SPRINGBOARD DATA SCIENCE
CAPSTONE PROJECT 2

Customer Lifetime Value Prediction

by Sanana Alizada

December 4th, 2022

Table of Contents

1. Introduction

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

It is essential to know customer's Customer Lifetime Value for several reasons:

- *To know how much the company should spend on customer acquisition;*

- *To identify customers with high Customer Lifetime Value and study them. This would create an opportunity for businesses to set a certain group of customers as a main target and build strategies around their behaviors to increase customer retention as well as profit margin;*

- *To build a customer-centered business model.*

Stakeholders are mainly the company's Sales, Marketing, and Customer Relationship Management teams. The need is relevant to them, because of the goals such as increasing sales, keeping the most profitable customers, customer retention, and establishing an effective platform. Sales & Marketing Team can use the results to build new Marketing Strategies, promotional campaigns or loyalty programs after knowing the interactions of the existing and potential most profitable customers. Operations Team can improve the efficiency of customer service. The Technical Team can use the results to focus on building a more user-friendly online platform.

Taking the above mentioned points into consideration, being able to accurately predict a customer's Customer Lifetime Value would be absolutely important for the business success. In this project we utilized publicly available customer data (IBM Watson Marketing Customer Value Analysis dataset)[1] and explored different machine learning models to accurately predict the Customer Lifetime Value.

Extensive implementation details can be found in this IPython notebook[2].

[1] IBM Watson Marketing Customer Value Data

[2] IPython Notebook

2. Approach

2.1. Data Acquisition and Wrangling

Data has been acquired from the website kaggle.com. Kaggle has various real world datasets readily available to download and utilize. For this project, IBM Marketing Customer Value Analysis dataset was used. More details about acquiring the data can be found in this **IPython notebook[3].**

There are 9134 rows and 24 columns in our dataset. There are non-categorical and categorical data.

Categorical data: Customer, State, Response, Coverage, Education, Effective To Date, EmploymentStatus, Gender, Location Code, Marital Status, Policy Type, Policy, Renew Offer Type, Sales Channel, Vehicle Class, Vehicle Size

Non-Categorical data: Customer Lifetime value, Income, Monthly Premium Auto, Months Since Last Claim, Months Since Policy Inception, Number of Open Complaints, Number of Policies, Total Claim Amount.

While preparing the data for analysis, we made sure that the both categorical and numerical variables are being correctly classified. Not every numerical value represents non-categorical variables and vice versa.

Duplicate rows, null values have been checked and dataset updated accordingly. Some of the columns that do not provide any meaningful information or have an enormous impact on model accuracy were removed in the later stages of the process. The cleaned dataset was then used for explorations.

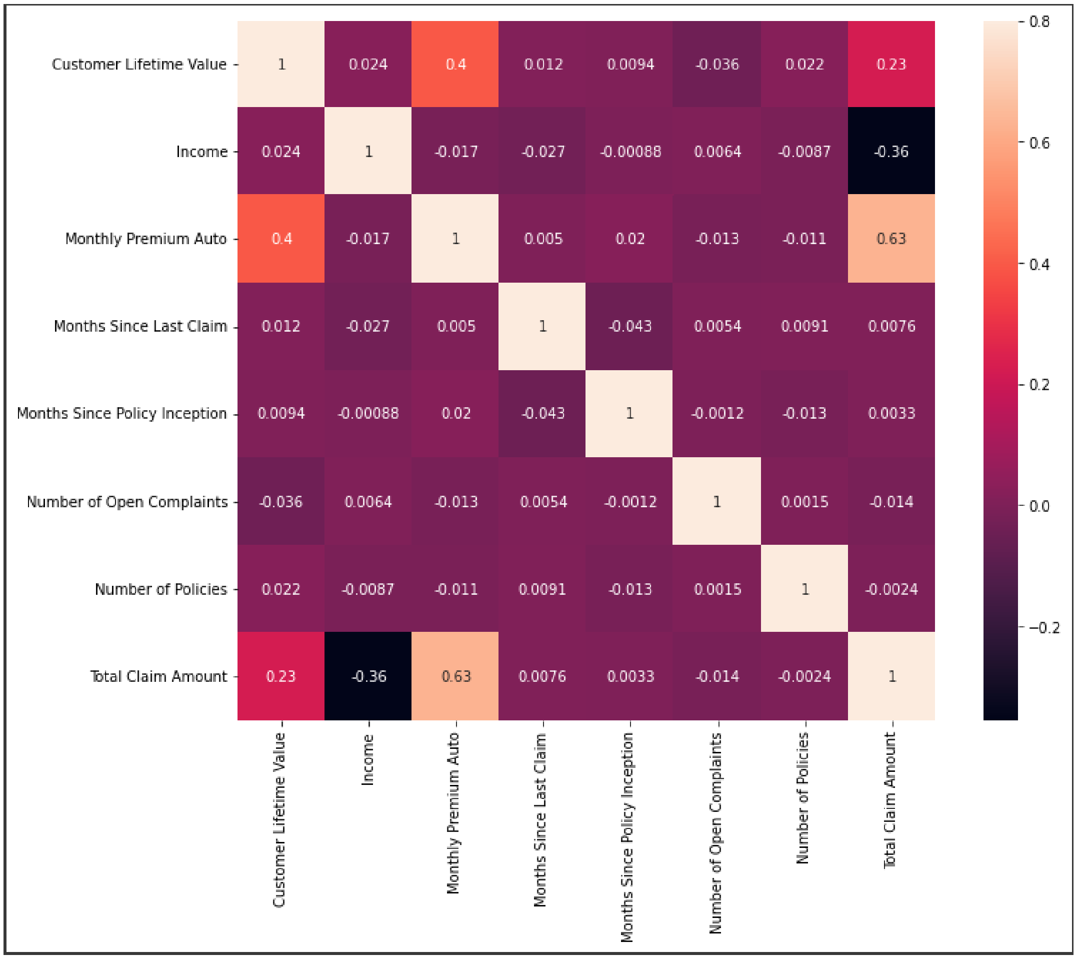
[3] IPython Notebook

2.2. Storytelling and Inferential Statistics

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

2.2.1. Numerical/Continuous variables exploration

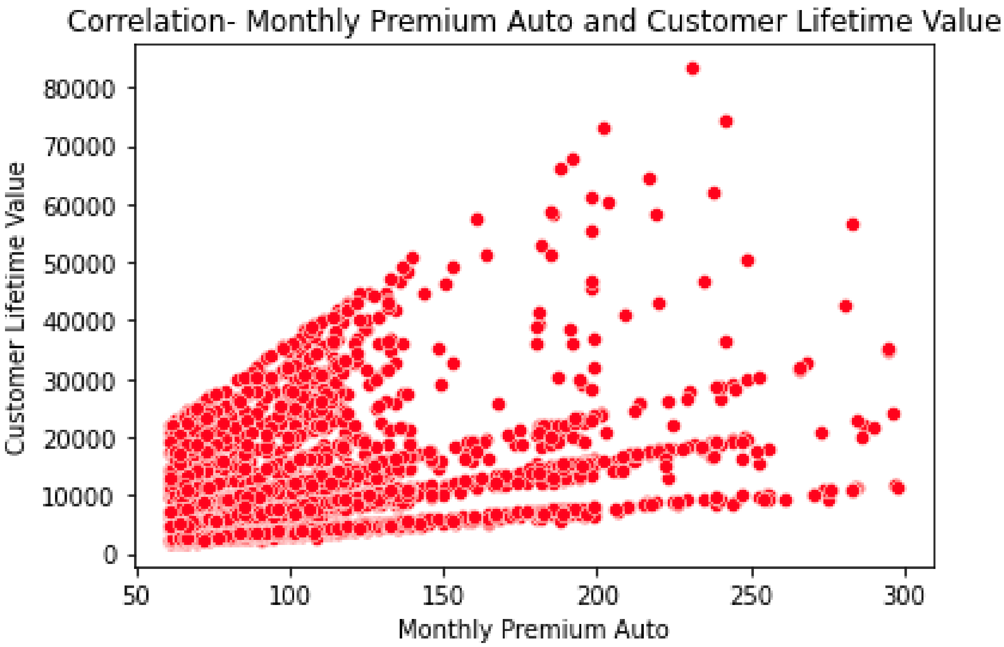
Considering this dataset consists of both numerical and categorical variables, looking into them separately was considered a more effective option. We utilized a heatmap to have a deeper understanding of the pairwise correlation among numerical variables and each with **Customer Lifetime Value:**



Based on the heatmap above, as a result, we found that specific to the dataset, 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).

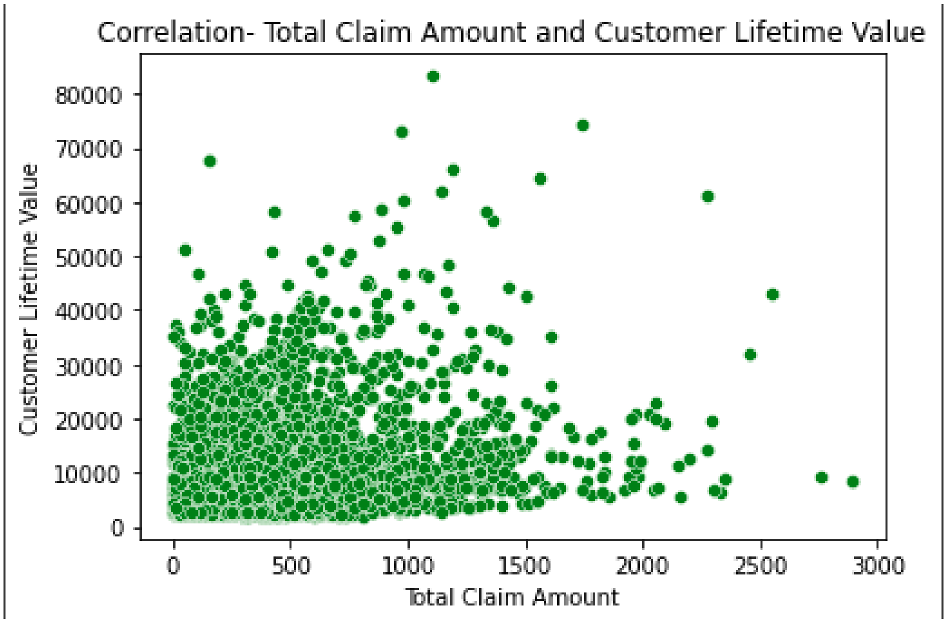
Monthly Premium Auto

- 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


Correlation- Monthly Premium Auto and Customer Lifetime Value

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

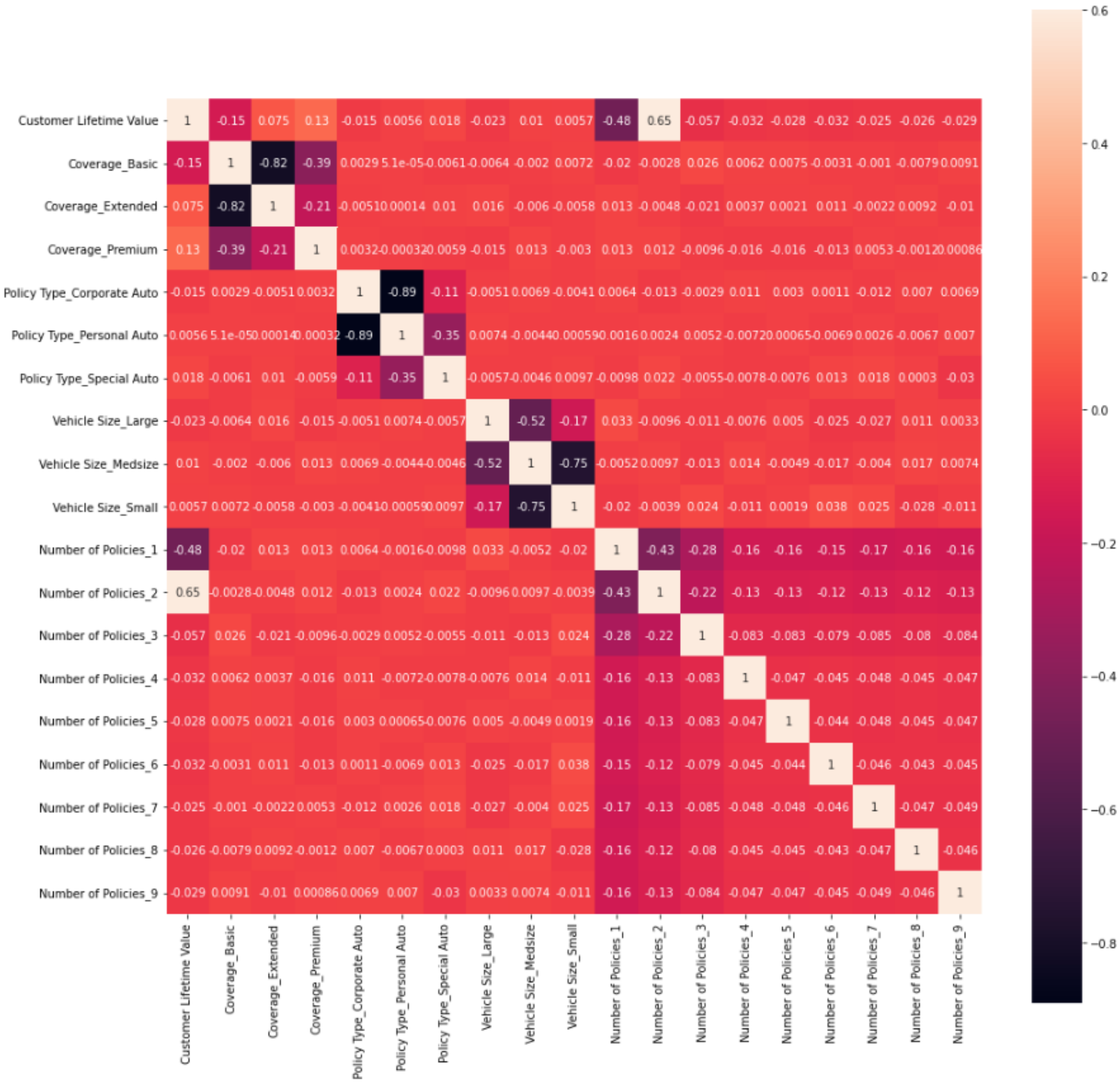


2.2.2. Categorical variables analysis

We utilized bar charts to demonstrate the relationship between Customer Lifetime Value (CLTV) and each categorical variable. Categorical variables are mostly the customer characteristics or descriptions.

- **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**: MedSize vehicle owners are the most valuable customers

We also converted categorical variables to dummies to build a heatmap to explore the correlation between each feature and Customer Lifetime Value.



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). Refer to this **notebook** for further details[4].

3. Baseline Modeling

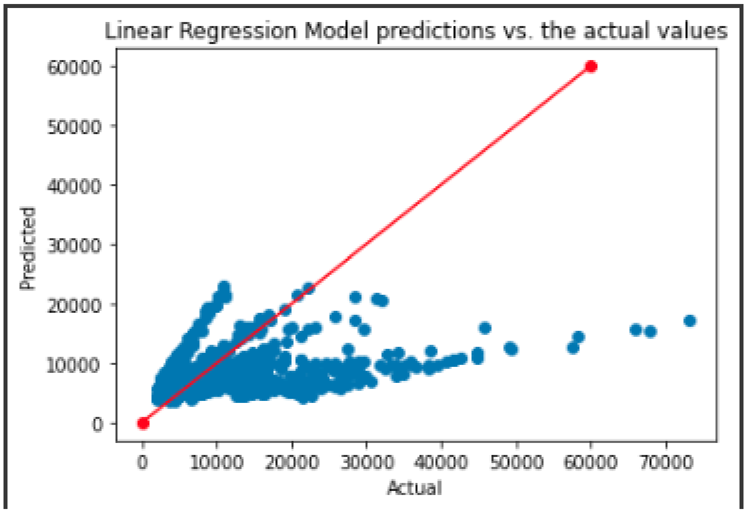
Before starting to decide and utilize different models, we prepared the dataset for the modeling stage. This process included, converting categorical values to dummy variables, to do the train (80%) and test (20%) split on dataset.

For this project, both Linear and Random Forest Regression algorithms were utilized. Linear Regression model was considered as a baseline model, where all features were considered.

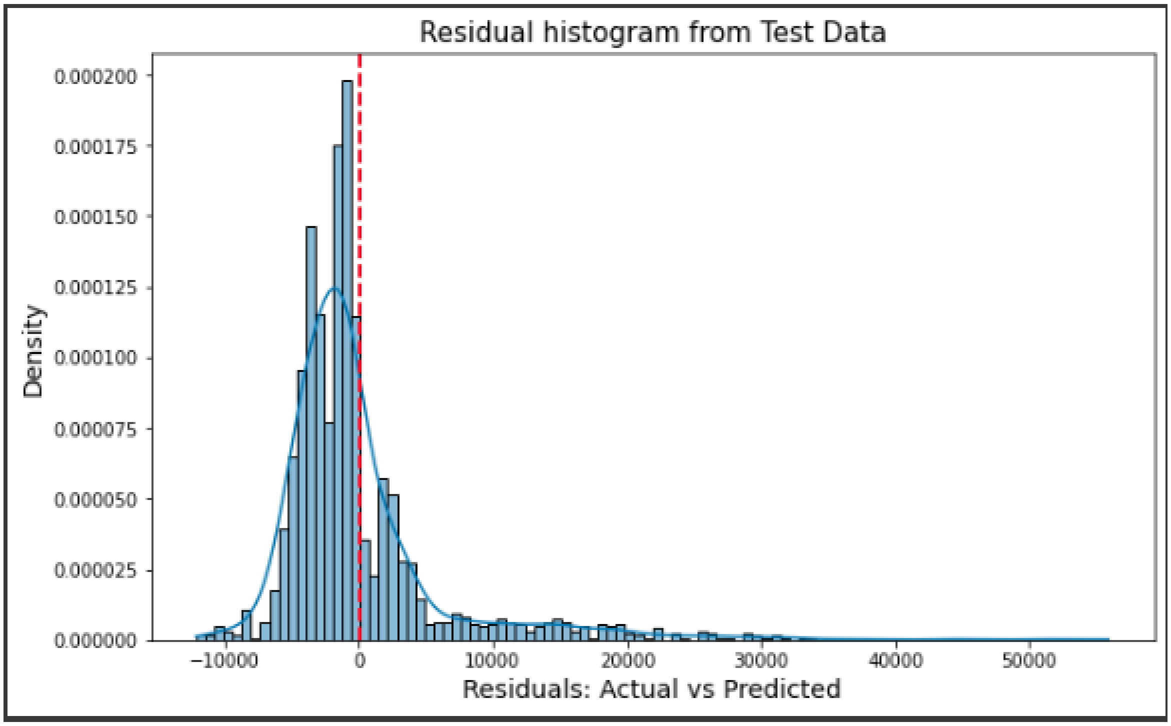
We used R2 (R-Squared) and Mean Absolute Percentage Error (MAPE) for model evaluation, and for the initial Linear Regression model the specific values are:

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

When we look at the Actual vs Predicted values of this model in both scatter plot and in residual histogram from test data:



[4] IPython Notebook



We definitely see that the residuals (and hence the error terms) are not normally distributed. On the contrary, the distribution of the residuals is quite skewed and there are extreme outliers. Several nuances in the dataset might have caused this inaccuracy. For instance, we might end up having too many variables after converting categorical variables to dummy variables. Also, we must take into the consideration that the Linear regression performs poorly when there are non-linear relationships.

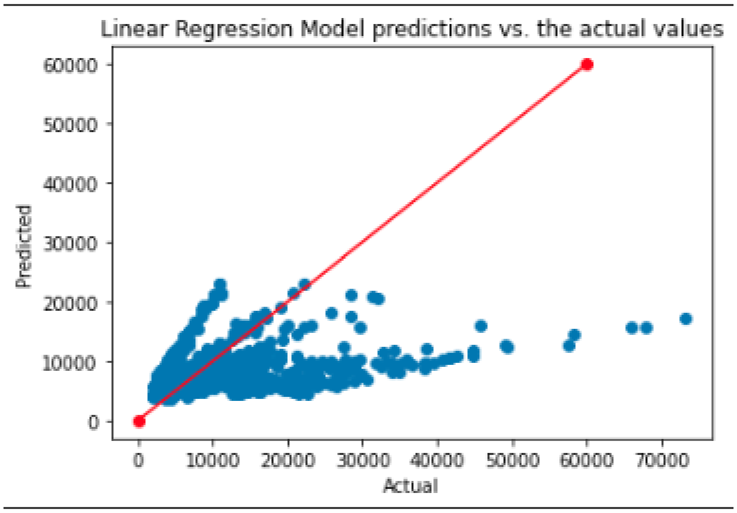
4. Extended Modeling

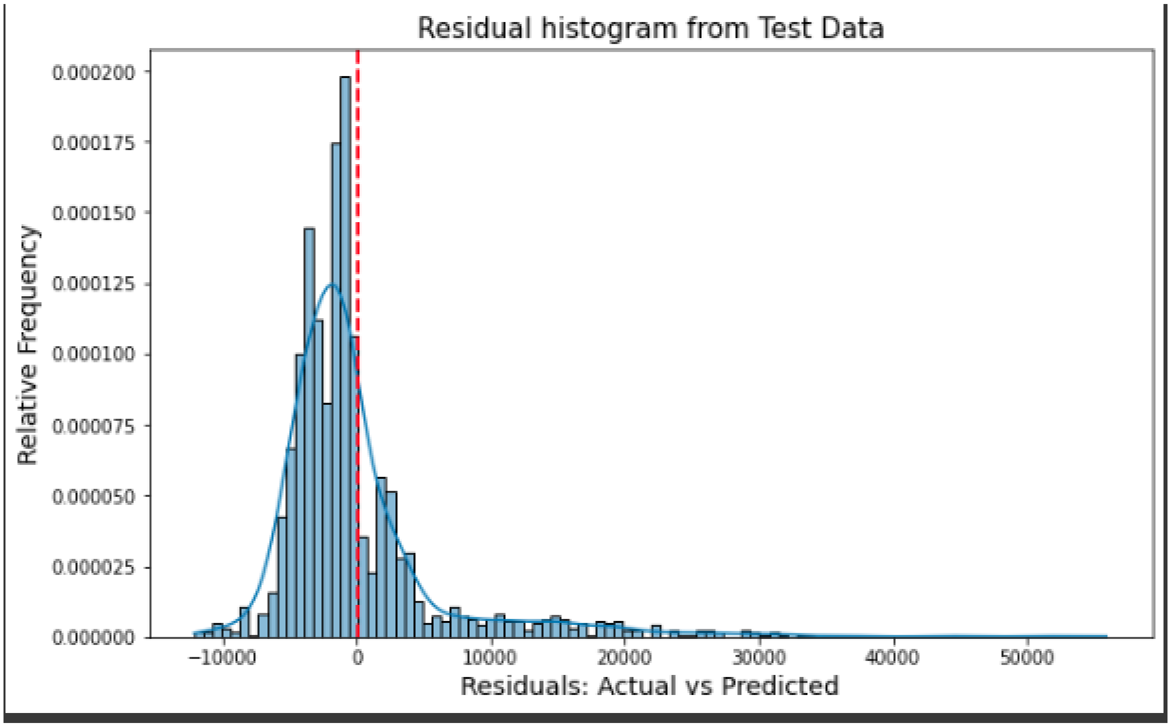
After careful consideration, we dropped some unnecessary variables to try to increase the model accuracy as well as overall performance. For more details about the variables have been dropped and the reasoning, please refer to this ***notebook.***

As a first extended model, Linear Regression was utilized again, since this time we had fewer features. R2 and MAPE are still used as evaluation metrics:

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

If we look at the scatter plot of Actual vs Predicted values as well as histogram of residuals:



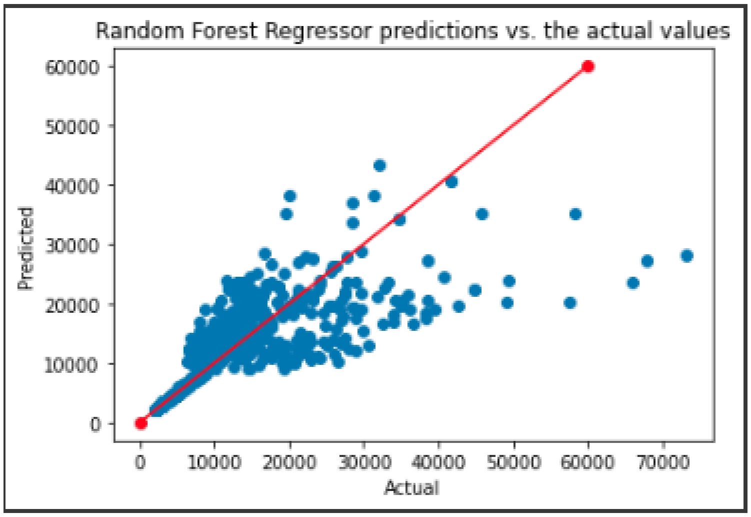


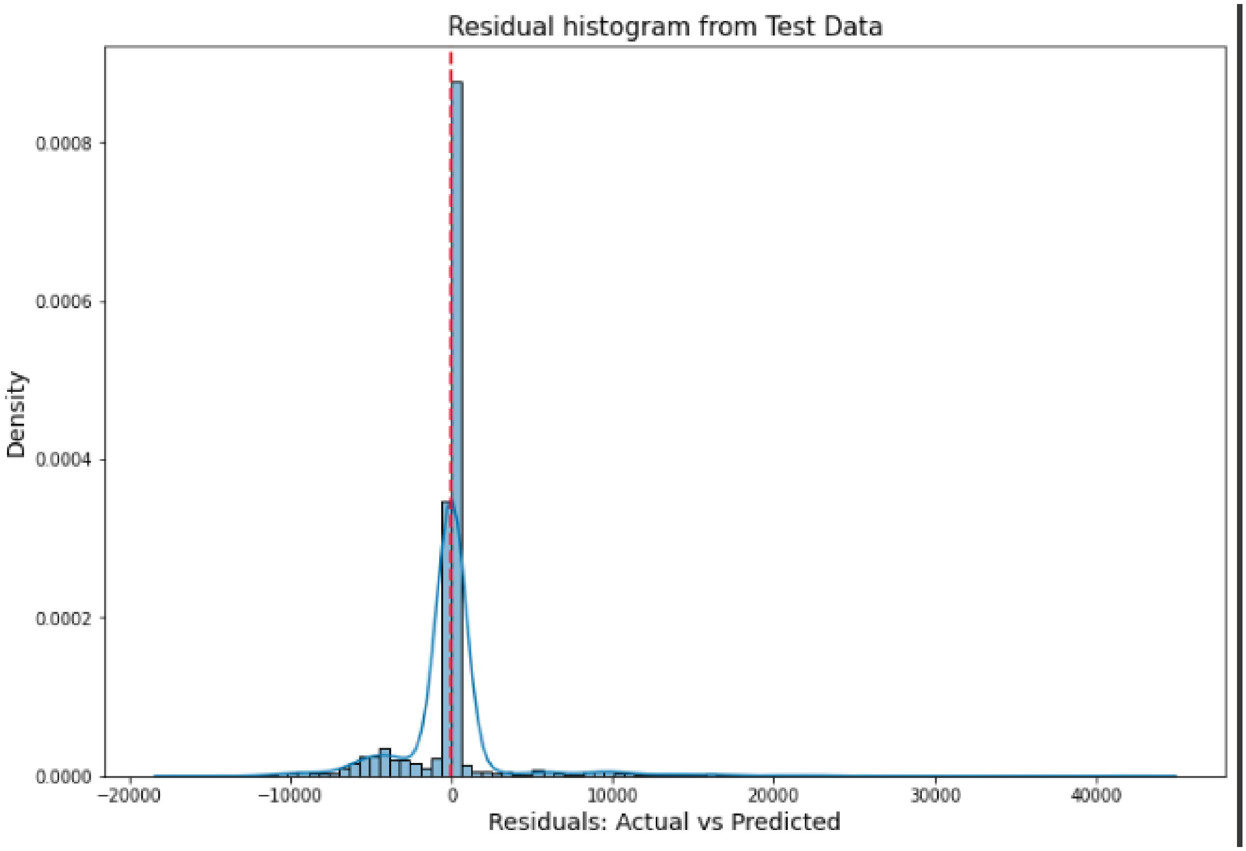
We can see that residuals are not normally distributed and residuals are skewed and there are extreme outliers.

As a second extended model, we utilized the Random Forest Regressor. Random Forest's nonlinear nature can give it a leg up over linear algorithms, making it a great option.

R2 and MAPE were used as evaluation metrics:

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





Per above, residuals are more normally distributed and with less outliers. Random Forest performed well with this dataset. Please refer to ***this*** notebook for more details and Random Forest Tuning.

5. Findings

We have created 3 models to determine the best model with highest R-square and lowest Mean Absolute Percentage Error.

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%

Taking the table above into consideration, Random Forest Regressor is the best model out of all models that have been tested. As mentioned before, Random Forest's nonlinear nature made it a most compatible option, and looking at the constraints of the dataset, 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.

Circling back to the original business problem, after analysis and prediction of Customer Lifetime Value, companies will have a better understanding of the characteristics of customers with high CLTV. This will play a significant role in marketing and product development, as well as, potentially increase the profit margin for the companies.

6. Conclusion and Future work

We originally wanted to be able to accurately predict Customer Lifetime Value for the businesses to have better strategies for the business and product development. Keeping that in mind, we tried to find the most suitable and accurate Machine Learning Algorithm to predict the Customer Lifetime Value by using different Regression algorithms. As a result, we discovered that the

Random Forest Regressor was the most accurate model for the prediction of the Customer Lifetime Value.

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.

7. Recommendations for the Clients

When we look into the feature importance according to the Random Forest Regressor, we do see that the Number of open complaints have a significant impact on CLTV:

Therefore:

- Give priority to ultimate customer service and to resolving the complaints in a timely manner; otherwise, it could decrease the Customer Lifetime Value;
- Monthly Premium and Number of Policies (especially 1 and 2) are the ultimate factors increasing CLTV;
- 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.

8. Consulted Resources

Datasource:

- kaggle.com[5]

Python libraries:

- Numpy[6];

[5] Kaggle Dataset

[6] NumPy Documentation

- Pandas[7];
- Scikit-learn[8];
- Seaborn[9];
- Matplotlib[10]

Books:

- "The Art of Statistics: How to Learn from Data" by David Spiegelhalter(2019).

[7] Pandas Documentation
[8] Scikit-learn Documentation
[9] Seaborn Documentation
[10] Matplotlib Documentation