GAICO

Customer Retention Strategy Optimization with CLV Prediction

JCDSJKTAM-006 | M. Ariq Fauzan

1

Introduction

Background and business problem.

2

Data & Analytical Foundation

Data understanding and EDA

3

Modeling & Evaluation

Modeling process and performance

4

Business Impact & Insights

Business Impact and key insights

5

Recommendation

For marketing and analytics department

GAICO

About GAICO

To date, it is the leading american vehicle insurance company.

In 2019, it has 15 million auto policies in force and insures more than 20 million vehicles, holding 12.8% of the auto insurance market in America. It is committed to serve and meet the need of drivers.

Business Problem

Insurance companies have very limited budget for marketing that must be optimized towards maintaining valuable customers.

Stakeholder: GAICO Marketing Strategy Team
My Role: GAICO Analytics Team

Goals

Determine customer's lifetime value using their historical profile data to identify high-value customers.

Objectives

1. Develop a machine learning model that accurately predicts CLV.
2. Discover the most significant feature that affect CLV.

GAICO

Executive Summary

Accurate Model Successfully Developed

A tuned **Gradient Boosting** model was able to predict the CLV with great accuracy and time **(MAPE 7,84% at 0.02s)**

Features that affects CLV

Strong positive relationship is found between customer financial strength and product engagement.

(Number of policy, premium, and customer income.)

Marketing Strategy

Model will help increase retention ROI through better customer prioritization.

Data Understanding

Feature	Description
Vehicle Class	Type of vehicle insured
Coverage	Coverage tiers ("Basic", "Extended", "Premium")
Renew Offer Type	Type of renewal offer presented.
Employment Status	Current employment status
Marital Status	Current marital Status
Education	Latest educational background
Number of Policies	Number of active insurance policies the customer holds
Monthly Premium	Monthly premium amount in US$
Total Claim Amount	Amount of claims made in US$
Income	Customer's annual income in US$
Customer Lifetime Value	Projected total revenue a customer will bring in USD.

- 5669 rows of auto-insurance data from 2019.
- One row represents one unique customer.

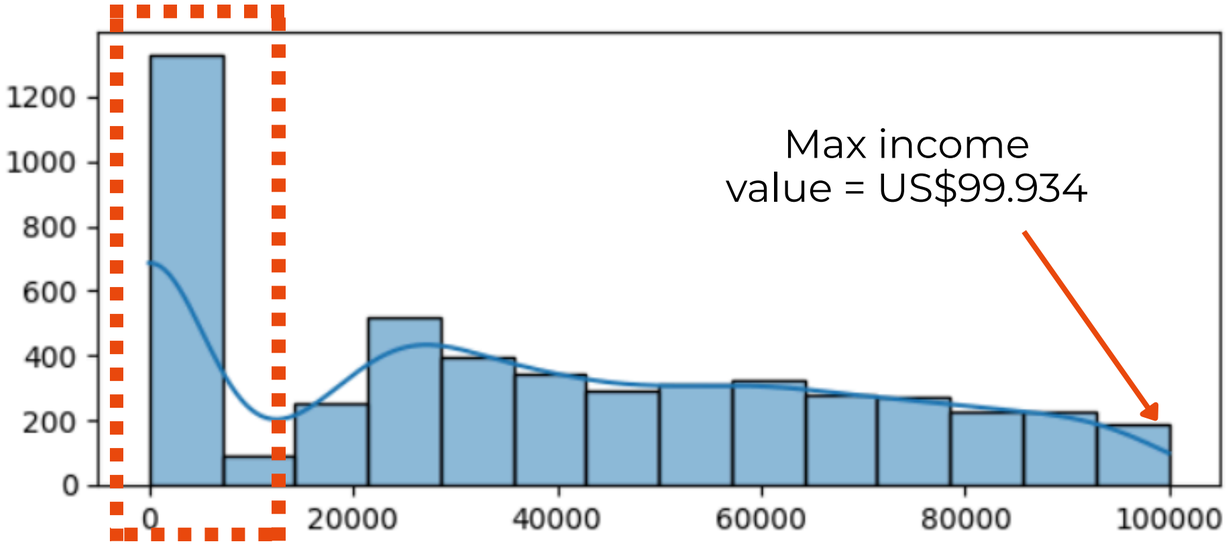
Categorical features

Numerical features

Target feature

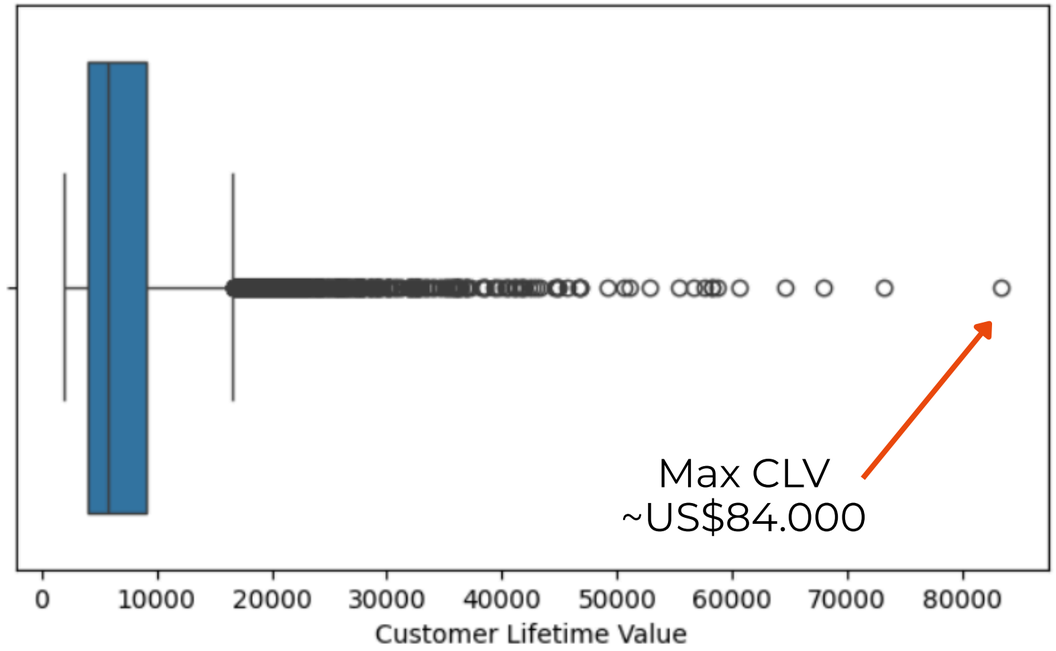
Exploratory Data Analysis

Income Distribution

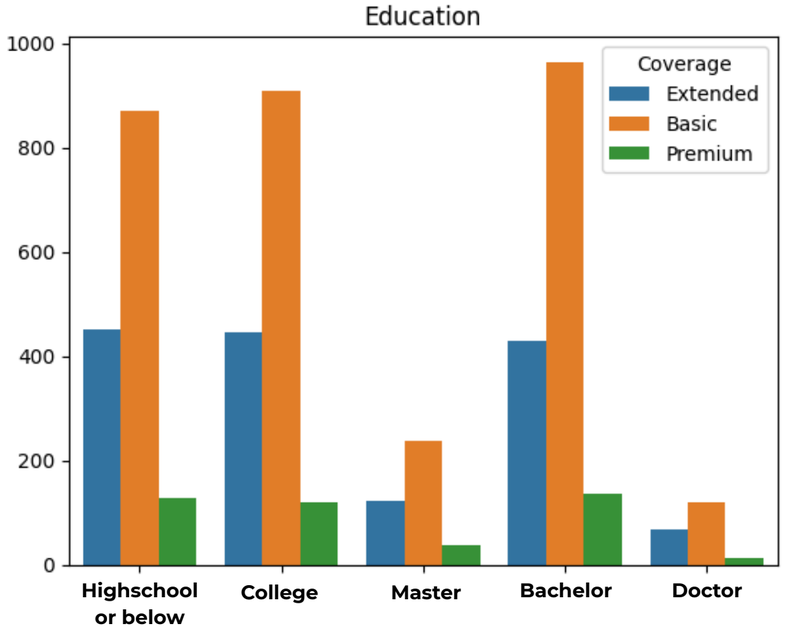


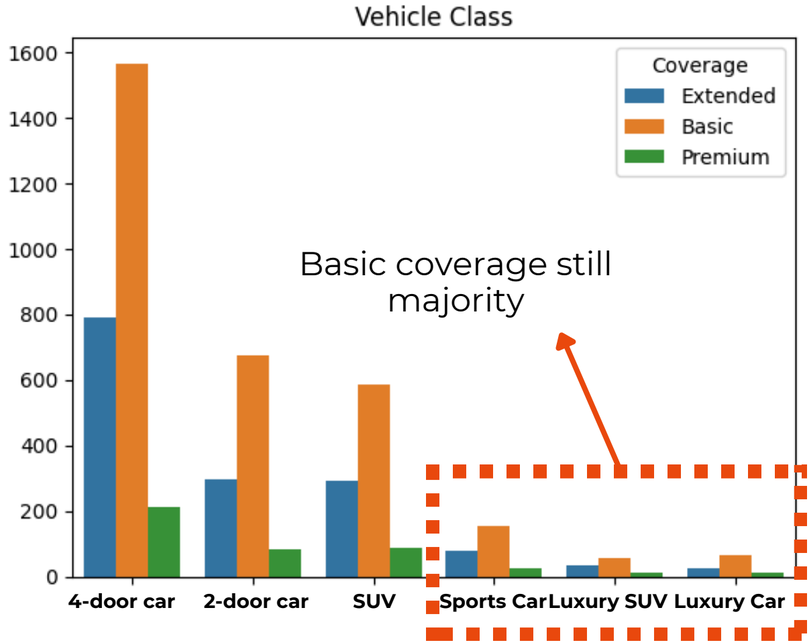
- Huge amount of ‘0’ income caused entirely by **“Unemployed” employment status.**
- Income value is **capped at US$100.000.**

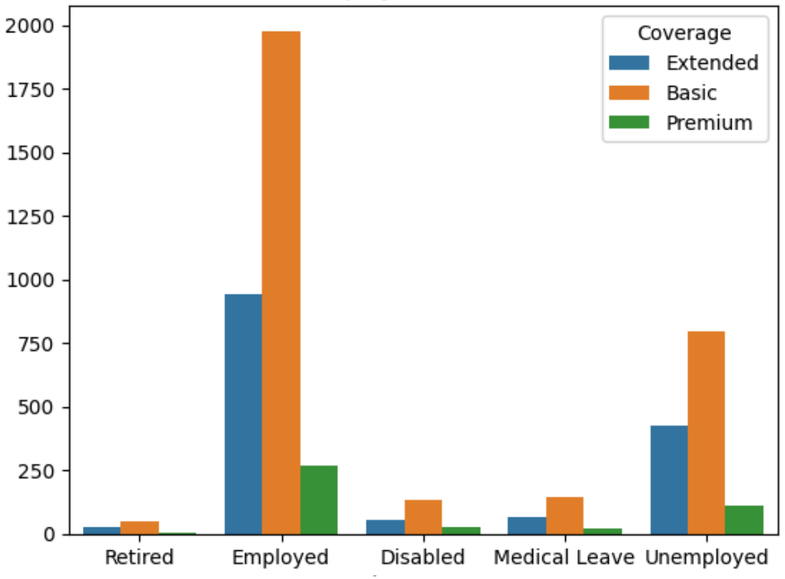
CLV Distribution

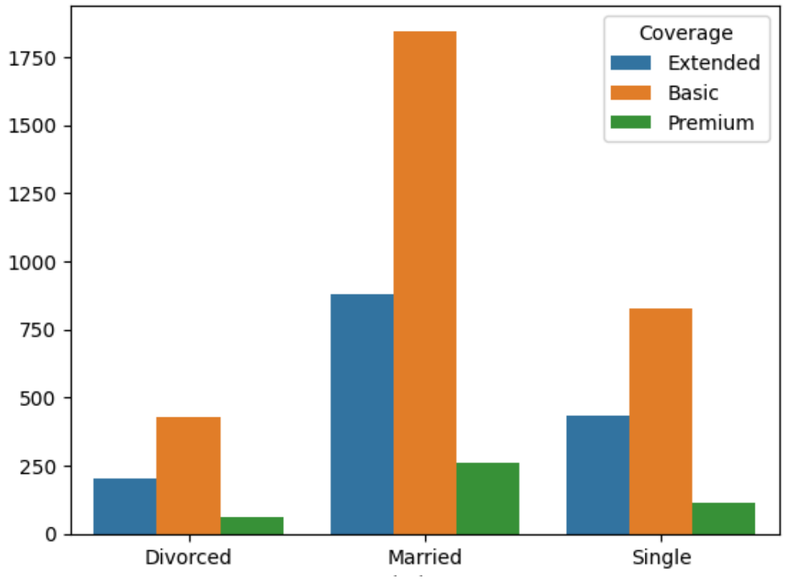


- Greatly **positive-skewed.**
- Could **ruin model performance** later.



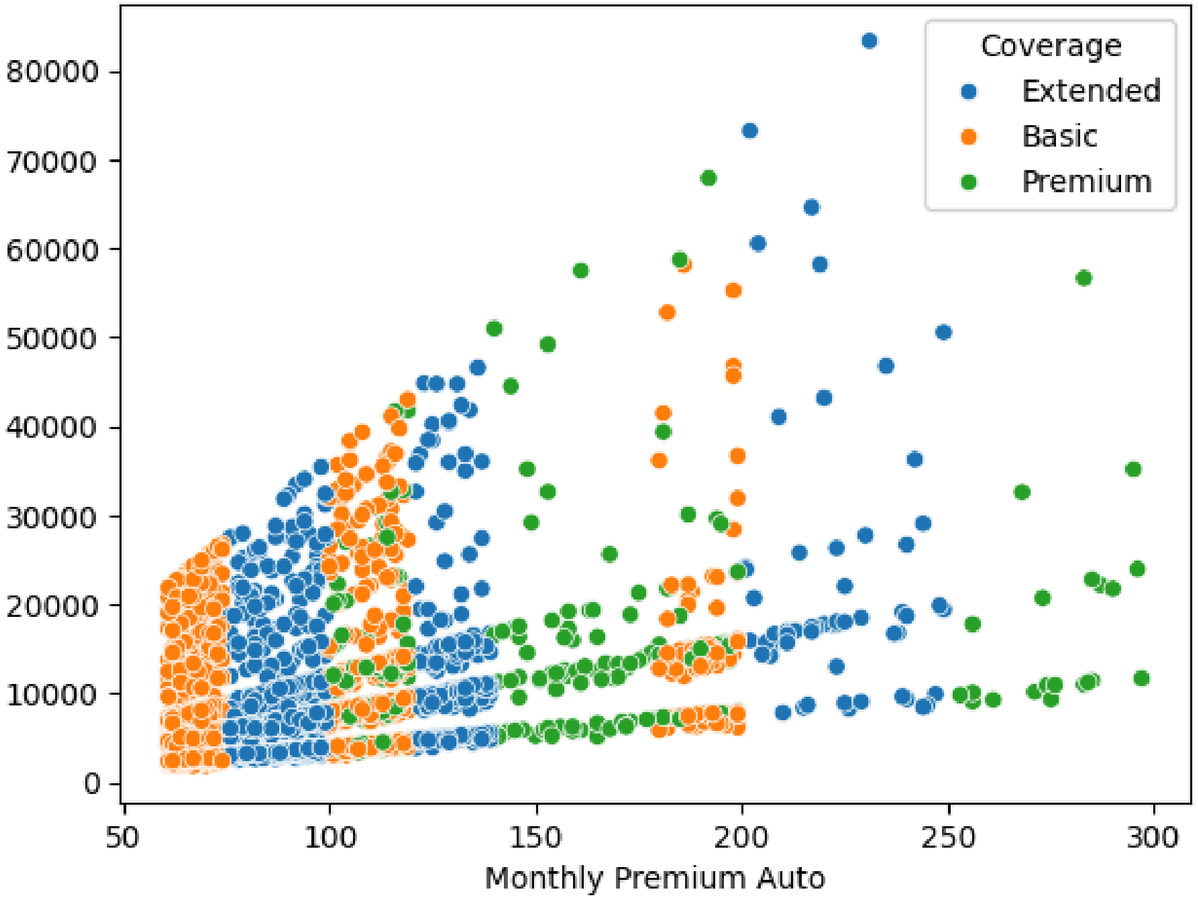


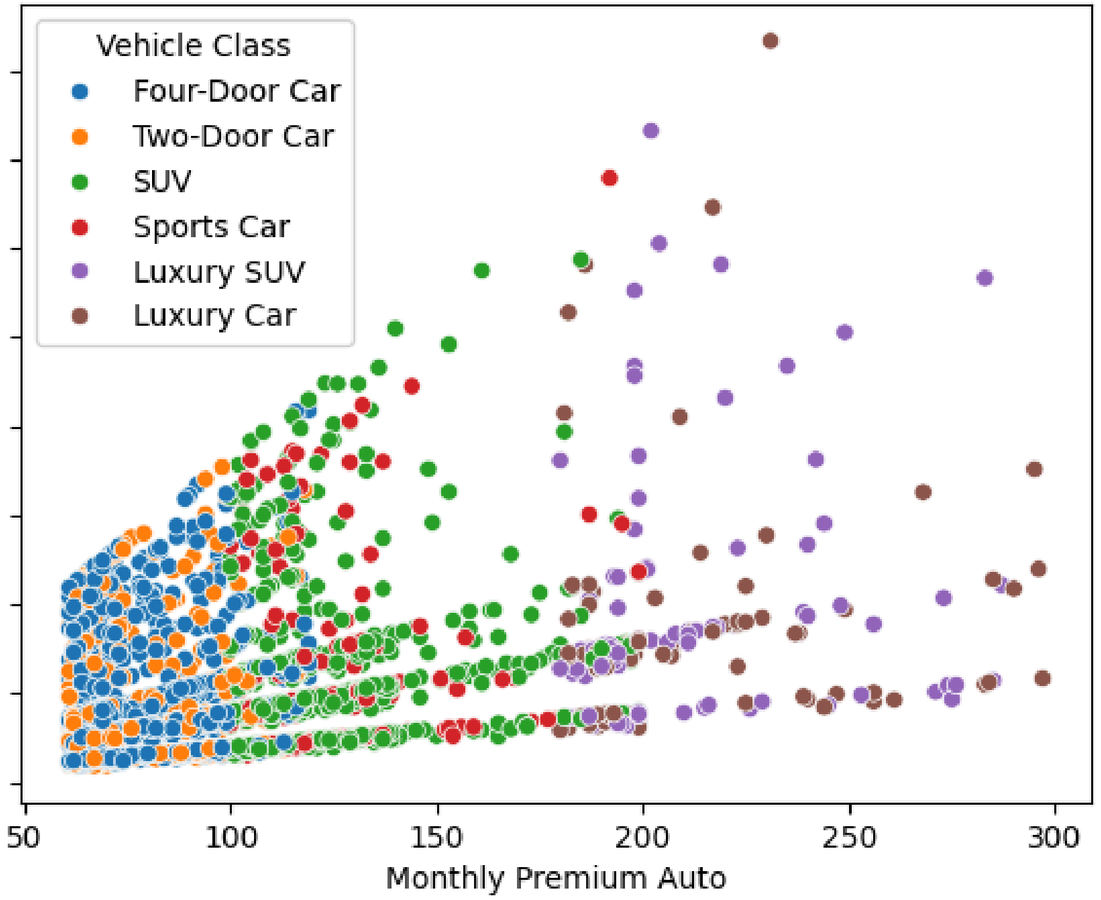




Across all categorical features,
“Basic” coverage dominates.

Premium vs Customer Lifetime Value



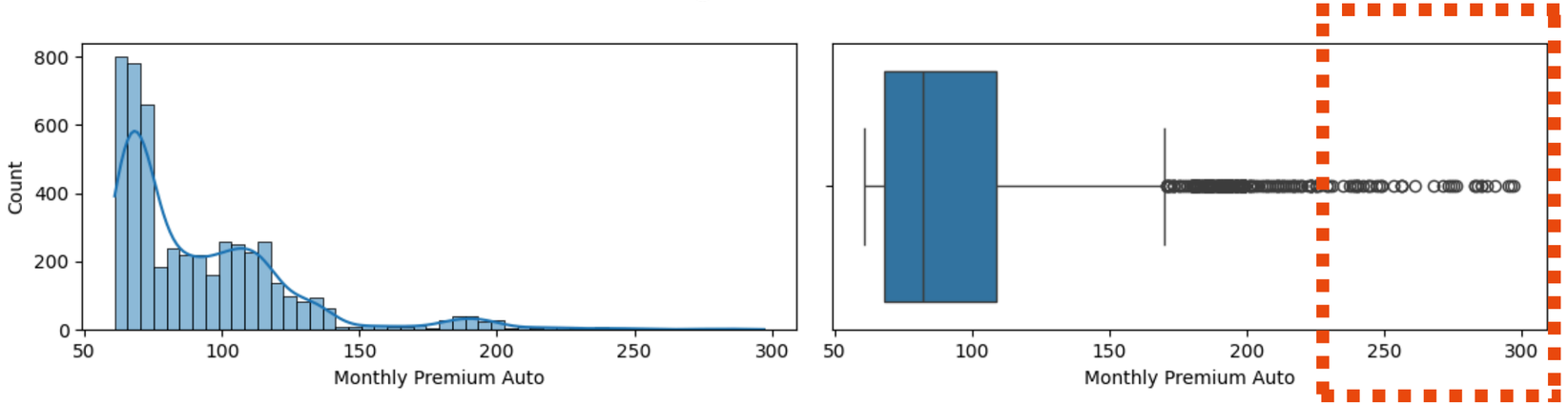


Clear **coverage pricing patterns visible** indicating each coverage plans have different set price rules.

- **4-door cars** Premium = US$60-100
- **2-door cars** Premium = US$60-110
- **SUV and Sports Car** Premium = US$100-180
- **Luxury SUV and Cars** Premium = US$180-270

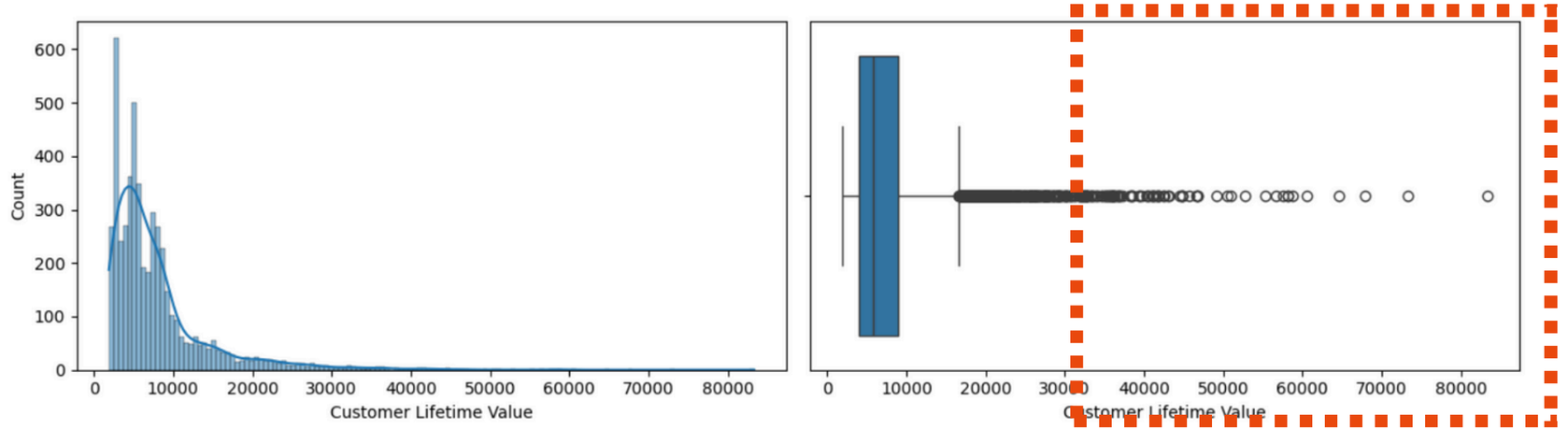
Outlier Treatment

Monthly Premium Auto



Highest monthly premium in 2019 is held by the state of Michigan at US$224.42. **Anything above will be dropped.**

Customer Lifetime Value



Removed CLV > US$30000 and any **unemployed customer with 0 income** because it is very unlikely and it ruins the model if used.

Feature Engineering Pipeline

Data
Outlier and duplicated treated outside pipeline.

Column Transformer
Encoder and scaler according to each feature

Log-transformer
Performs log transformation on target value for better fitting.

Model

Encoder

- **One-hot** — For < 5 unique values. Offer Type, Employment, and Marital Status
- **Binary** — For > 5 unique values. Only used on Vehicle Class.
- **Ordinal** — For ordinal data. Used on coverage and education

Scaler

Robust Scaling is used on **all numerical data** because most features have extreme outliers. Robust scaler is **resistant towards outliers.**

Model Selection

Model Pool

- Linear Regression
- KNearestNeighbor
- Decision Tree
- Bagging
- Random Forest
- ADABoost
- XGBoost
- Gradient Boost

Performed Cross-validation with **K-fold = 5, random_state = 42**

Best Model = ✨ Gradient Boost

RMSE

$2416

Good to evaluate models that wants to achieve high accuracy. Perfect for prediction models.

R^2

75.46%

Captures proportion of variance. Can be used as a reliability metric for the model.

MAPE

7.84%

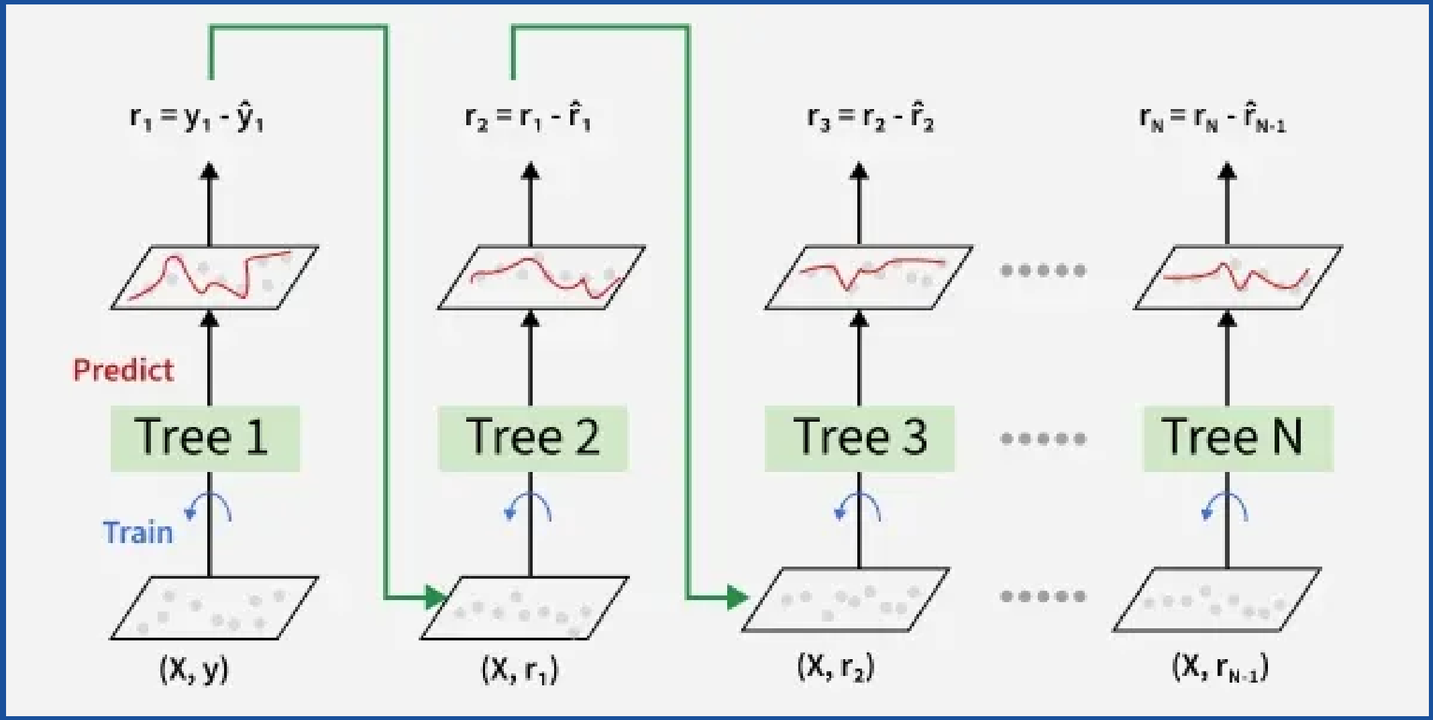
Easily interpretable absolute error percentage metric for stakeholders.

Time

0.02s

Measures time elapsed fitting and predicting. Essential metric for cost-oriented strategies.

How does Gradient Boost Work?



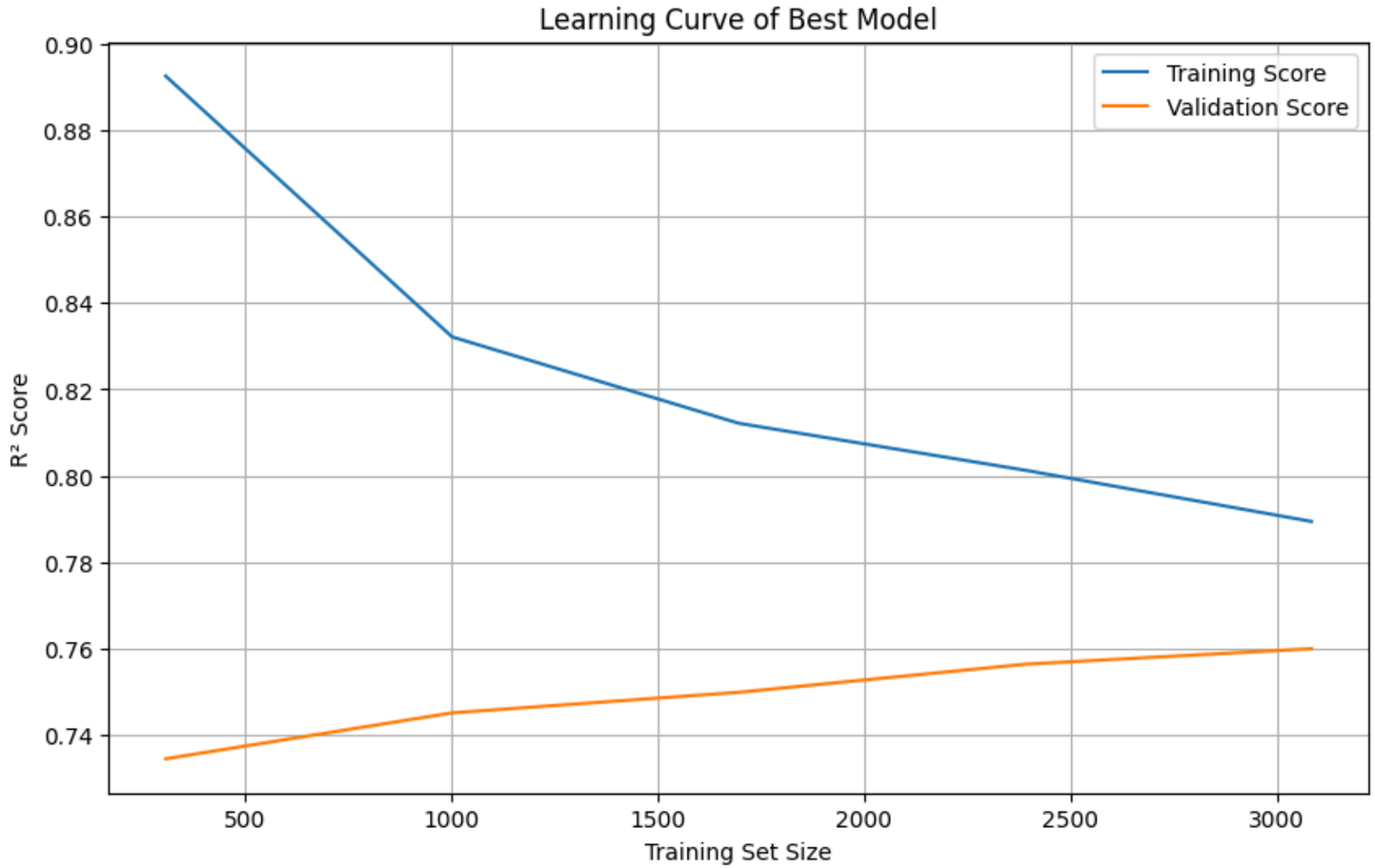
1. Model starts making **simple prediction** (i.e. average CLV)
2. Model checks the errors
3. Decision Tree trained only to fix that error
4. Model keeps adding tree to fix error.
5. Rinse and repeat!

Creating many weak models that **keeps learning from error.**

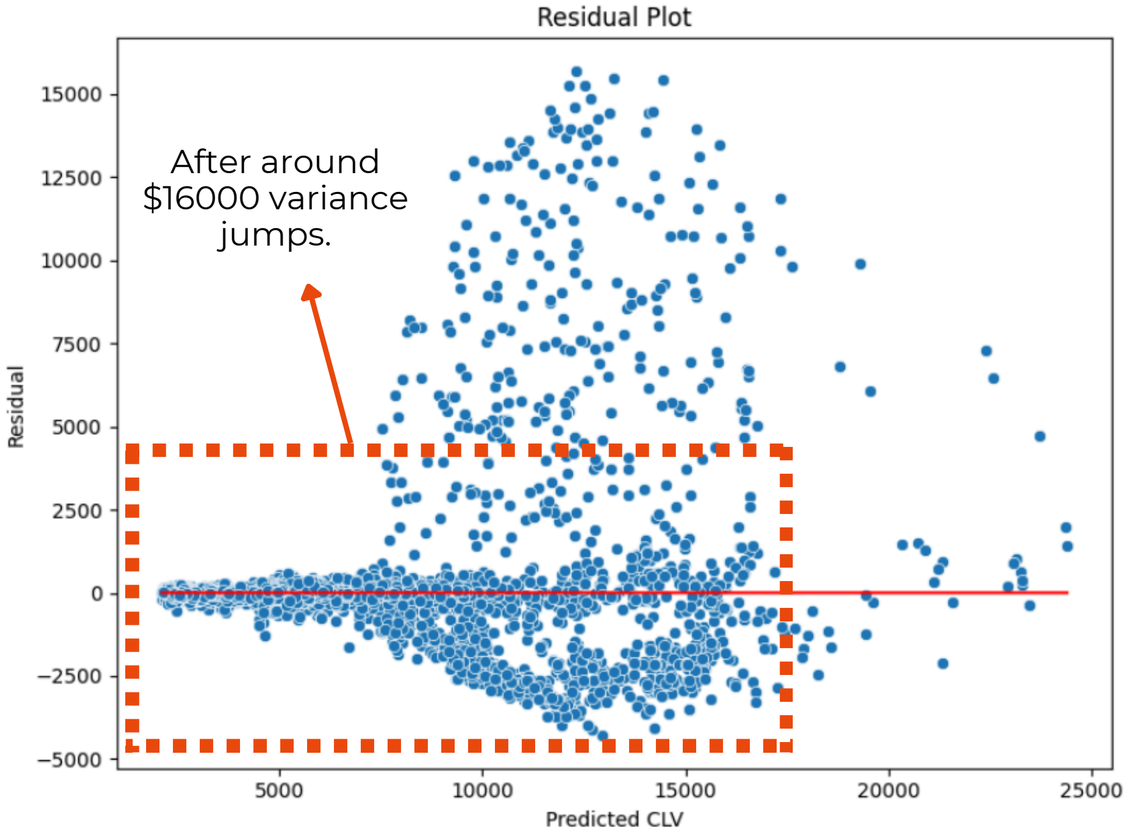
Model Best Parameter

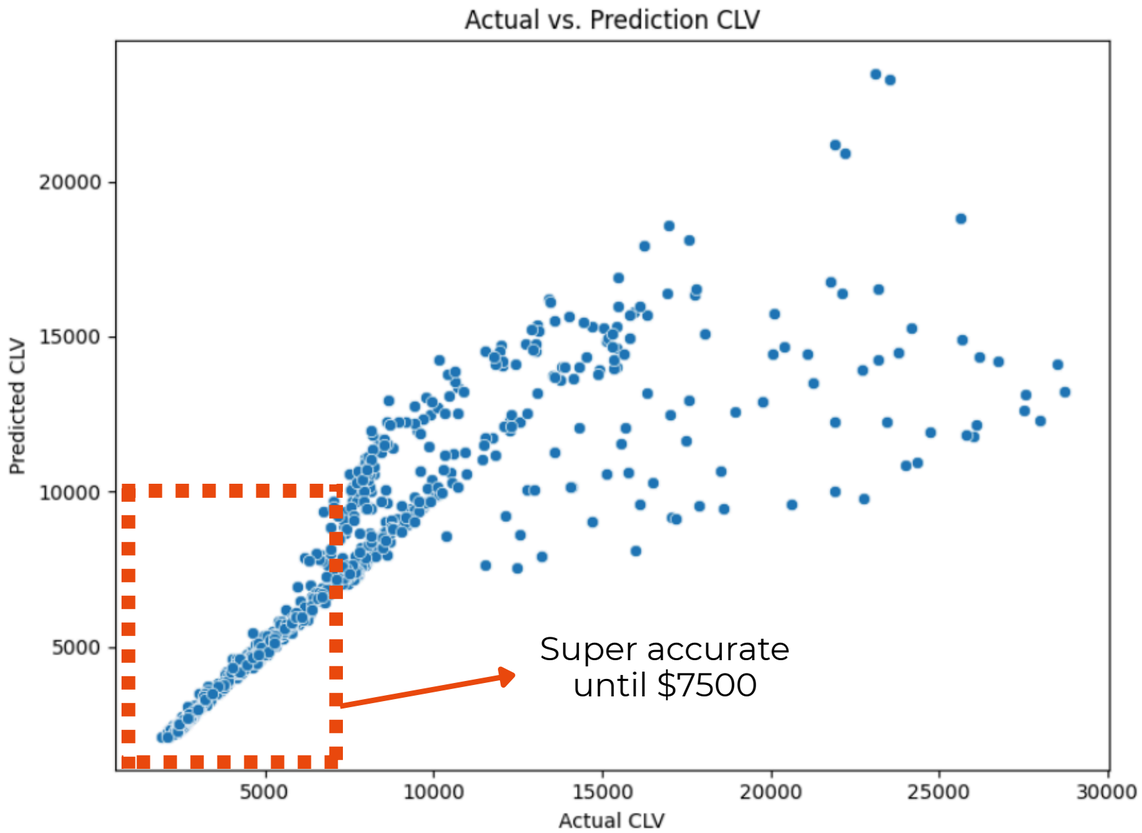
Hyperparameter	Best Value
N_estimators	600
Min_samples_split	15
Min_samples_leaf	1
Max_features	None
Max_depth	None
bootstrap	True

- Model is decently generalized.
- Model would benefit from **larger training data.**



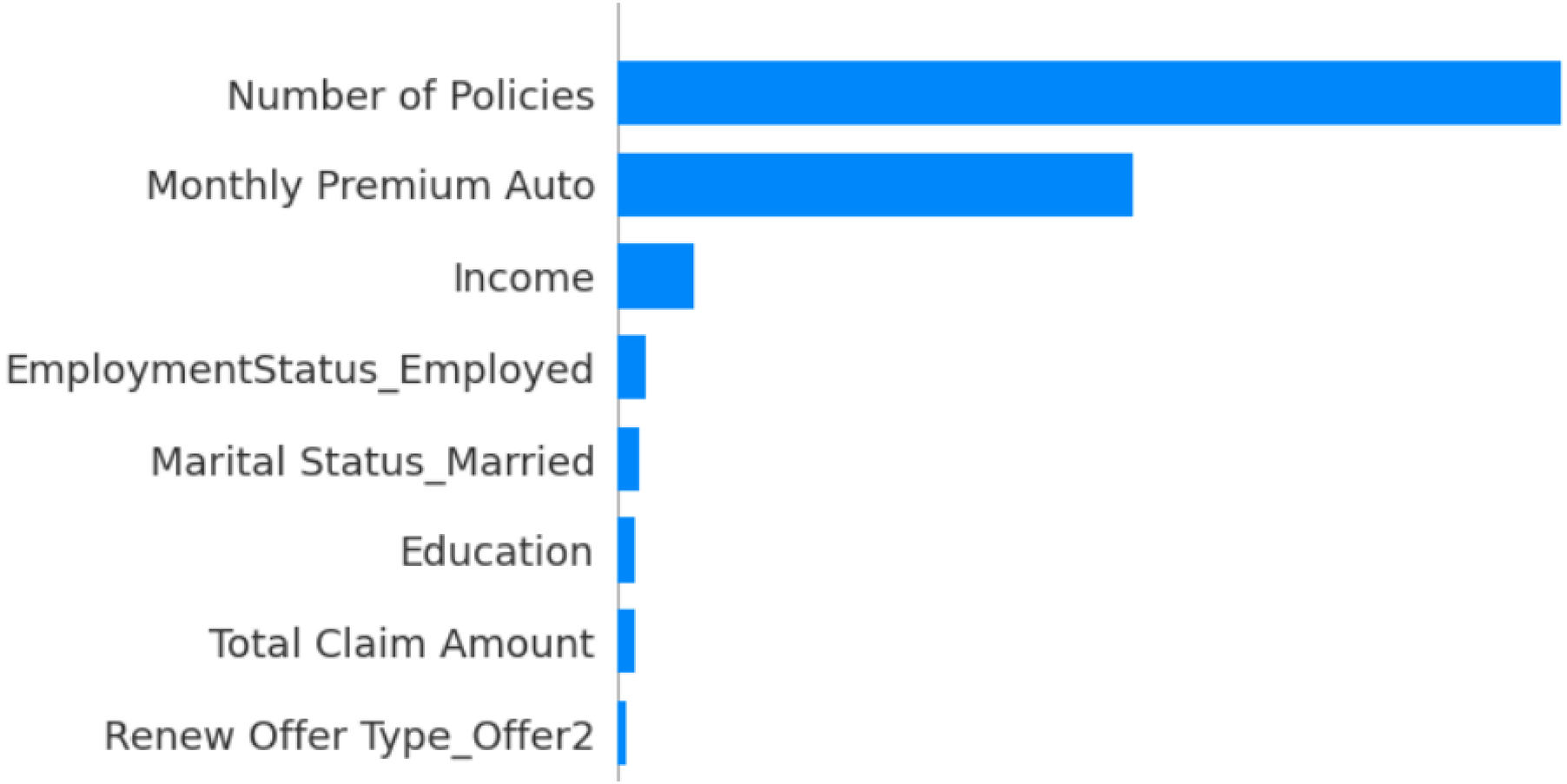
Residual Analysis





- Model is able to accurately predict CLV up until **~ US$10000** before the residual variance gets very unstable.
- After around the upper boundary of our CLV data (US$16624.75), the model **can no longer reliably perform.**

Feature Importance



- Number of policies result is suspicious and unreliable at 0.4 SHAP value.
- Premiums are indeed the main way auto insurance companies make their revenue.
- **Customer that are wealthier, with high-achieving careers and multiple responsibilities** generally have higher CLV.

Business Impact - ROI Improvement Case

Before Model Utilization

- Marketing spend **distributed evenly** across all customers.
- Broadly targeted retention campaign strategy = **Financially inefficient**
- 20% of customers contribute ~80% of total CLV, but they are **not prioritized.**

After Model Utilization

- Model **predicts CLV** and marketing team **flags high-CLV** or **high-churn-risk** customers
- Marketing spend reallocated, directing **more budget** toward these customers.
- Targeted approach improve retention ROI by **better customer prioritization.**

For now, model only predicts customers with **annual income of < US$100000,** and **Monthly Premium auto < US$224.42.**

Model only accurately predicts when **CLV < US$7500**, above US$16624.75 it is completely unreliable.

Model Limitations

Model does not capture and understand **customer tenure,** which is the most significant driver of overall CLV.

Conclusion

ML Perspective

- Gradient Boost model is produces prediction with least error and time.
- Performance are driven by features related to customer financial strength and product engagement

Business Perspective

- Utilizing this model will increase retention ROI through better customer prioritization.
- Model is able to predict most common CLV, but not super high-CLV customers

Recommendation

For Further CLV Modeling:

- Collect more data (**tenure, claim frequency/severity, and acquisition cost**) to improve model accuracy.
- **Add more high-value customer** samples to expand model effective range.
- Consider **binning CLV** (e.g., 0, normal, ultra-high) to enable more stable modeling and deeper segment-specific insights.
- **Regularly retrain** model with updated data

For Business/Marketing Team:

- Allocate more retention budget toward **proven high-value customers.**
- Identify low-value segments to **reduce marketing spend**.
- Encourage **policy bundling**
- Implement claim-reduction strategies to increase CLV.
- Develop customer segmentation models to support targeted and cost-efficient marketing actions.

GAICO

Thank you!