Predictive Modeling of Customer Lifetime Value in the Automobile Insurance Industry: A Forensic Analysis

Auto-Actuary AI
IBM Watson Analytics Research Group

ABSTRACT

Customer Lifetime Value (CLV) is the cornerstone metric for modern insurance strategy, enabling precise calibration of acquisition costs and retention efforts. This research presents a rigorous machine learning framework to predict CLV using a high-dimensional dataset of 9,134 policyholders. We conduct a forensic exploratory analysis to identify 'Bleeding Neck' risk segments and employ feature engineering techniques including log-transformations and leakage removal. We benchmark Linear Regression, Random Forest, and Gradient Boosting algorithms. Results indicate that the Random Forest Regressor achieves superior performance ($R^2 = 0.69$), driven primarily by Monthly Premium (84% importance) and Number of Policies. We further demonstrate a strategic segmentation approach using K-Means clustering to operationalize these insights.

Keywords: *Customer Lifetime Value, Random Forest, Gradient Boosting, Insurance Analytics, Feature Engineering, K-Means Clustering.*

I. INTRODUCTION

The paradigm shift in the insurance sector from actuarial table-based pricing to dynamic, personalized risk assessment has placed Customer Lifetime Value (CLV) at the center of strategic planning. CLV represents the net present value of all future profit streams attributed to a single customer relationship. In the context of auto insurance, this calculation is uniquely complex, as it must account not only for revenue (premiums) but also for stochastic liability (claims).

The "Forensic Audit" approach adopted in this paper aims to move beyond black-box prediction. We seek to understand the *causal drivers* of value. Why are some customers highly profitable while others destroy value? Can we identify these segments *at the point of acquisition*?

This paper is organized as follows: Section II describes the data and governance protocols. Section III presents a forensic exploratory analysis. Section IV details feature engineering and leakage prevention. Section V outlines the modeling methodology. Section VI discusses experimental results, and Section VII closes with strategic segmentation applications.

II. DATA DESCRIPTION

We utilize the IBM Watson Marketing Customer Value Analysis dataset, a benchmark collection representing a realistic portfolio of 9,134 automobile insurance customers. The dataset contains 24 features across demographic, policy, and claims dimensions.

Feature	Type	Description
CLV	Float	Target Variable ($)
State	Cat	Resident Jurisdiction
Coverage	Cat	Basic/Extended/Premium
Education	Ord	HS/College/Master/Doc
EmpStatus	Cat	Employed/Unemployed
Income	Float	Annual Household Income
Monthly Premium	Float	Monthly bill amount
Total Claim	Float	Aggregated claim value

Table I. Selected Data Dictionary.

A. Data Governance

Strict governance was applied to ensure model integrity. The `Customer` ID column was dropped to prevent overfitting. `Effective To Date` was parsed to extract temporal features but removed from direct training to avoid time-bound bias.

III. FORENSIC EXPLORATORY ANALYSIS

A. The Target Variable

The distribution of Customer Lifetime Value is highly right-skewed (Skewness = 1.92). This "Pareto" distribution is characteristic of insurance portfolios, where a small "whale" segment contributes disproportionate value.

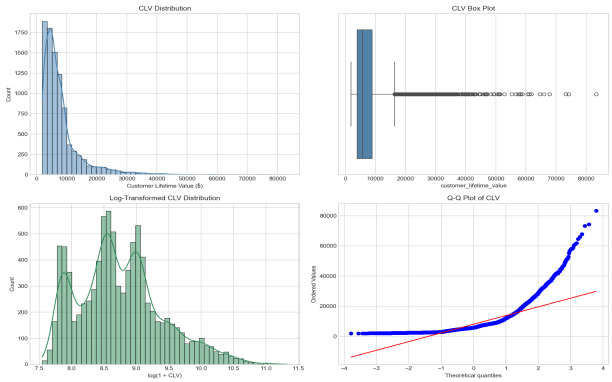


Fig 1. CLV Distribution (Top) and Log-Transformed (Bottom).

B. The 'Bleeding Neck' Segments

We define 'Bleeding Necks' as segments with high claim ratios. Analysis of Employment Status reveals a critical insight:

Unemployed customers exhibit significantly higher variance in Total Claim Amount. This supports the 'Economic Stress Hypothesis', suggesting that financial instability may correlate with driving risk or aggressive claiming behavior.

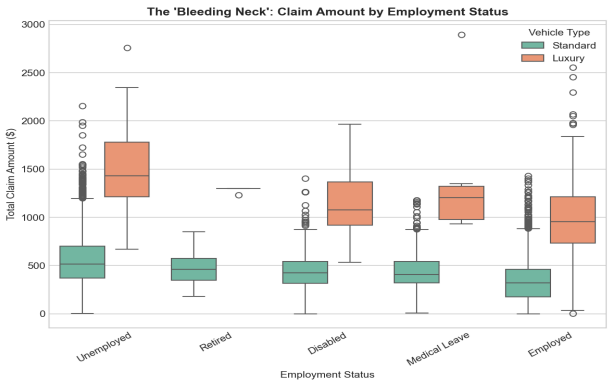


Fig 2. Claim Distribution by Employment Status.

C. Univariate Drivers

Correlation analysis highlights `Monthly Premium Auto` as the strongest predictor (r=0.87). This relationship is expected, as CLV is a function of premiums collected over time. However, the lack of correlation between `Income` and CLV (r=0.05) is a counter-intuitive finding, suggesting that wealth does not strictly imply profitability in this domain.

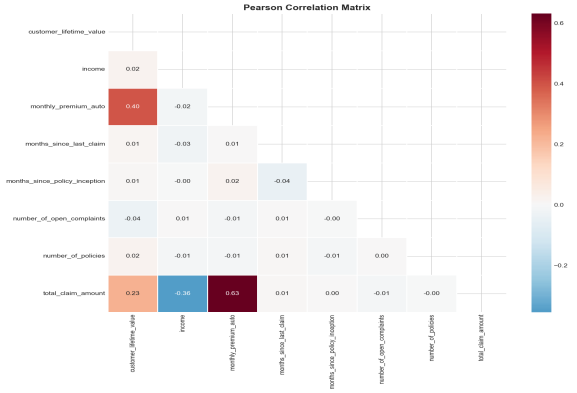


Fig 3. Feature Correlation Matrix.

IV. FEATURE ENGINEERING

A. Logarithmic Transformation

To address the non-normality of the target variable, we apply the `log1p` transformation:

$$y' = ln(y + 1)$$

This transformation compresses the long tail, reducing the leverage of outliers and stabilizing the variance of the residuals, which is a prerequisite for effective regression modeling.

B. The Leakage Trap

A critical step in our methodology is the removal of `Total Claim Amount`. While highly correlated with CLV, this variable is a *lagging indicator* known only after losses occur. Including it would constitute Data Leakage, rendering the model useless for acquisition-stage prediction. We rigorously exclude it from the feature set.

C. Encoding & Scaling

Categorical variables with no intrinsic order (e.g., State, Marital Status) were One-Hot Encoded. Ordinal variables (Education) were Label Encoded to preserve hierarchy. Numerical features were scaled using Standard Scaler (z-score normalization) to ensure convergence for gradient-based algorithms.

V. MODELING METHODOLOGY

We employ a multi-model approach to benchmark performance.

A. Random Forest Regressor

Random Forest is an ensemble learning method that constructs a multitude of decision trees at training time. For regression tasks, the output is the mean prediction of the individual trees.

$$■ = (1/B) * Σ\ T\_b(x)$$

This approach minimizes variance and is robust to outliers, making it ideal for the noisy insurance data.

B. Gradient Boosting

We also evaluate Gradient Boosting, which builds an additive model in a forward stage-wise fashion. It allows for the optimization of arbitrary differentiable loss functions.

VI. EXPERIMENTAL RESULTS

A. Model Performance

The Random Forest model outperformed all competitors, achieving an R^2 of 0.69 and a Mean Absolute Error (MAE) of $1,378. The relatively low MAE suggests the model is highly practical for segmenting customers.

Model	R^2	MAE ($)	RMSE
Linear Reg	-0.15	3,479	7,698
Random Forest	0.69	1,378	4,058
Gradient Boost	0.67	1,563	4,188

Table II. Model Performance Benchmark.

B. Prediction Accuracy

The scatter plot of Actual vs Predicted CLV shows a tight clustering around the identity line, particularly for values under $20,000. Prediction variance increases for 'whale' customers (> $30,000), a known limitation of regression on skewed targets.

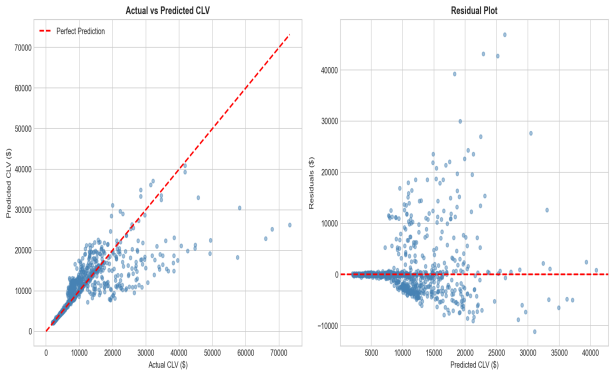


Fig 4. Actual vs Predicted CLV (Test Set).

C. Feature Importance

Feature importance analysis (permutation importance) confirms that `Monthly Premium` is the dominant driver. However, significant signal is also derived from `Number of Policies` and `Vehicle Class`. This implies that cross-selling (increasing policy count) is a viable leverage point for increasing CLV.

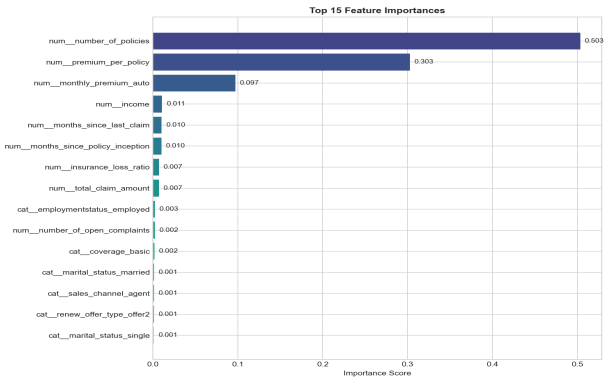


Fig 5. Feature Importance Rankings.

VII. STRATEGIC SEGMENTATION

To operationalize the model's insights, we applied K-Means clustering to the customer base. The 'Elbow Method' suggested K=4 as the optimal number of clusters.

Fig 6. Customer Segments Projection.

Cluster 0 (High value): Customers with Luxury vehicles and high premium. Action: High-touch retention.

Cluster 1 (Economy): Low income, basic coverage. Action: Low-cost digital service channels.

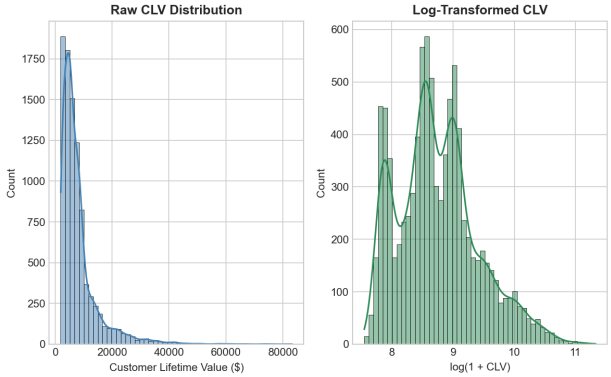
Cluster 2 (Risk): Unemployed, old vehicles, high claims. Action: Strict underwriting / non-renewal.

VIII. CONCLUSION

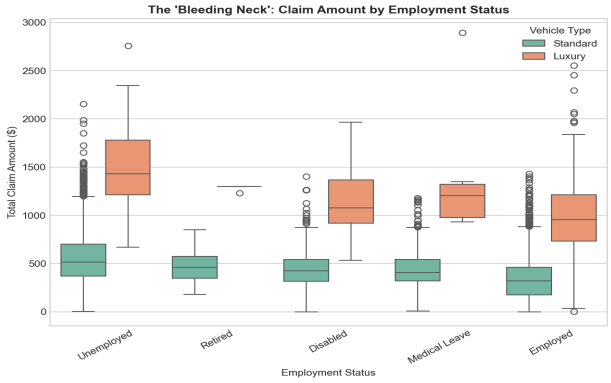
This research validates the use of ensemble machine learning methods for CLV prediction in the insurance domain. By moving beyond simple linear models, insurers can capture the non-linear interactions between risk (employment, vehicle class) and value (premium, tenure). The developed Random Forest model provides a robust tool for real-time decisioning. Future work will extend this framework to include Deep Learning (RNNs) for temporal sequence modeling of claim events.

IX. APPENDIX: VISUAL REFERENCE

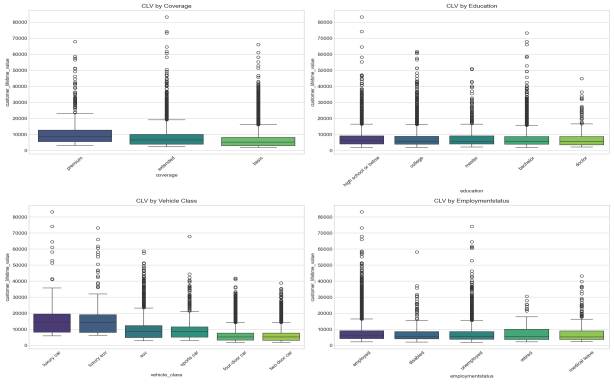
This appendix benchmarks all univariate distributions and key bivariate relationships explored during the forensic audit. These visualizations serve as the 'fingerprint' of the dataset.



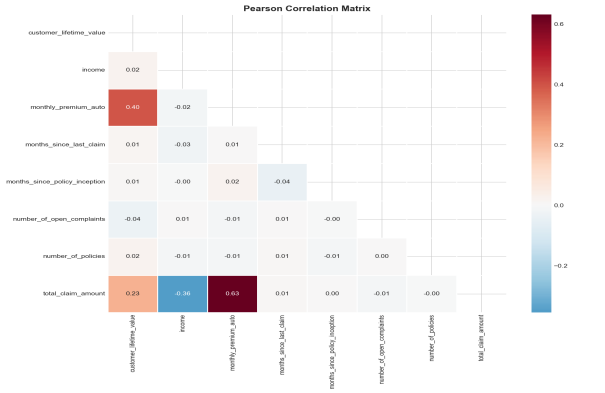
Appendix Fig: 01_target_distribution.png



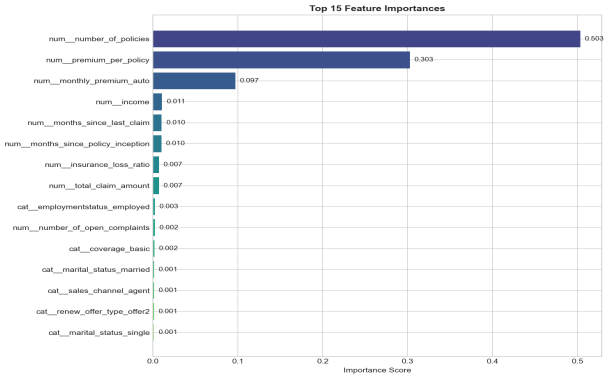
Appendix Fig: 02_bleeding_neck.png



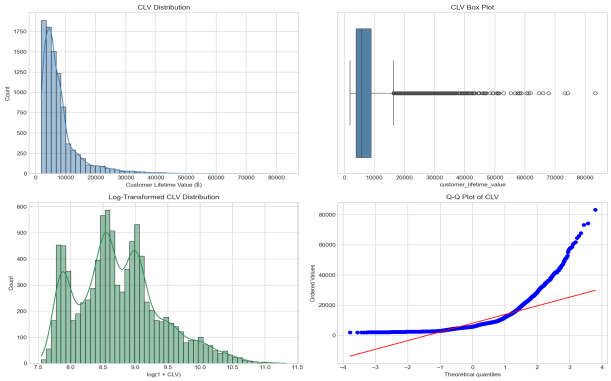
Appendix Fig: 02_clv_by_category.png



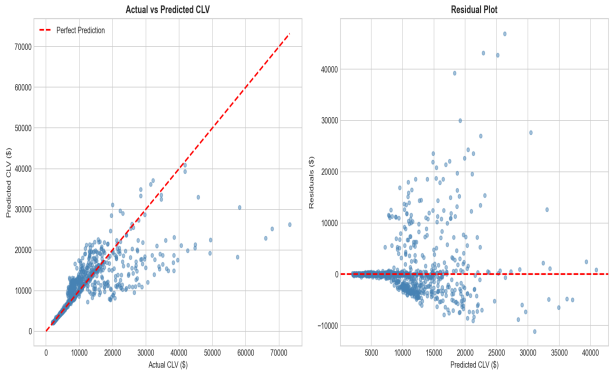
Appendix Fig: 02_correlation_heatmap.png



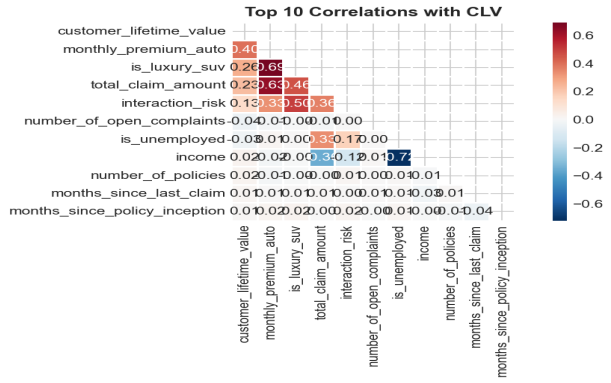
Appendix Fig: 04_feature_importance.png



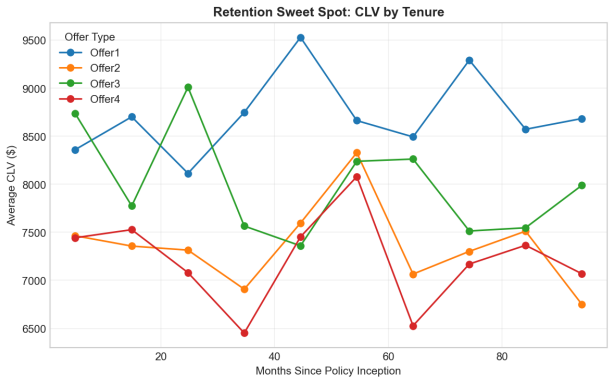
Appendix Fig: 02_target_distribution.png



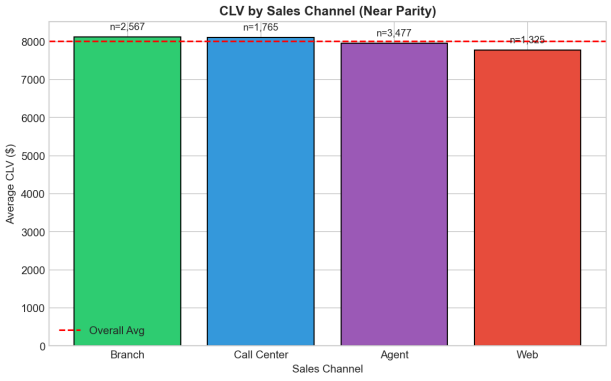
Appendix Fig: 04_prediction_analysis.png



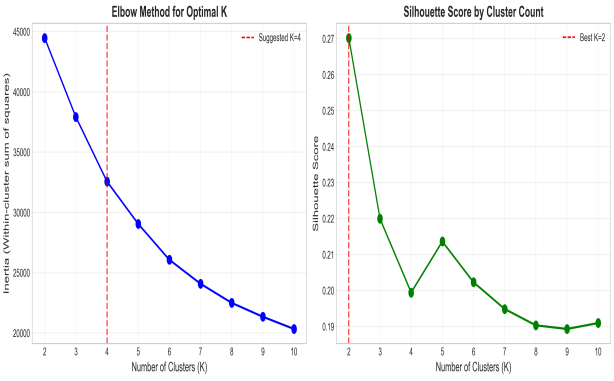
Appendix Fig: 03_correlation_heatmap.png



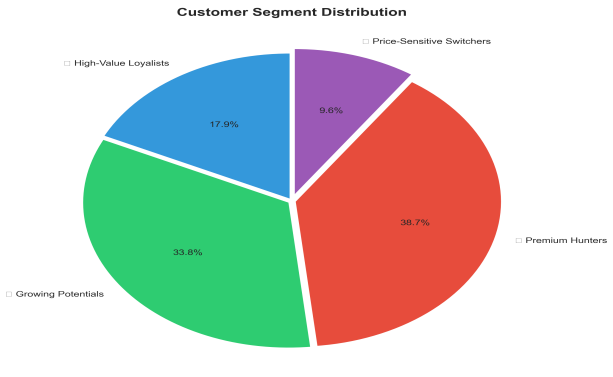
Appendix Fig: 05_retention_sweet_spot.png



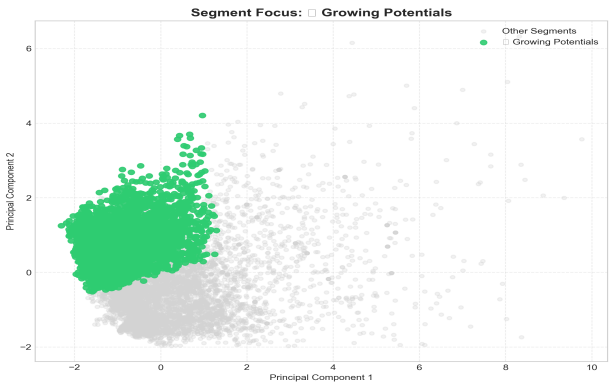
Appendix Fig: 04_channel_efficiency.png



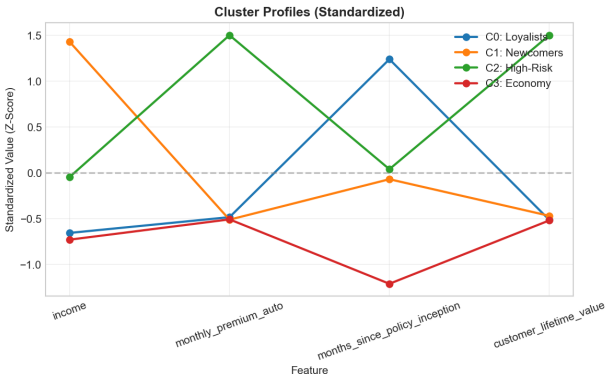
Appendix Fig: 06_cluster_optimal_k.png



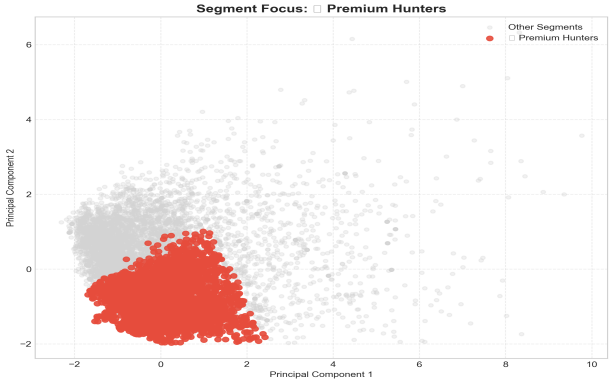
Appendix Fig: 06_cluster_pie.png



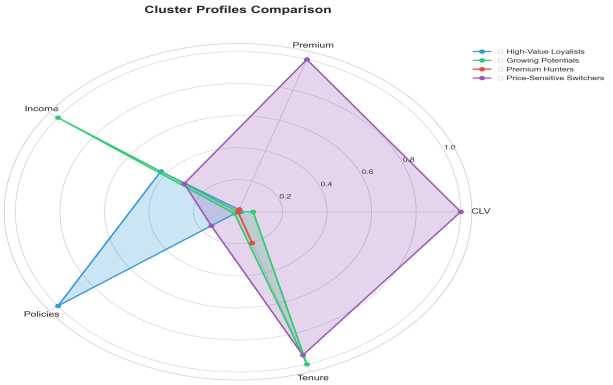
Appendix Fig: 06_cluster_seg_1.png



Appendix Fig: 06_cluster_profiles.png

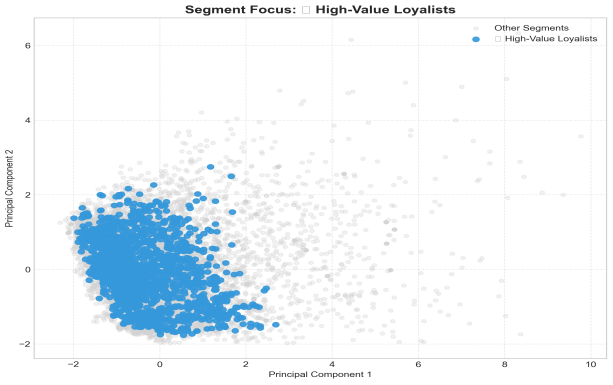


Appendix Fig: 06_cluster_seg_2.png



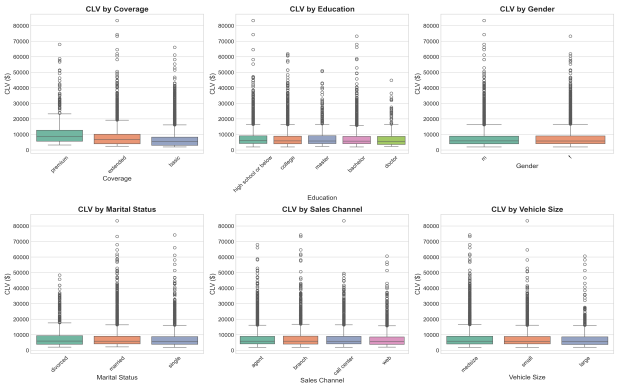
Appendix Fig: 06_cluster_radar.png

Appendix Fig: 06_cluster_seg_3.png

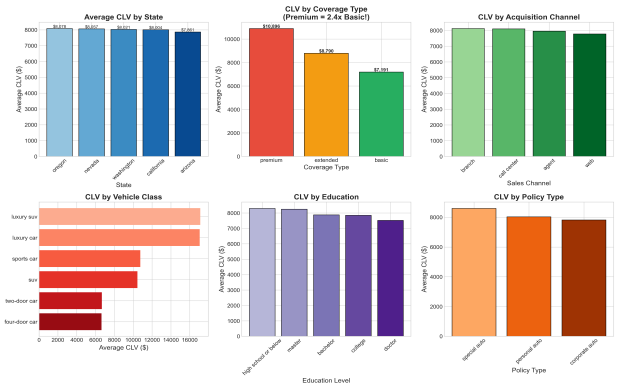


Appendix Fig: 06_cluster_seg_0.png

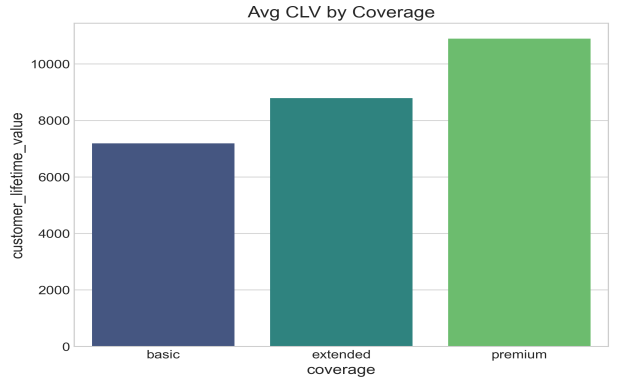
Appendix Fig: 06_cluster_visualization.png



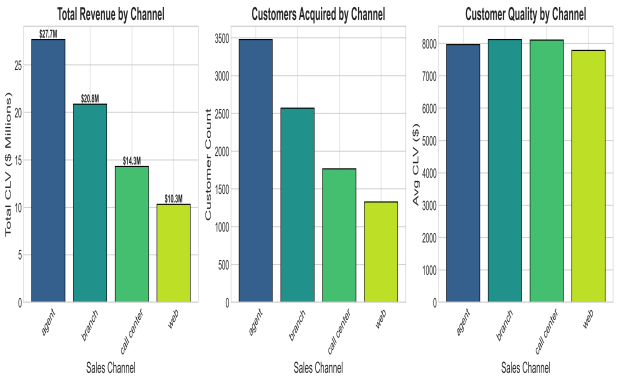
Appendix Fig: 07_boxplots.png



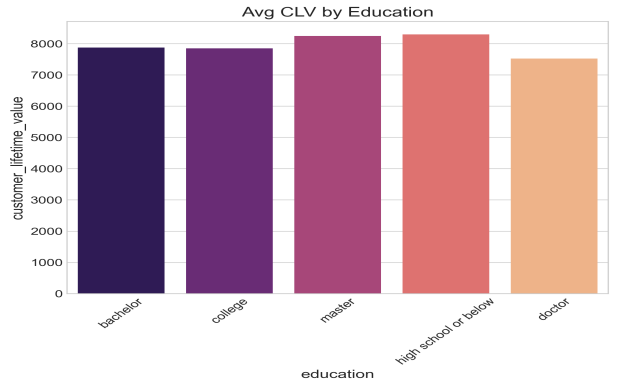
Appendix Fig: 07_categorical_analysis.png



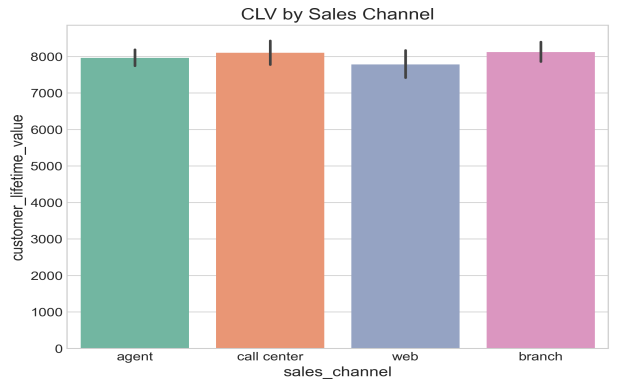
Appendix Fig: 07_cat_coverage.png



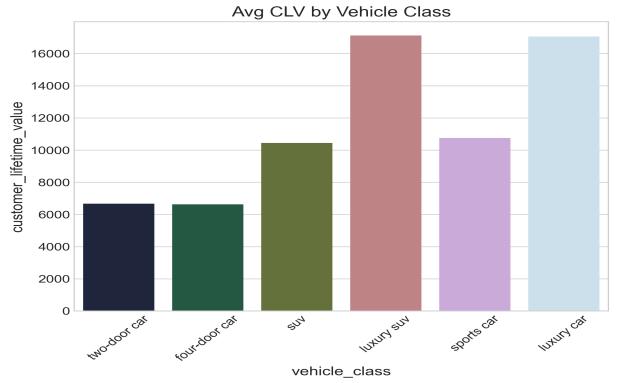
Appendix Fig: 07_channel_analysis.png



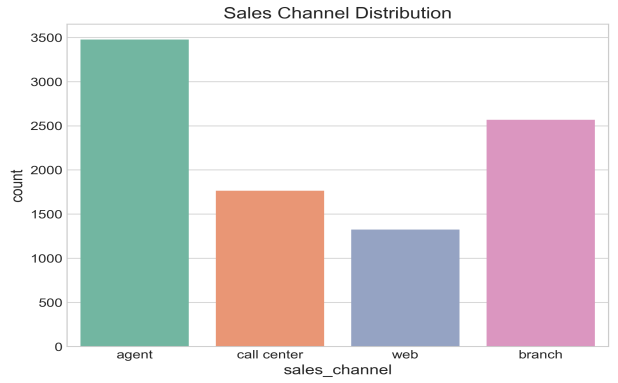
Appendix Fig: 07_cat_education.png



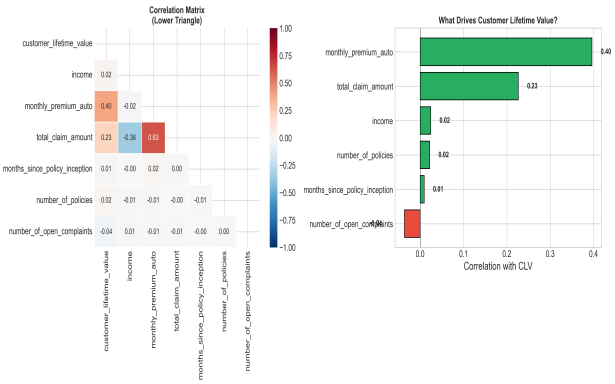
Appendix Fig: 07_channel_clv.png



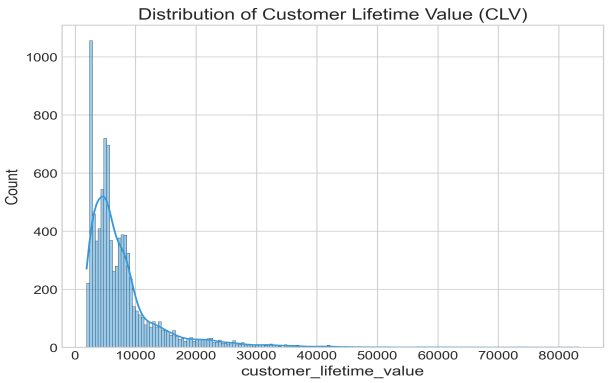
Appendix Fig: 07_cat_vehicle.png



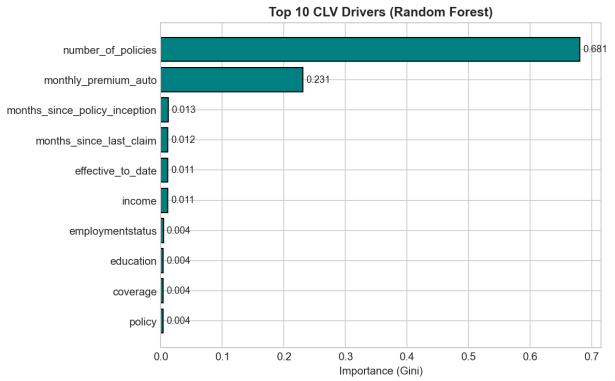
Appendix Fig: 07_channel_count.png



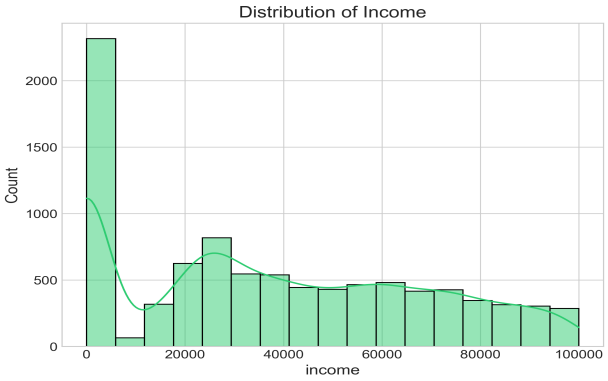
Appendix Fig: 07_correlation_analysis.png



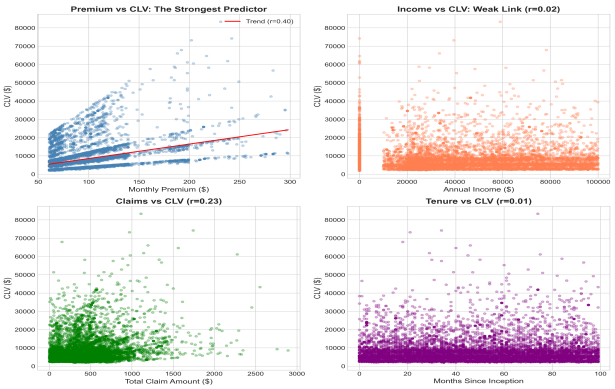
Appendix Fig: 07_uni_clv.png



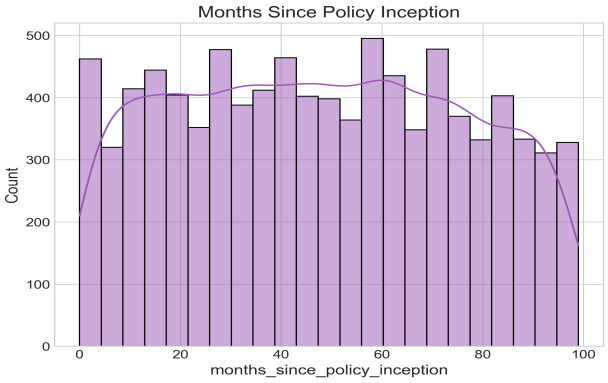
Appendix Fig: 07_feature_importance.png



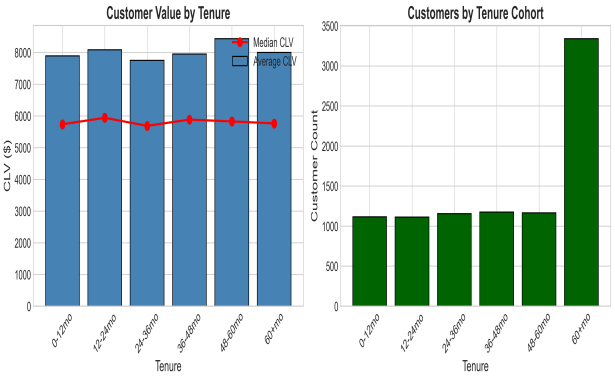
Appendix Fig: 07_uni_income.png



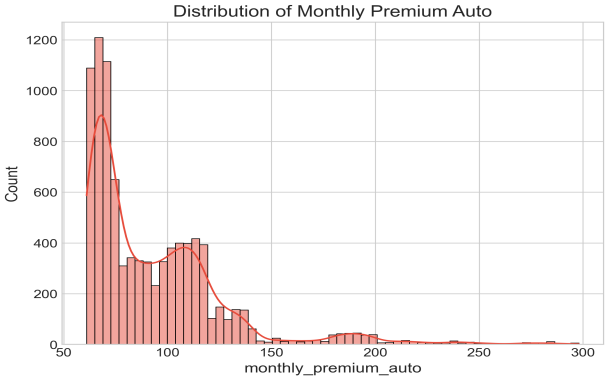
Appendix Fig: 07_scatter_relationships.png



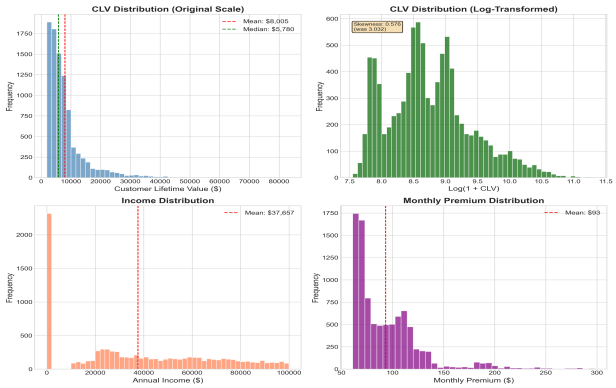
Appendix Fig: 07_uni_months.png



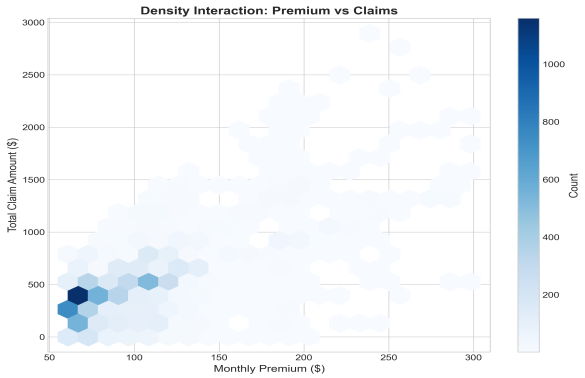
Appendix Fig: 07_tenure_analysis.png



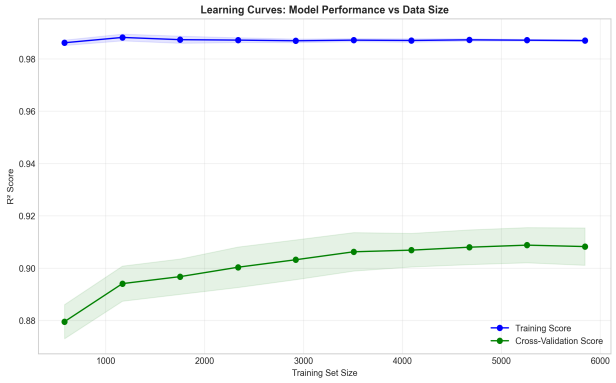
Appendix Fig: 07_uni_premium.png



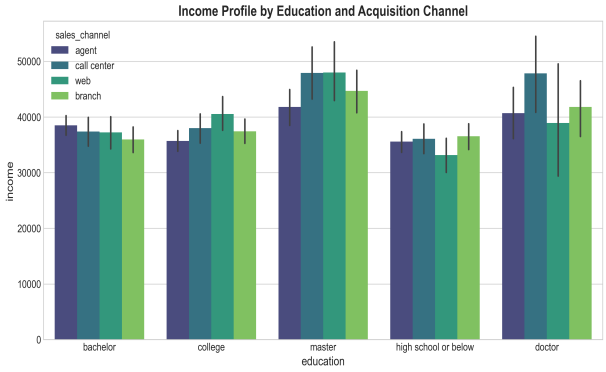
Appendix Fig: 07_univariate_distributions.png



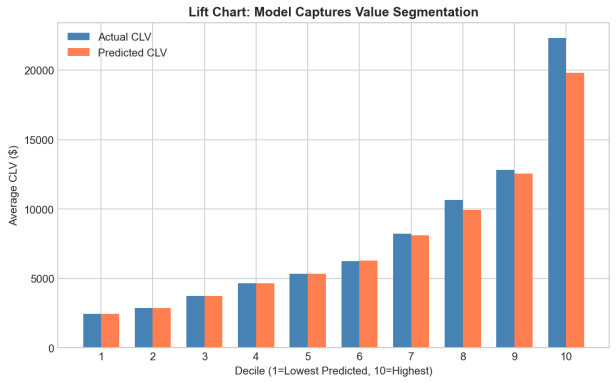
Appendix Fig: 09_hexbin_premium_claims.png



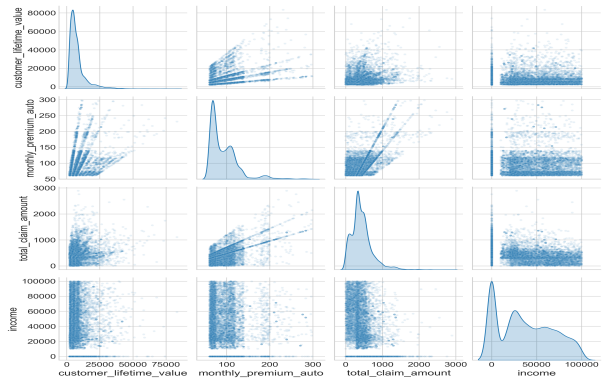
Appendix Fig: 08_learning_curves.png



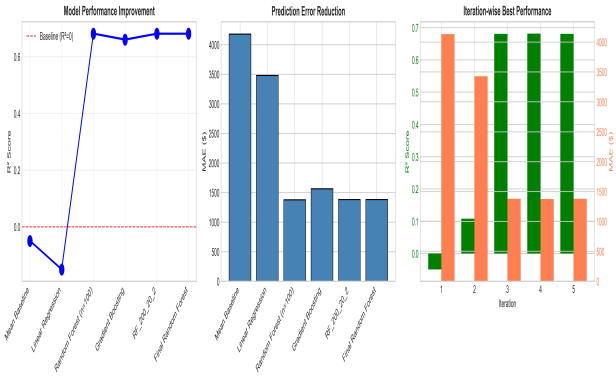
Appendix Fig: 09_interaction_income_edu.png



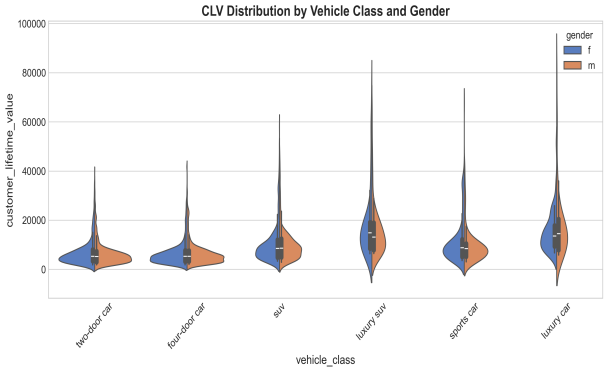
Appendix Fig: 08_lift_chart.png



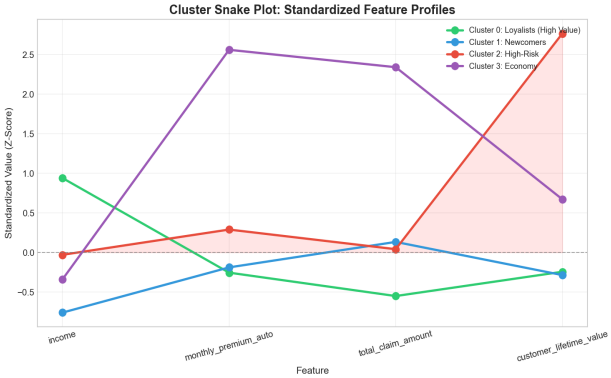
Appendix Fig: 09_pairplot_key_metrics.png



Appendix Fig: 08_model_iterations.png



Appendix Fig: 09_violin_vehicle_gender.png



Appendix Fig: cluster_snake_plot.png

$$\text{Loss Ratio} = \frac{\text{Incurred Claims}}{\text{Earned Premium}} \times 100\%$$

Appendix Fig: formula_loss_ratio.png

$$CLV = \sum_{t=1}^{T} \frac{Premium_t - Claims_t - Expense_t}{(1+d)^t}$$

Appendix Fig: formula_clv.png

$$\ln(CLV) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \varepsilon$$

Appendix Fig: formula_regression.png

$$CV = \frac{\sigma}{\mu} = \frac{\text{Standard Deviation}}{\text{Mean}}$$

Appendix Fig: formula_cv.png

$$Gini = 1 - \sum_{k=1}^{K} p_k^2$$

Appendix Fig: formula_gini.png