Customer Lifetime Value Prediction

by Sanana Alizada

December 4th, 2022

Mentor: AJ Sanchez, Phd

Overview

Customer Lifetime Value can be described as the amount of money customers spend on a company's product or services.

- **Stakeholders:** Sales & Marketing, Product Team, Website design and Engineering Team
- **Goal:** Predict Customer Lifetime Value to help business:
    - Better strategize
    - Improve customer retention
    - Increase Customer Lifetime Value

Data Acquisition and Exploration

Data has been acquired from the website kaggle.com. IBM Marketing Customer Value Analysis dataset was used. Dataset consists of both Categorical and Numerical features

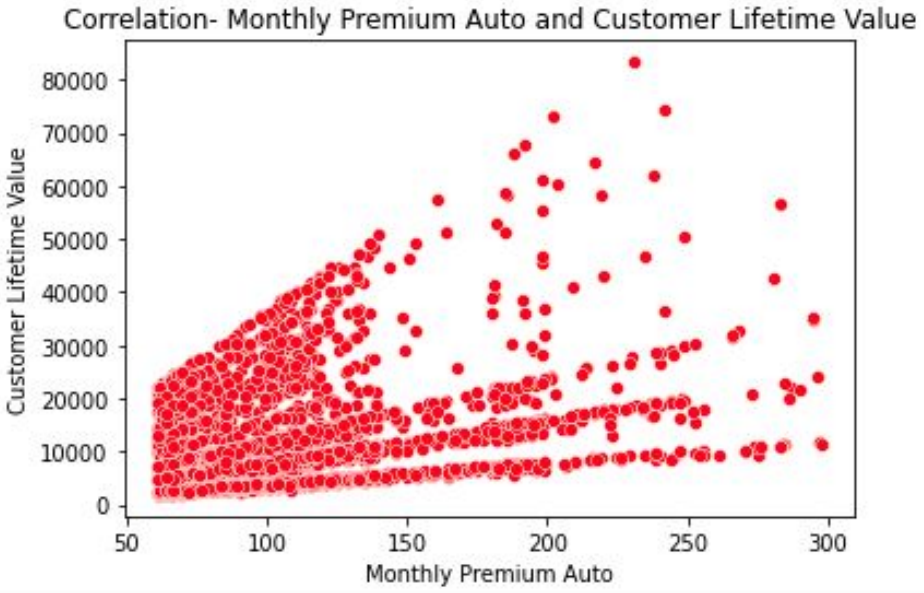
Customer Lifetime Value is the dependent variable which is being predicted. Firstly, we started exploring the correlations between Customer Lifetime Value and other variables.

Numerical/Continuous variables exploration

Customer Lifetime Value is positively correlated with **Monthly Premium Auto**(amount customer pays monthly to the company) and **Total Claim Amount**(higher the monthly premium amount customer pays, higher the claim amount is).

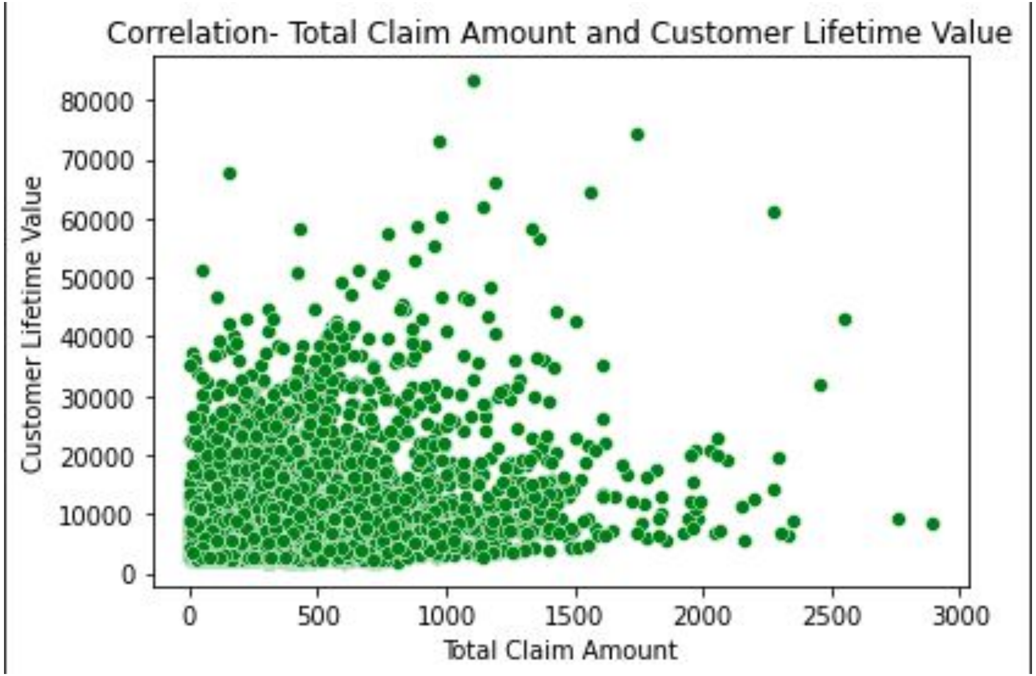
Monthly Premium Auto

- Maximum Monthly Premium Amount (MPA) is 298 and the minimum MPA is 61
- Mean of MPA is 93.21929 and the Median is 84.00
- The Standard Deviation is 34.40797
- Skewness is positive



Total Claim Amount

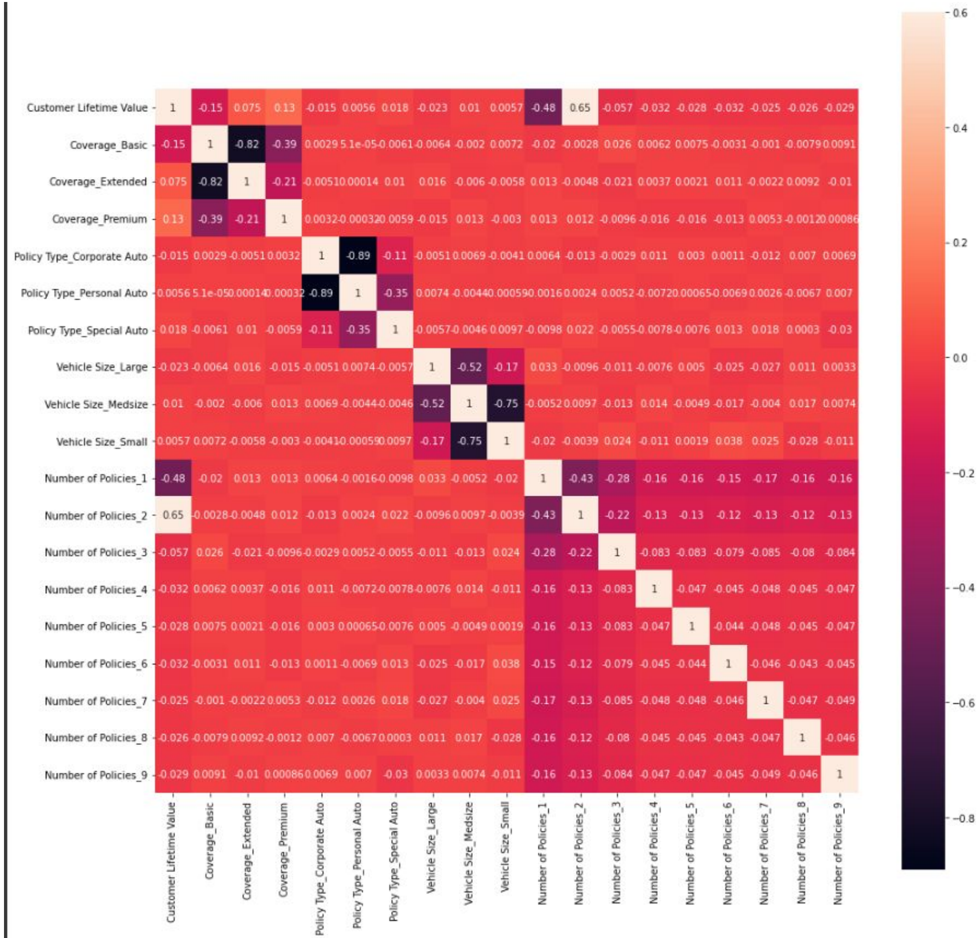
- Maximum Total Claim Amount (TCA) is **293** and the minimum TCA is **0.099007**
- Mean of TCA **434.088794** and the Median is **383.945434**
- The Standard Deviation is **290.500092**
- Skewness is positive
- Not as skewed or as long tailed as Monthly Premium Amount



Categorical Variables Analysis

According to our heatmap from different policy numbers(levels), Policy number 1 and 2 is positively correlated with Customer Lifetime Value. While the majority of the variables are positively correlated with our target variable, it looks like the type of policy of different customers has a cardinal impact on the Customer Lifetime Value.

This exploration helped us in the modeling stage to determine the most important variables to keep, since there was a risk where model accuracy would be affected by having too many dummy variables (we used both Linear and Logistic Regression algorithms, but more on them in the modeling part).



Categorical Variables Analysis

- **State**: California residents are the most valuable customers in comparison to other states
- **Response**: Customers responded "No" to the marketing initiatives are the most valuable to the company
- **Coverage**: Basic coverage is the most chosen option, therefore most valuable
- **Education**: Level of education does not affect the lifetime value of the customer
- **Employment status**: Employed customers are the most valuable to the company
- **Gender**: Gender has no impact on Customer LifeTime value
- **Marital status**: Married customers are the most valuable to the company
- **Policy Number**: Customer with Policy number 2 are the most valuable to the company
- **Policy Type**: Personal policies are the most valuable
- **Renew Offer Type**: Customers preferring Offer 1 are the most valuable ones to the company
- **Sales Channel**: Call Center is not performing well compared to other channels throughout the country (in terms of high value customers). While customers utilize Agent as a channel are the most valuable to the company.
- **Vehicle Class**: Customers having Four-Door Car are the most valuable
- **Vehicle Size**: MidSize vehicle owners are the most valuable customers

Baseline Modeling

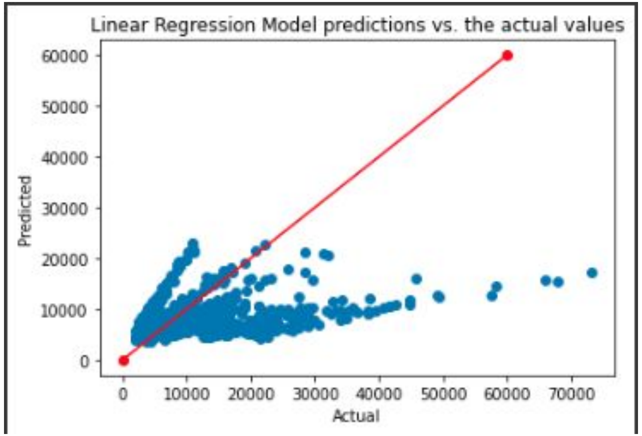
Train Data: 0.8
Test Data: 0.2

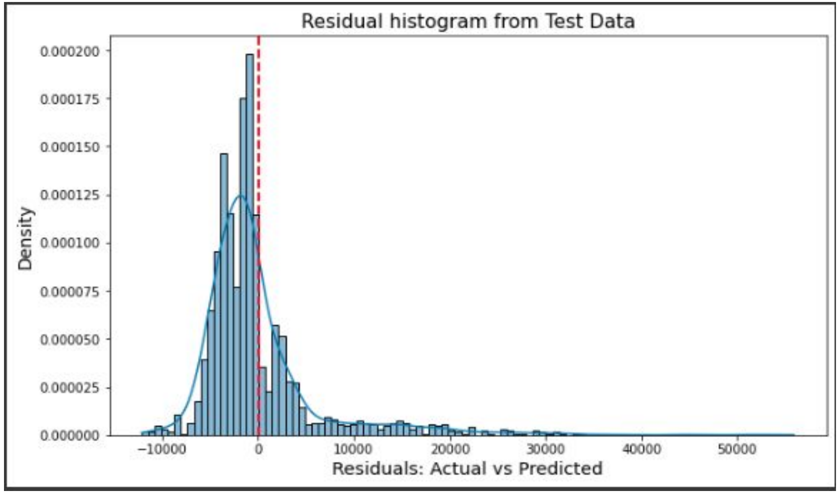
Linear Regression model was considered as a baseline model.

We used R2 and Mean Absolute Percentage Error for model evaluation and for the initial Linear Regression model:

- **R2= 0.16**
- **MAPE= 61.41**

<u>We can see that residuals are not normally distributed and residuals are skewed and there are extreme outliers.</u>

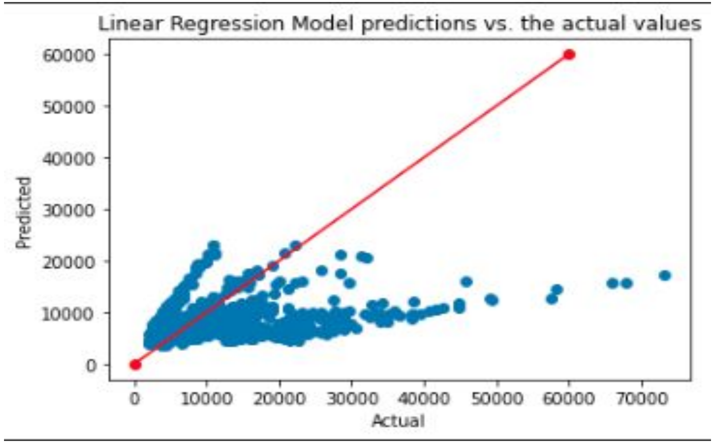


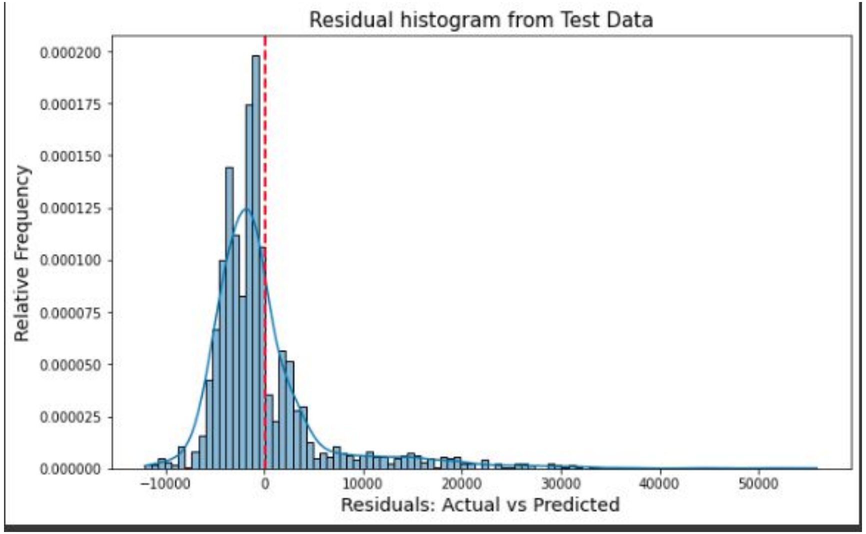


Extended Modeling

Linear Regression with fewer features

- **R2= 0.16**
- **MAPE= 61.47**



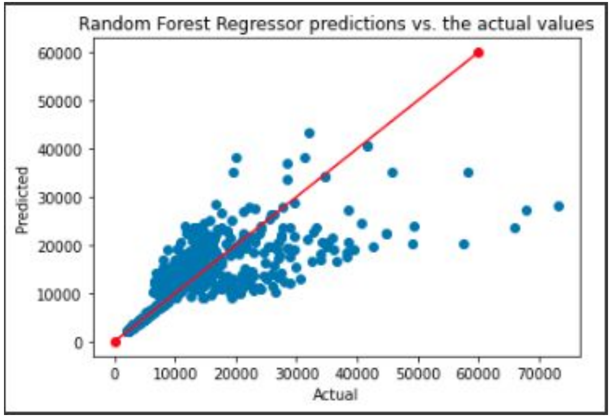


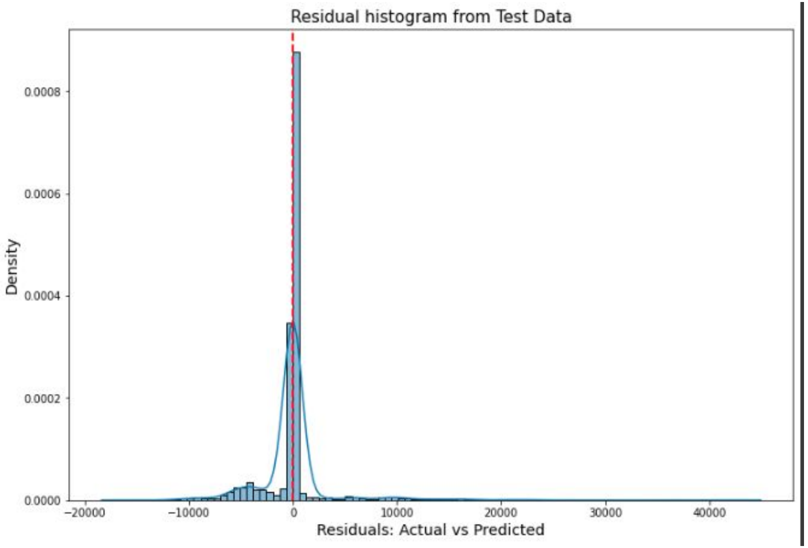
<u>Residuals are not normally distributed and residuals are skewed and there are extreme outliers.</u>

Extended Modeling

Random Forest Regressor

- **R2= 0.69**
- **MAPE= 10.19**





<u>Residuals are more normally distributed and with less outliers.</u>

Findings

Random Forest Regressor is the best model out of all models that have been tested:

- Its nonlinear nature made it a most compatible option.
- 0.69 R2 is a pretty good score, meaning 69% of the variance can be explained.
- MAPE value is 10% means that the average difference between the forecasted value and the actual value is 10% of the actual value, which is considered an acceptable accuracy.

Model	R-square	MAPE
Linear Regression (with all features)	0.16	61.41%
Linear Regression (with fewer features)	0.16	61.47%
Random Forest Regressor	0.69	10.19%
Random Forest Regressor (tuned)	0.68	11.02%

Future Work

As a future work, I would take below actions:

- I would test more algorithms to find out if there is any other algorithm that is more accurately predicting CLTV;
- I would also use classification algorithms to determine characteristics of customer with different levels of CLTV;
- I would use more than 1 dataset to increase the complexity and to multi dimensionalize the results, optimize the model performance.

Recommendations

- Give priority to ultimate customer service and to resolving the complaints in a timely manner; otherwise, it could decrease the Customer Lifetime Value;
- Targeting married employed customers might be an effective way to increase CLTV
- Creating a promotional campaign specific to California residents with midsize vehicles also might cause an increase in CLTV.