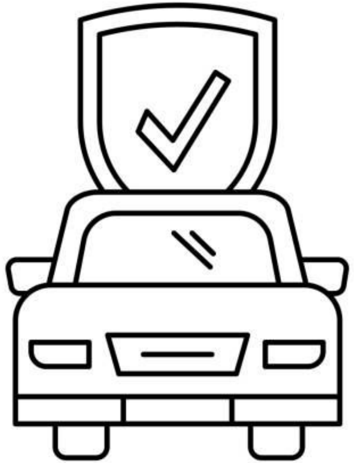
Collecting Analyzing Large Data - MGTA 452

Customer Lifetime Value Prediction for SureDrive Auto



Srihari Nair

Agenda

1. Problem Statement
2. Exploratory Data Analysis
3. Feature Engineering
4. Model Building & Evaluation
5. Model Interpretation & Selection
6. Results & Conclusion

Problem Statement:

Prediction of the <u>Customer Lifetime Value</u> (CLV) for an Auto Insurance company

SureDrive Auto Insurance has observed a ***decline in customer retention*** over the past few months.

The company expects us to Predict CLV for future customers based on a dataset of their existing customers

Why?

- Understanding CLV is crucial for businesses as it guides their investment strategies in customer acquisition and retention.

- This analysis will enable our client to create targeted strategies to enhance customer engagement, retention, and drive growth.

Data Source - https://www.kaggle.com/code/dktalaicha/predict-customer-life-time-value-clv

Exploratory Data Analysis

```
Data columns (total 24 columns):
 #   Column                         Non-Null Count  Dtype  
---  ------                         --------------  -----  
 0   Customer                       9134 non-null   object 
 1   State                          9134 non-null   object 
 2   Customer Lifetime Value        9134 non-null   float64
 3   Response                       9134 non-null   object 
 4   Coverage                       9134 non-null   object 
 5   Education                      9134 non-null   object 
 6   Effective To Date              9134 non-null   object 
 7   EmploymentStatus               9134 non-null   object 
 8   Gender                         9134 non-null   object 
 9   Income                         9134 non-null   int64  
 10  Location Code                  9134 non-null   object 
 11  Marital Status                 9134 non-null   object 
 12  Monthly Premium Auto           9134 non-null   int64  
 13  Months Since Last Claim        9134 non-null   int64  
 14  Months Since Policy Inception  9134 non-null   int64  
 15  Number of Open Complaints      9134 non-null   int64  
 16  Number of Policies             9134 non-null   int64  
 17  Policy Type                    9134 non-null   object 
 18  Policy                         9134 non-null   object 
 19  Renew Offer Type               9134 non-null   object 
 20  Sales Channel                  9134 non-null   object 
 21  Total Claim Amount             9134 non-null   float64
 22  Vehicle Class                  9134 non-null   object 
 23  Vehicle Size                   9134 non-null   object 
dtypes: float64(2), int64(6), object(16)
```

- **Key Variables:** State, Coverage, Education, Number of Policies, Employment Status, Income, Monthly Premium, Policy details, Total Claim Amount, and more.

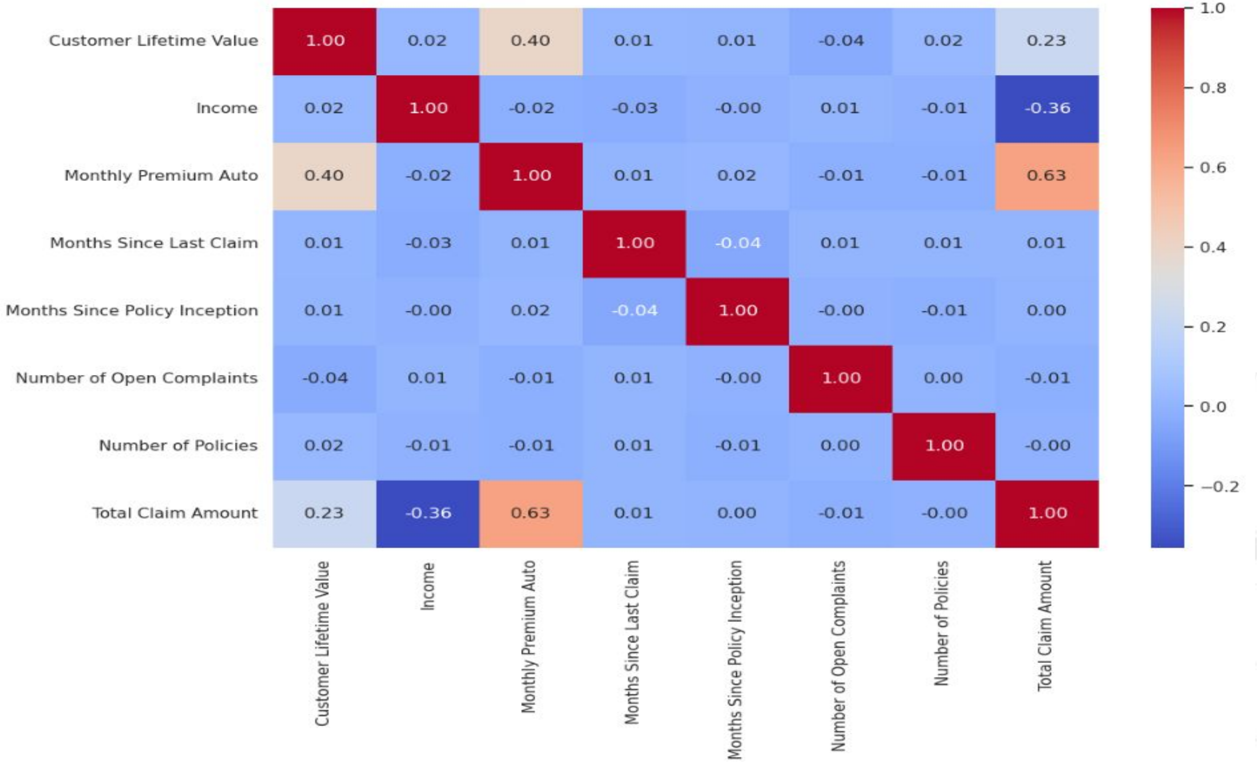
- **Continuous Variables:** Income, Monthly Premium Auto, Months Since Last Claim, Number of Policies, Total Claim Amount, etc.

- **Data Quality:** No null values, ensuring data integrity and reliability.

- **Dataset Size and Diversity:** 9,134 observations with 24 variables, including a mix of categorical and continuous data.

Exploratory Data Analysis

Correlation Matrix



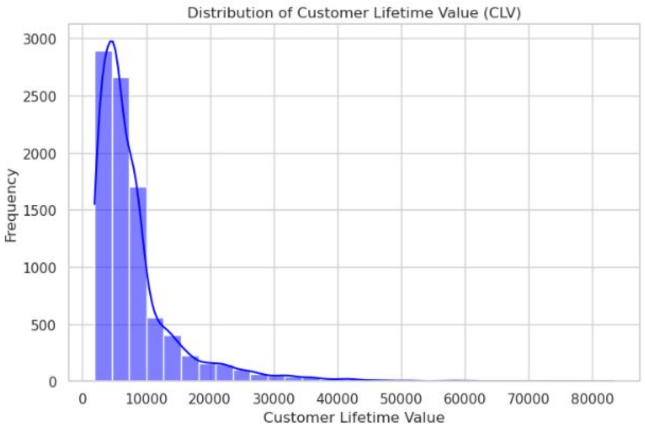
Strong Positive Correlation:

- **‘Total Claim Amount**' & '**Monthly Premium Auto**'
- '**Total Claim Amount**' & '**Number of Policies**'

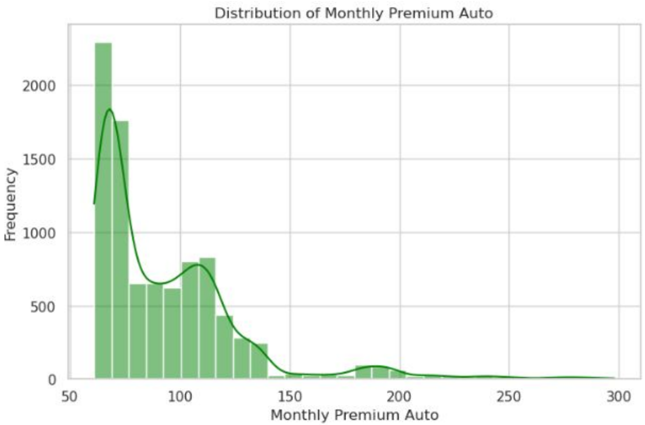
Negative Correlation:

- '**Income**' & '**Months Since Last Claim**'
- ‘**Income**' & '**Monthly Premium Auto**’

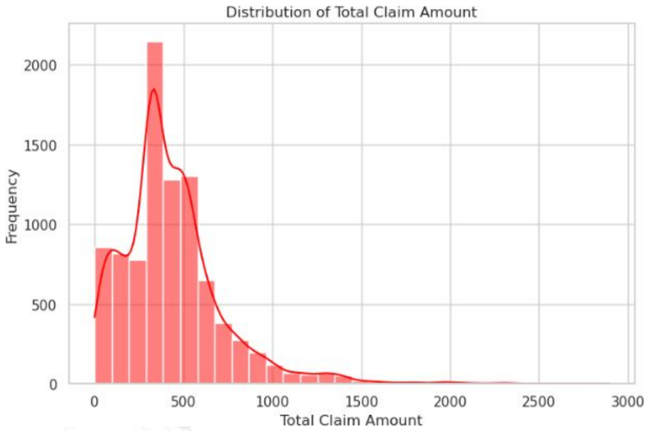
Exploratory Data Analysis



CLV: Skewed right



Monthly Premium: Right-skewed



Total Claim: Right-skewed low values

Income: Moderately right-skewed, excludes zero

Exploratory Data Analysis

CLV vs Monthly Premium

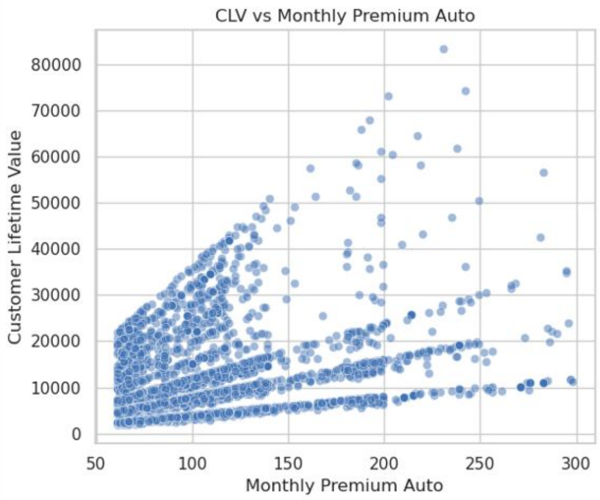
Monthly Premium is directly proportional to CLV.
Variability in CLV grows with higher premiums.

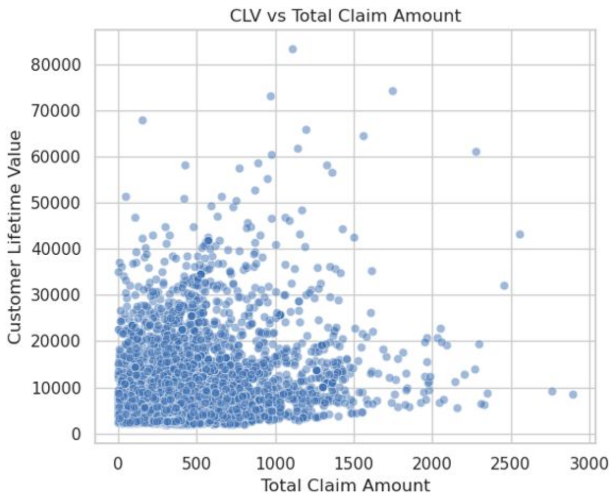
CLV vs Total Claim Amount

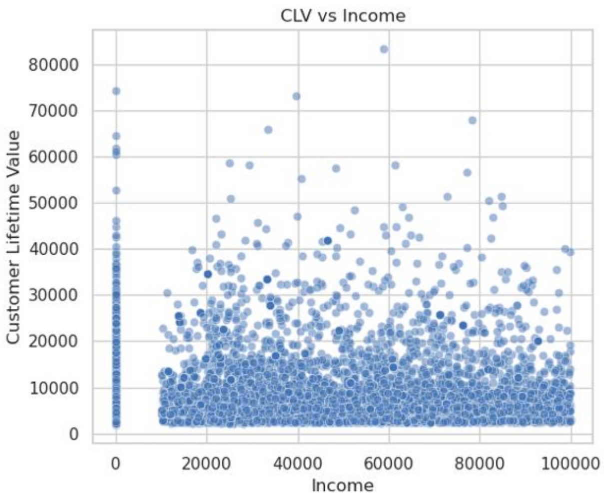
Total Claim Amount is almost directly proportional to CLV.
Multiple outliers exist for this correlation

CLV vs Income

No clear relationship between CLV & Income
CLV varies widely across all income levels







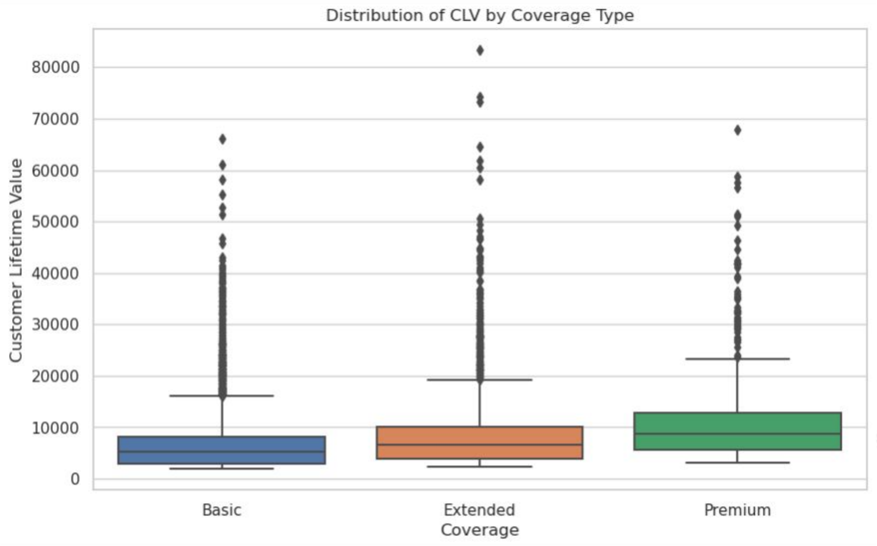
Exploratory Data Analysis

CLV Distribution Across Coverage Types

"Basic": Lowest CLV

"Extended": Higher median

"Premium": Widest range, most variability.

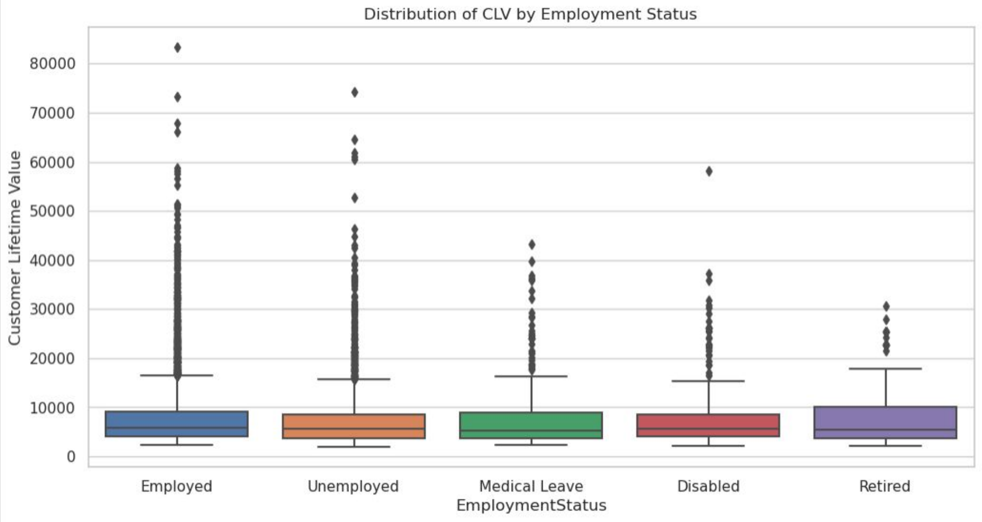


CLV Variation by Employment Status

"Employed": Stable CLV range,

"Unemployed": Higher median CLV,

"Retired": Broad, varied CLV distribution.

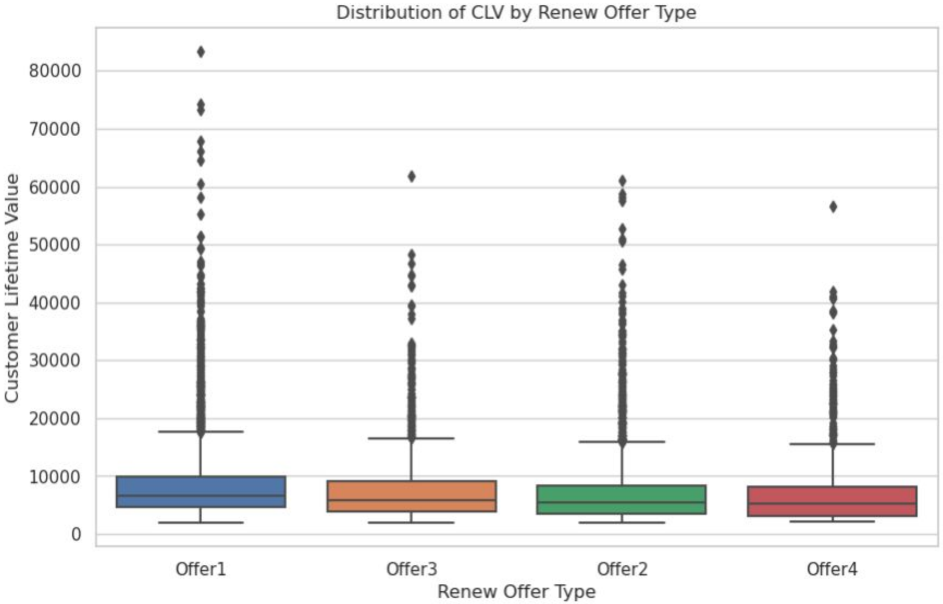


Exploratory Data Analysis

CLV Distribution by Renew Offer Type

"Offer1 & Offer2": Higher median CLV

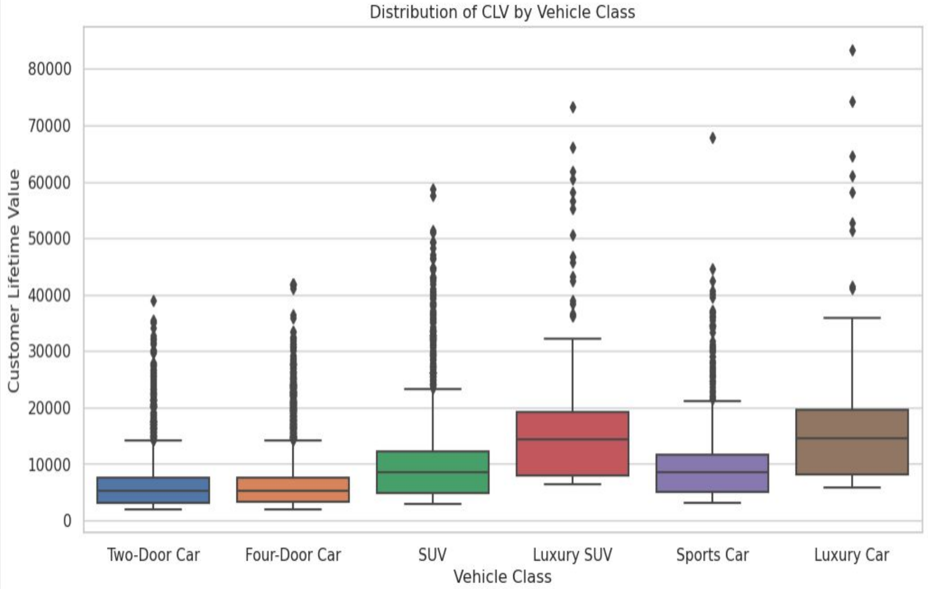
"Offer3 & Offer4": Lower Indicates retention effectiveness



CLV Distribution by Vehicle Class

Luxury Cars: Higher CLVs

Two-Door, Four-Door Cars: Lower CLVs



Feature Engineering

Data Cleaning for Model Training

Dropping redundant columns : Removing unnecessary or duplicate columns from the dataset to simplify the model and improve performance.

- **Dropped Date variables,** since we are not performing time series analysis
- **Dropped Customer No. & Names,** since it was not adding value to features

One Hot Encoding : Converting categorical variables into a binary (True or False) matrix to allow for proper analysis in machine learning models.

```
# One-hot encoding of categorical variables
data_encoded = pd.get_dummies(data, drop_first=True)

# Displaying the first few rows of the encoded data
data_encoded.head()
```

Number of Policies	Total Claim Amount	Customer_AA11235	Customer_AA16582	...	Sales Channel_Branch	Sales Channel_Call Center	Sales Channel_Web	Vehicle Class_Luxury Car
1	384.811147	False	False	...	False	False	False	False
8	1131.464935	False	False	...	False	False	False	False
2	566.472247	False	False	...	False	False	False	False
7	529.881344	False	False	...	False	True	False	False

Feature Engineering

Data Cleaning for Model Training

Interaction Matrix : Created to capture the combined effect of two or more variables on the dependent variable, an effect that is not simply additive

```
# Creating interaction features between Income and Coverage
for coverage_type in ['Coverage_Extended', 'Coverage_Premium']:
    interaction_feature_name = f'Income_{coverage_type}'
    data_encoded[interaction_feature_name] = data_encoded['Income'] * data_encoded[coverage_type]

# Displaying the first few rows of the updated dataset
data_encoded[['Income', 'Coverage_Extended', 'Coverage_Premium', 'Income_Coverage_Extended', 'Income_Coverage_Pre
```

	Income	Coverage_Extended	Coverage_Premium	Income_Coverage_Extended	Income_Coverage_Premium
0	56274	False	False	0	0
1	0	True	False	0	0
2	48767	False	True	0	48767
3	0	False	False	0	0
4	43836	False	False	0	0

Feature Engineering

Feature Selection

- **Importance Ranking:** We leveraged Random Forest to rank features based on their importance, helping us identify key predictors for our model.
- **Dimensionality Reduction:** The initial feature selection, is reducing dimensionality and simplifying our model.
- **Model-based Feature Engineering:** Random Forest has guided our model-based feature engineering, especially in creating polynomial and interaction features.

Feature	Importance
Number of Policies	0.466433
Monthly Premium Auto	0.253064
Months Since Last Claim	0.043022
Total Claim Amount	0.037320
Months Since Policy Inception	0.035491
Income	0.027280
Income_Coverage_Extended	0.011561
Education_High School or Below	0.005807
Number of Open Complaints	0.005530
Sales Channel_Branch	0.005120
Gender_M	0.004979
Renew Offer Type_Offer2	0.004785
Location Code_Urban	0.004395
Marital Status_Married	0.004184
Education_College	0.004168
Response_Yes	0.004063
Vehicle Size_Medsize	0.003873
Policy_Personal L2	0.003830
Sales Channel_Web	0.003735
Education_Master	0.003654

***Importance** tells us what is the percentage contribution of each of our variables to a unit change in the dependent variable
For our model building we considered variables only with a **threshold of 0.02,** these variables explained around **80%+ of variability in Y**

Model Building

Initial Model - Linear Regression

We have run a **Linear Regression** model using variables with **6 highest importance scores** as our **explanatory variables**, making predictions on our **dependent variable, i.e. - Customer Lifetime Value**

Method used -

- Employed a systematic iteration of feature combinations to optimize model accuracy.
- Selection criteria focused on minimizing Mean Absolute Error (MAE) and maximizing adjusted R-squared value.

Model	MAE	Adjusted R2
model_Monthly Premium Auto_Total Claim Amount	3983.10	0.155
model_Monthly Premium Auto_Total Claim Amount_Customer Lifetime Value	3988.12	0.153
model_Monthly Premium Auto_Total Claim Amount_Months Since Policy Inception	3983.91	0.149

Model Building

Final Model - Random Forest

Why we chose Random Forest to improve our model?

- **Handles Overfitting Well:** Reduces overfitting through averaging multiple decision trees.
- **Works with Categorical and Numerical Data:** Effectively processes both types of data without extensive preprocessing.
- **Robust to Outliers and Non-linear Data:** Performs well with datasets that have outliers or non-linear relationships.
- **No Need for Feature Scaling:** Eliminates the need for input feature normalization or standardization.

Model	MAE	Adjusted R2
model_Number of Policies_Monthly Premium Auto_Income	1529.11	0.649
model_Number of Policies_Monthly Premium Auto_Total Claim Amount	1636.99	0.635
model_Number of Policies_Monthly Premium Auto_Months Since Policy Inception	1639.74	0.621

Model Interpretation & Selection

Model - 1

X- Feature: Number of Policies, Monthly Premium, **Income**

Y- Feature: Customer Lifetime Value

Metric	Value
Mean Absolute Error	**1528.18**
Adjusted R2	**0.649**

Model - 2

X- Feature: Number of Policies, Monthly Premium, **Total Claim Amount**

Y- Feature: Customer Lifetime Value

Metric	Value
Mean Absolute Error	**1637.66**
Adjusted R2	**0.635**

Model - 3

X- Feature: Number of Policies, Monthly Premium, **Months Since Policy Inception**

Y- Feature: Customer Lifetime Value

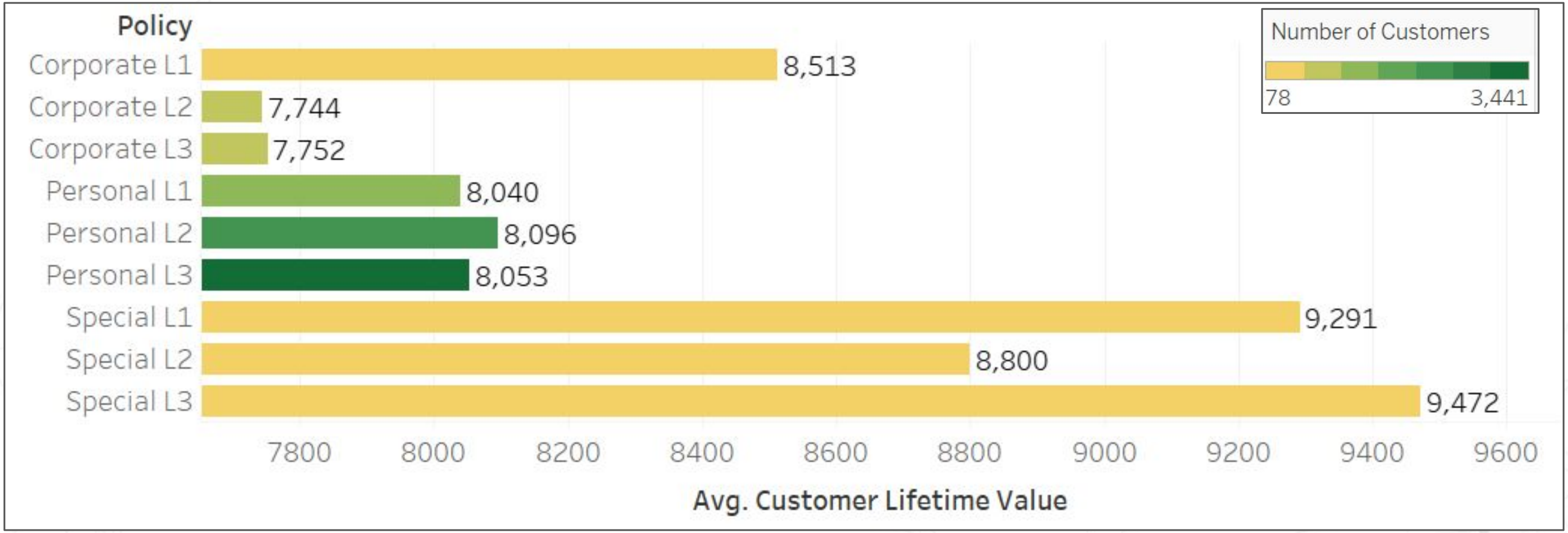
Metric	Value
Mean Absolute Error	**1634.70**
Adjusted R2	**0.621**

SureDrive Auto has received Customer Lifetime Value predictions for 100 prospects, derived from three distinct models

The selection of an appropriate model by SureDrive Auto will be guided by their specific business objectives and contextual considerations.

Results & Conclusion

- **Our analysis reveals that while 'Special' policy types have a smaller customer base, they consistently yield higher Customer Lifetime Values (CLVs).**
- **In contrast, 'Personal' policies, despite being the most popular, exhibit lower customer retention rates.**



Length of Bar - Average Customer Lifetime Value

Thank You

Group I