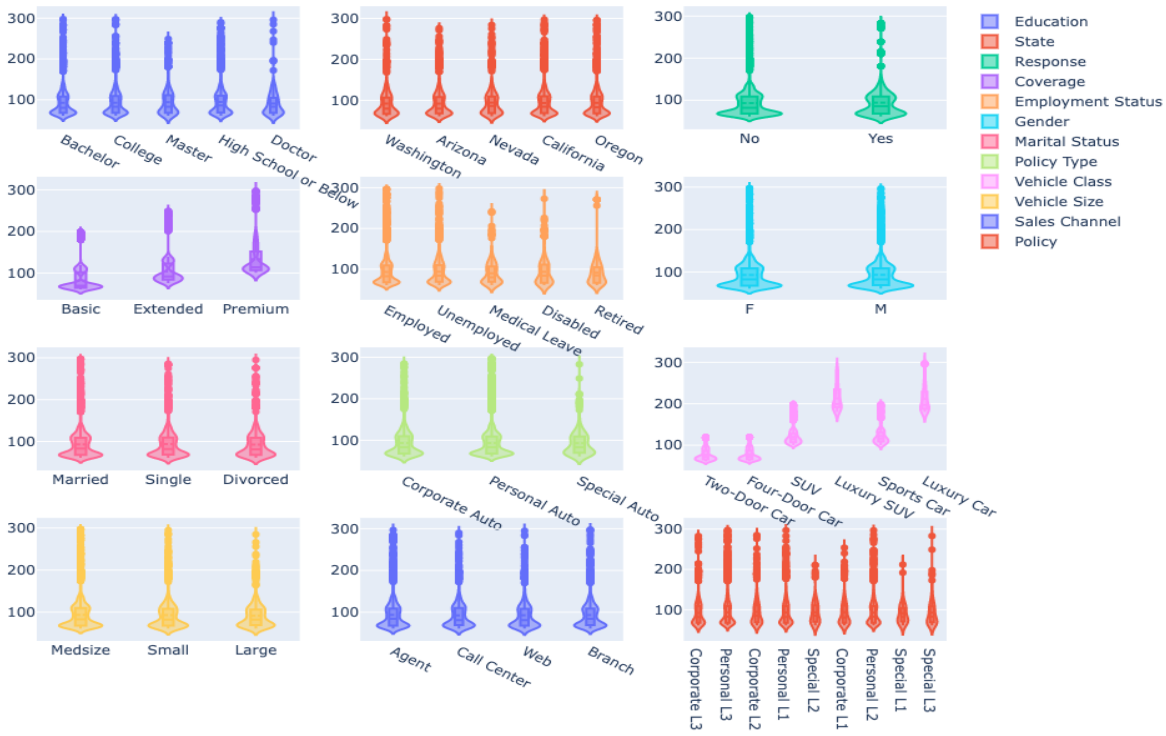


Figure 3

	Monthly Premium Auto
Accident Severity in State	-0.003687
Number of Accidents in State	0.006633
Customer Lifetime Value	0.396262
Income	-0.016665
Monthly Premium Auto	1.000000
Months Since Last Claim	0.005026
Months Since Policy Inception	0.020257
Number of Open Complaints	-0.013122
Number of Policies	-0.011233
Total Claim Amount	0.632017

Figure 4

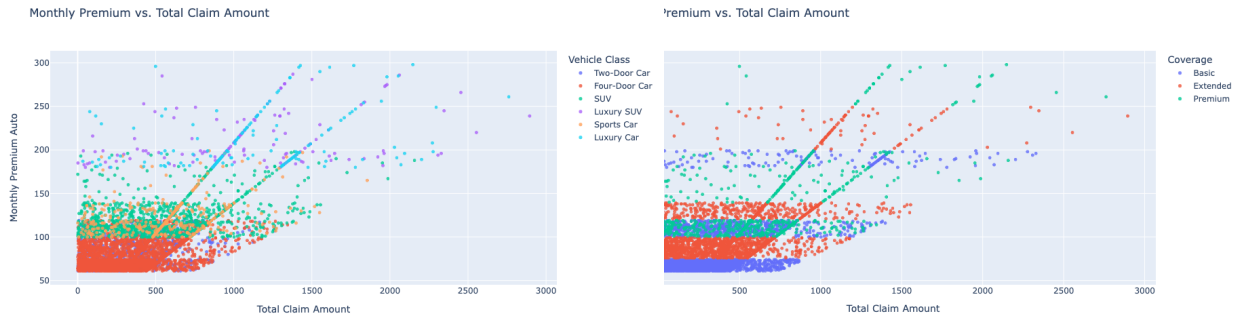


Figure 5

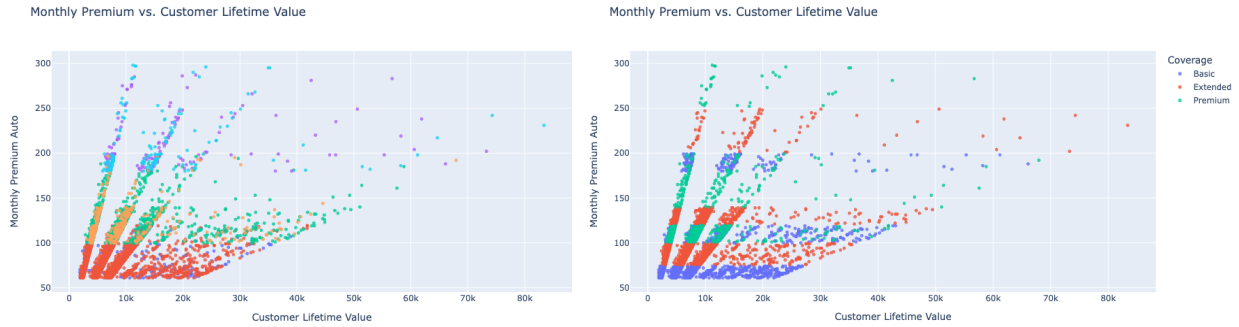


Figure 6

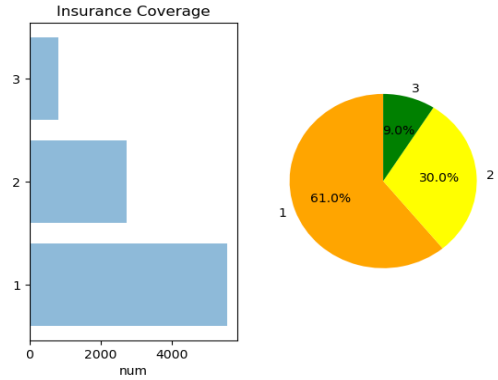


Figure 7

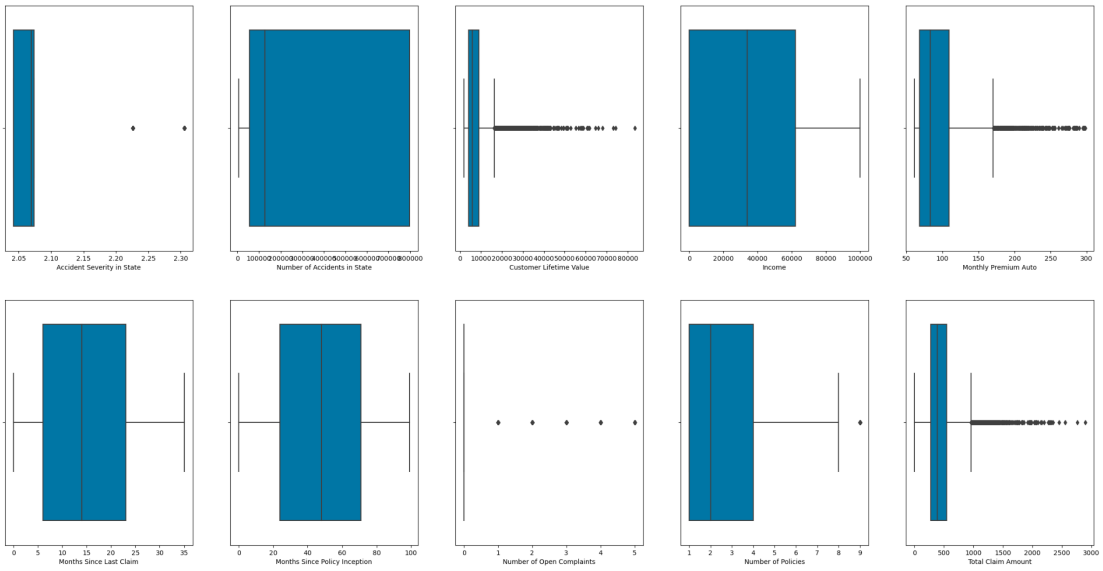


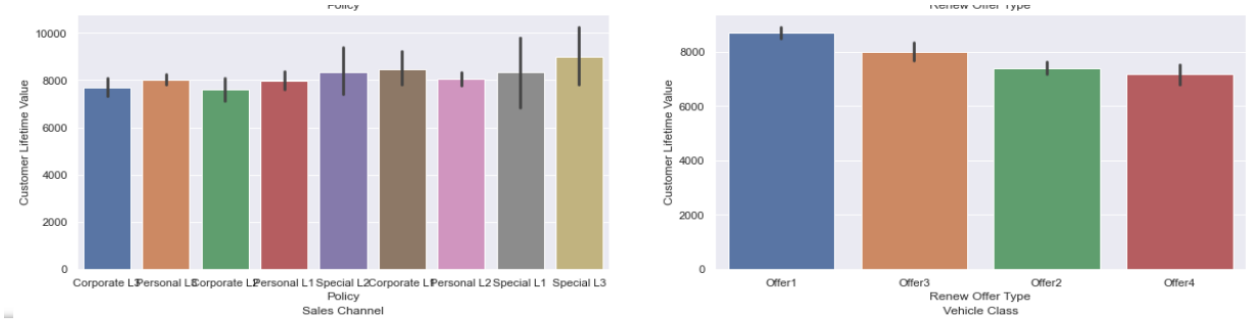
Figure 8

	Basic	Extend	Premium
Basic	1632	40	0
Extend	293	524	11
Premium	0	182	59

Figure 9

	Basic	Extend	Premium
Basic	1626	46	1
Extend	214	559	41
Premium	0	65	189

Figure 10



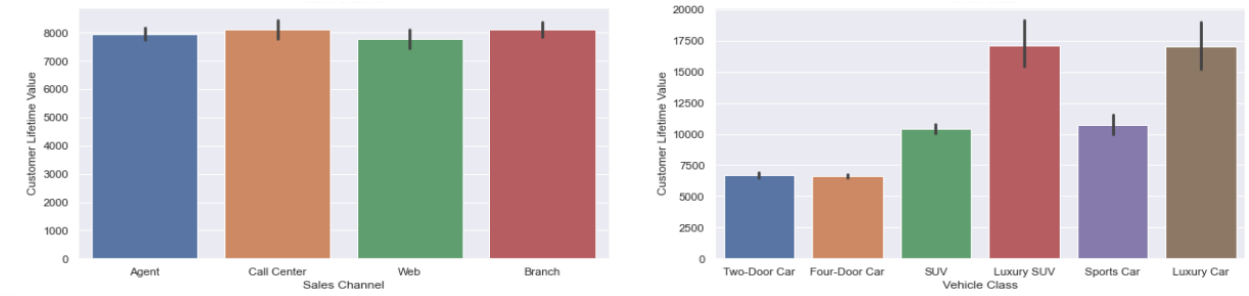


Figure 11

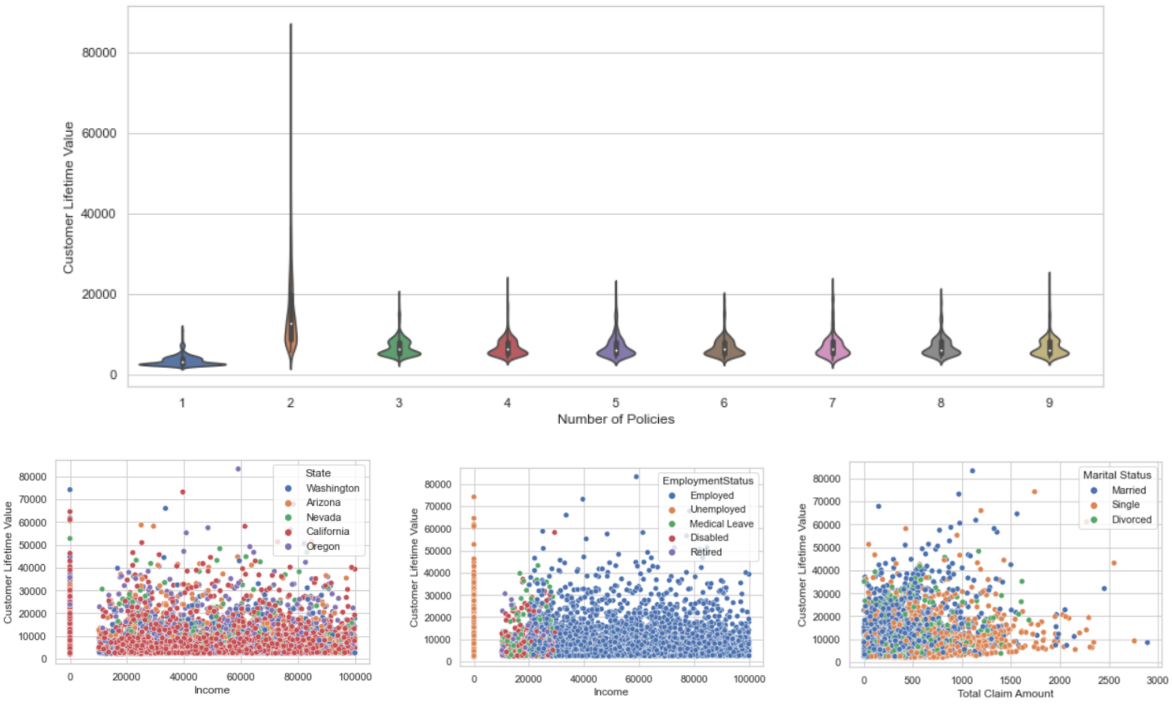
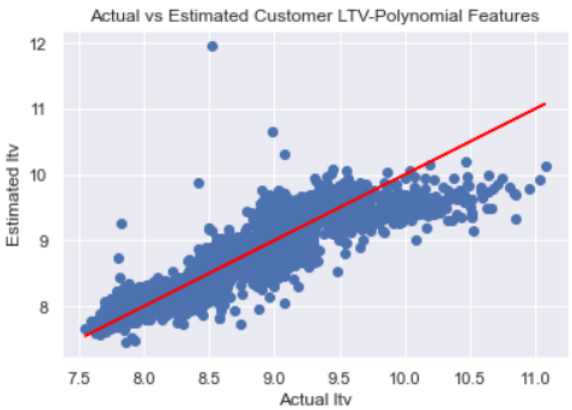


Figure 12

OLS Regression Results

Dep. Variable:	Customer Lifetime Value	**R-squared:**	0.782
Model:	OLS	**Adj. R-squared:**	0.776
Method:	Least Squares	**F-statistic:**	114.2
Date:	Mon, 05 Dec 2022	**Prob (F-statistic):**	0.00
Time:	14:10:14	**Log-Likelihood:**	-1569.9
No. Observations:	6850	**AIC:**	3560.
Df Residuals:	6640	**BIC:**	4995.
Df Model:	209		
Covariance Type:	nonrobust		




```
Mean Absolute Error (MAE)      : 0.24811304750055949
Mean Sq. Error (MSE)           : 0.1092379552169062
Root Mean Sq. Error (RMSE)     : 0.3305116567035211
Mean Abs. Perc. Error (MAPE)   : 2.771177647112457
```

Figure 13

Variable Name	Coefficient	p-value
Total Claim Amount	-0.1087	0
Total Claim Amount^2	0.0802	0
Marital Status_Married	0.084	0
Marital Status Single	0.0517	0.027
Number of Policies 2	0.2282	0
Number of Policies 3	0.1314	0
EmploymentStatus_Medical Leave	0.1302	0.007
EmploymentStatus_Retired	-0.2526	0.004
Accident Severity in State Total Claim Amount	-0.2119	0

Figure 14

Variable Name	Coefficient	p-value
Total Claim Amount	-0.1087	0
Total Claim Amount^2	0.0802	0
Marital Status_Married	0.084	0
Marital Status Single	0.0517	0.027
Number of Policies 2	0.2282	0
Number of Policies 3	0.1314	0
EmploymentStatus_Medical Leave	0.1302	0.007
EmploymentStatus_Retired	-0.2526	0.004
Accident Severity in State Total Claim Amount	-0.2119	0

Figure 15

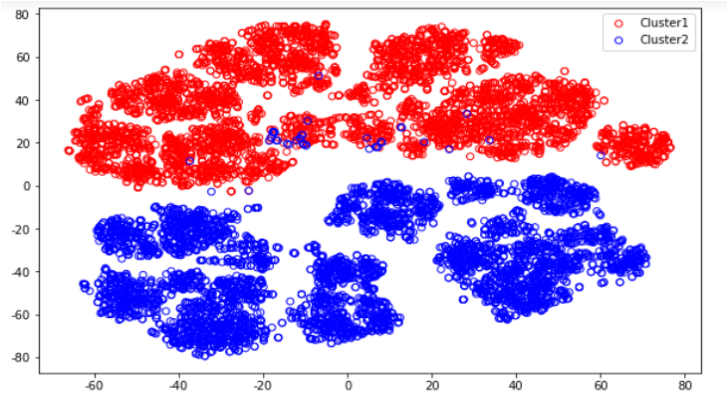


Figure 16

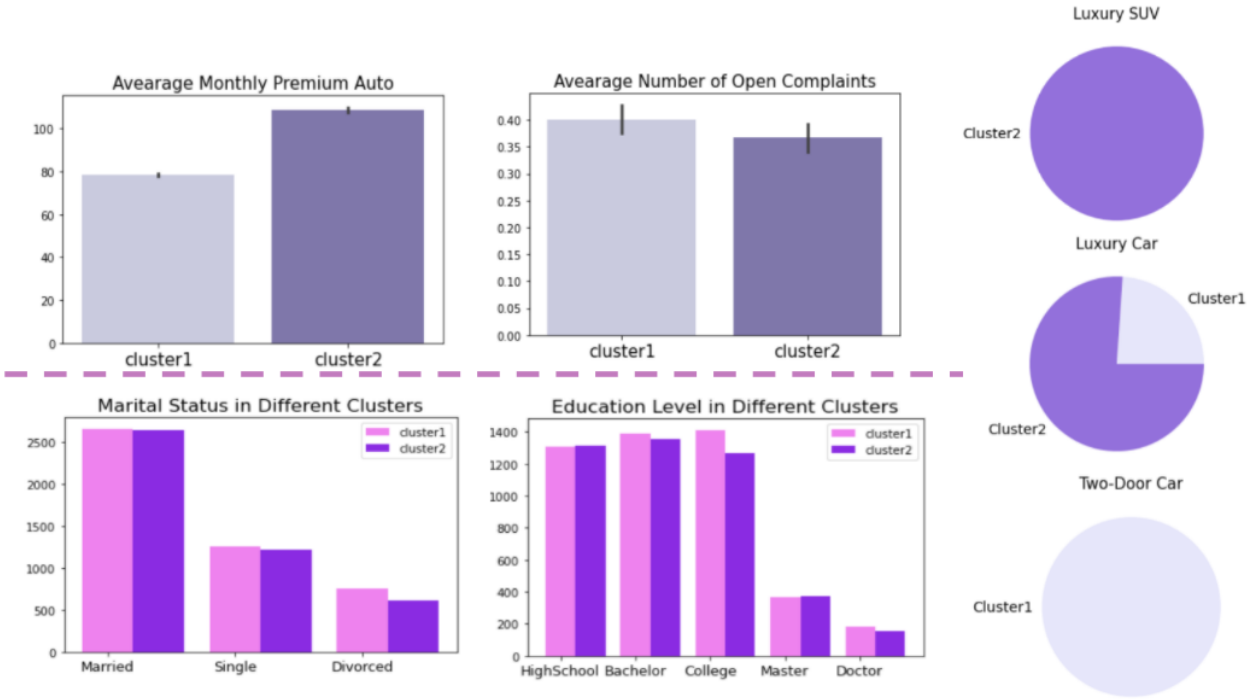


Figure 17

Mean Absolute Error	4.97
Mean Squared Error	43.02
R Squared Score	0.96

Figure 18