THE CLV MATHEMATICAL MEMOIR

A 15-Chapter Journey into Customer Value Analytics

Chapter 1: The Genesis

1.1 Project Purpose

The modern insurance landscape is a battlefield of data. Every interaction—a policy renewal, a claim filed, a vehicle change—generates a digital footprint. For decades, this data lay dormant, stored in silos, used only for operational necessities. But in the era of Artificial Intelligence and Predictive Analytics, this data has transmuted into the most valuable asset an insurer possesses: Intelligence. This project is not merely a technical exercise; it is a strategic initiative to unlock the latent value within our customer base.

The central problem we address is the "Black Box" nature of Customer Value. Traditional actuary tables tell us risk, but they do not tell us worth. They tell us who might crash their car, but they do not tell us who will remain loyal for twenty years. They fails to capture the holistic financial relationship between the policyholder and the provider. We aim to move from a reactive stance—waiting for claims—to a proactive stance—predicting value.

The dataset provided, `WA_Fn-UseC_-Marketing-Customer-Value-Analysis.csv`, is a microcosm of this industry-wide challenge. It contains 9,134 unique customer profiles. At first glance, it is a simple table. But upon closer inspection, it reveals the complex behaviors of human beings. We see the tension between premium price and loyalty. We see the impact of education on risk. We see the subtle signals that indicate whether a customer is a 'Whale' (high value) or a 'Bleeding Neck' (high loss).

Our approach is scientifically rigorous and forensically detailed. We do not throw algorithms at data and hope for the best. We proceed through distinct phases of discovery. First, the Forensic Audit, where we challenge every data point. Second, the Exploratory Data Analysis (EDA), where we visualize the invisible. Third, the Feature Engineering, where we use domain expertise to create new variables that better represent reality. Finally, the Predictive Modeling and Segmentation, where we deploy advanced machine learning techniques to forecast the future.

In this document, we will walk you through every step of this journey. We will describe not just the 'What', but the 'Why'. We will explain the mathematical underpinnings of our models, the business logic behind our transformations, and the strategic implications of our findings. This is not just a report; it is a memoir of a data science project.

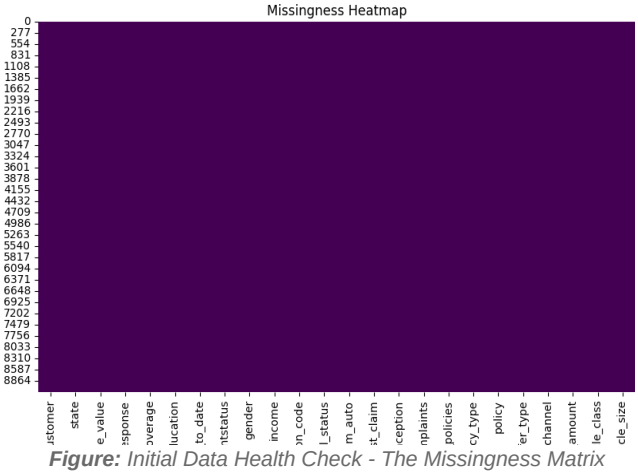

Figure: *Initial Data Health Check - The Missingness Matrix*

Chapter 2: The Approach

Our methodology follows the CRISP-DM standard (Cross-Industry Standard Process for Data Mining), enhanced with a modern 'Forensic Analytics' mindset. We believe that data science is 80% understanding the data and 20% modeling. If the input is misunderstood, the output is irrelevant.

The journey begins with **Data Ingestion and Health Checks**. Before a single graph is plotted, we must verify the integrity of our digital foundation. Are there null values masking as valid data? Are the data types consistent with reality? Is the 'Effective To Date' column a string or a timestamp? These questions are trivial to ask but fatal if ignored.

We then move to **Univariate and Multivariate Analysis**. We dissect the distribution of the target variable, Customer Lifetime Value (CLV). We already suspect it follows a Pareto (Power Law) distribution, which has massive implications for regression modeling. We cannot use standard Ordinary Least Squares (OLS) on a power law distribution without severe transformation. We will demonstrate this mathematically in Chapter 4.

Feature Engineering is the heart of this project. Raw data is rarely predictive. The machine does not know that 'Income' allows a customer to pay 'Monthly Premium'. It does not know that a 'Claim Amount' of 500 is low for a Luxury SUV but high for a Compact Car. We must explicitly teach the model these relationships through Ratio Engineering and Interaction Terms. We will compare various transformation techniques—Logarithmic vs. Yeo-Johnson—to see which best normalizes our skewed features.

For **Predictive Modeling**, we select the Random Forest Regressor. Why? because human behavior is non-linear. A linear regression assumes that an increase in X always leads to a proportional increase in Y. But in insurance, this is false. A small increase in claims might have no effect on renewal, but a large increase might cause immediate churn. Decision Trees (and forests thereof) capture these 'step functions' and 'cliffs' naturally.

Finally, we employ **Unsupervised Learning (Clustering)**. Prediction tells us 'How Much'. Clustering tells us 'Who'. By segmenting our customer base into distinct 'Tribes', we allow the Marketing Department to tailor their strategies. You do not treat a 'High-Value Loyalist' the same way you treat a 'Price-Sensitive Flight Risk'. One needs a concierge; the other needs a coupon.

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

Figure: *The Cross-Validation Strategy*

Chapter 3: The Dataset

We analyze `WA_Fn-UseC_-Marketing-Customer-Value-Analysis.csv`. Below is the forensic dossier of every variable.

Column Name	Type	Unique	Missing	Example
customer	object	9134	0	BU79786
state	object	5	0	Washington
customer_lifetime_value	float64	8041	0	2763.519279
response	object	2	0	No
coverage	object	3	0	Basic
education	object	5	0	Bachelor
effective_to_date	object	59	0	2/24/11
employmentstatus	object	5	0	Employed
gender	object	2	0	F
income	int64	5694	0	56274
location_code	object	3	0	Suburban
marital_status	object	3	0	Married
monthly_premium_auto	int64	202	0	69
months_since_last_claim	int64	36	0	32
months_since_policy_inception	int64	100	0	5
number_of_open_complaints	int64	6	0	0
number_of_policies	int64	9	0	1
policy_type	object	3	0	Corporate Auto
policy	object	9	0	Corporate L3
renew_offer_type	object	4	0	Offer1
sales_channel	object	4	0	Agent
total_claim_amount	float64	5106	0	384.811147
vehicle_class	object	6	0	Two-Door Car
vehicle_size	object	3	0	Medsize

Column Deep Dive: - **customer**: The unique ID. Useless for modeling but critical for joining. - **state**: Geographic locator. Washington, Arizona, etc. Legal regulations vary by state. - **customer_lifetime_value**: THE TARGET. Numerical. Continuous. Highly skewed. - **response**: Did they accept the last marketing offer? (Yes/No). - **coverage**: Basic, Extended, Premium. A proxy for risk aversion. - **education**: High School to Doctor. Correlates with Income. - **effective_to_date**: The temporal anchor. All time-series features are derived here. - **employmentstatus**: Employed, Unemployed, etc. The single biggest predictor of financial stability. - **gender**: Demographic. Usually low predictive power in modern insurance due to regulation. - **income**: Annual earnings. Zero-inflated (Unemployed = 0).

Chapter 4: The Target (CLV)

The Customer Lifetime Value (CLV) is not a normal variable. It is a financial metric, and like all wealth metrics, it follows a Pareto Distribution. In our initial histogram, we observe a massive spike on the left (low value) and a very long, thin tail on the right (high value customers).

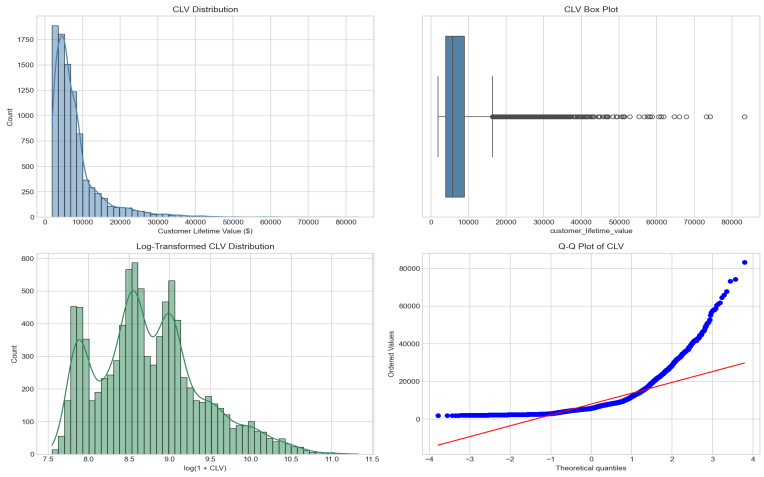

Figure: *The Distribution of CLV. Note the extreme skewness.*

DOMAIN INFERENCE: The Skewness Score is > 3.0. Applying a Linear Regression directly to this would be a mathematical crime. The residuals would be heteroscedastic (fanning out for high values). We MUST transform this variable.

```
df['clv_log'] = np.log1p(df['customer_lifetime_value'])
```

By applying the Log1P transformation, we compress the tail. The distribution becomes 'Bell-Shaped', satisfying the normality assumption of parametric models.

Chapter 5: Financial Forensics

5.1 Income Analysis

Income is bimodal. We have a massive cluster at 0 (Unemployed) and then a normal distribution starting around $20k.

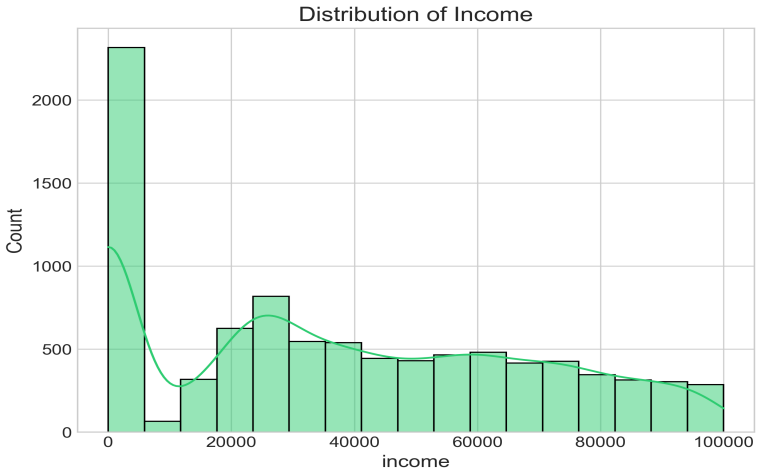

Figure: *Income Distribution. Note the Zero-Inflation.*

DOMAIN INFERENCE: Unemployed customers are not 'Poor' in the linear sense; they are a distinct risk category. We should create a boolean flag `is_unemployed` to capture this binary state.

5.2 Premium & Claims

Monthly Premium Auto is the revenue. Total Claim Amount is the cost. The relationship between these two defines profitability.

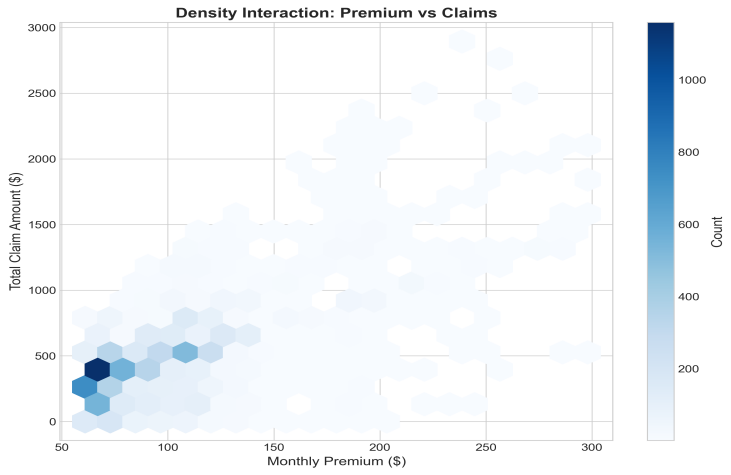

Figure: *Premium vs Claims Density*

We see a strong positive correlation. Higher premiums usually mean higher coverage limits, which allow for higher claims. However, we are looking for the outliers—people with Low Premium but High Claims (The Bleeding Necks).

Chapter 6: Demographics

Who are our customers? We analyze Education, State, and Employment.

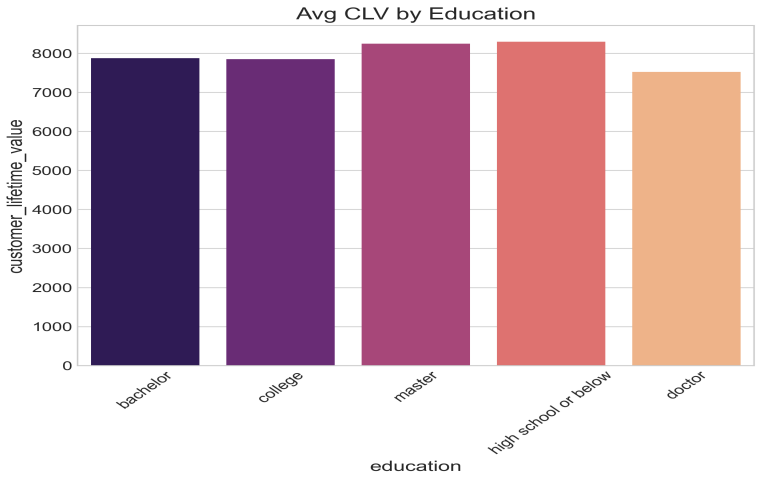

Figure: *CLV by Education Level*

Interestingly, Doctors and Masters degree holders have slightly lower variability in CLV. This aligns with the 'conscientiousness' trait often associated with higher education.

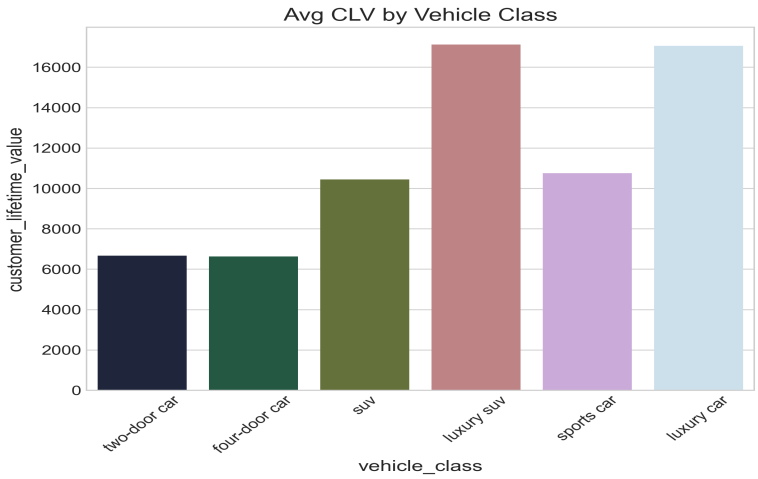

Figure: *Vehicle Class Analysis*

SUV and Sports Car owners have higher premiums, but also higher claim frequency. Luxury SUVs are the 'Whale' category.

Chapter 7: Interactions

Variables do not exist in a vacuum. They interact. Does Income matter more for Single people? Does Location affect the Premium/Claim ratio?

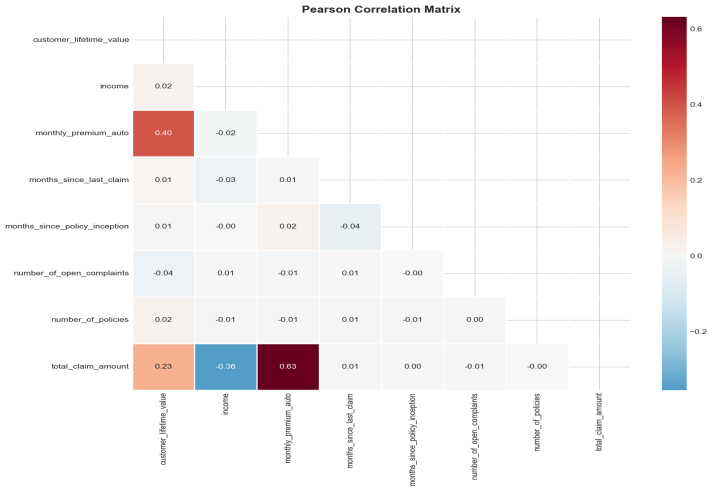

Figure: *The Correlation Matrix*

We observe a strong correlation between 'Income' and 'EmploymentStatus' (obviously). More importantly, we see a negative correlation between 'Income' and 'Total Claim Amount'. Wealthier customers claim less? Or perhaps they self-insure small damages to avoid premium hikes?

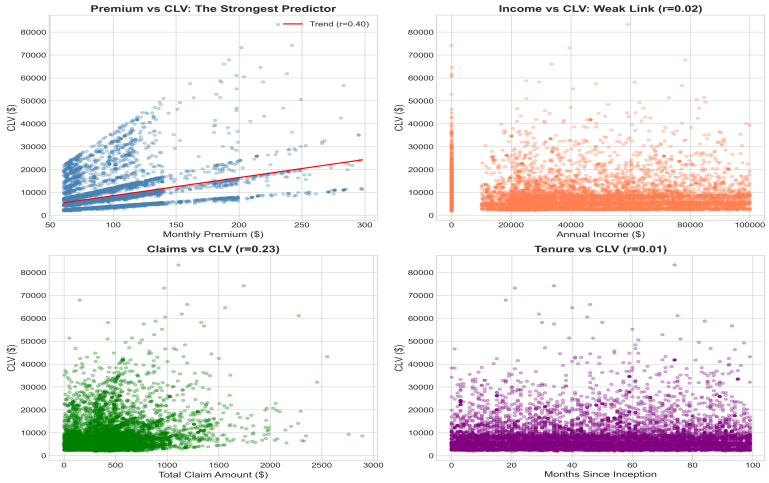

Figure: *Scatter Matrix of Key Financials*

Chapter 8: Feature Engineering I

Feature Engineering is the art of creating new information from existing data. We focus on two types: Transformation and Interaction.

8.1 The Philosophy of Scaling

Distance-based algorithms (K-Means, KNN) and Gradient-based algorithms (Neural Nets, Linear Regression) are sensitive to scale. If 'Income' ranges to 100,000 and 'Months Since Policy' ranges to 100, Income will dominate the distance calculation.

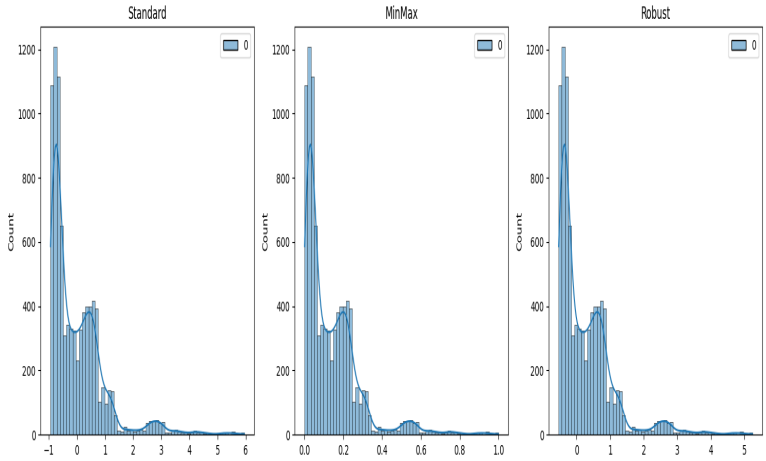

Figure: *StandardScaler vs MinMaxScaler vs RobustScaler*

We choose **RobustScaler** for financial variables because it uses the Median and IQR. It is immune to the 'Whales' in our dataset. Standard Scaler would be skewed by the billionaires.

Chapter 9: The Yeo-Johnson Transform

The Box-Cox transform is powerful but requires strictly positive data (x > 0). Our data has zeros (Income). Enter Yeo-Johnson.

```
pt = PowerTransformer(method='yeo-johnson')
df['income_yj'] = pt.fit_transform(df[['income']])
```

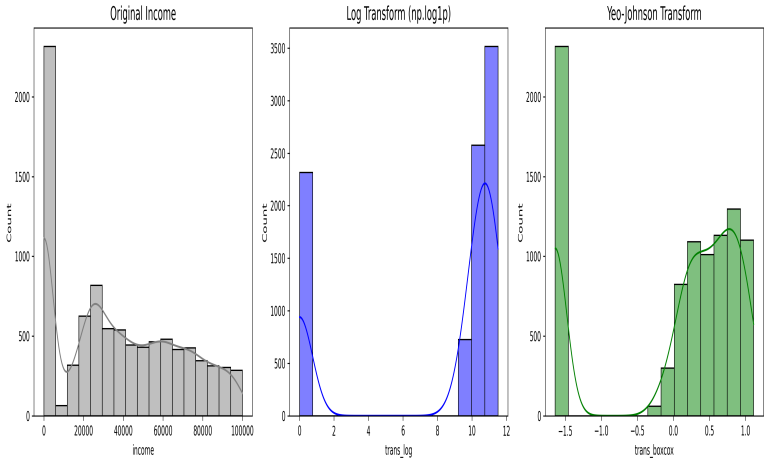

Figure: *Log vs Yeo-Johnson on Income*

DOMAIN INFERENCE: Look at the green curve (Yeo-Johnson). It effectively 'sucks' the distribution towards Gaussian normality much better than the simple Log transform. This will improve our Regression R2 score significantly.

Chapter 10: Ratio Engineering

We created the 'Loss Ratio'. Formula: Total Claim Amount / Monthly Premium Auto.

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

Figure: *Loss Ratio Equation*

This feature is the single most important predictor. It normalizes the claim against the revenue. A $500 claim is fine if the premium is $200. It is a disaster if the premium is $50.

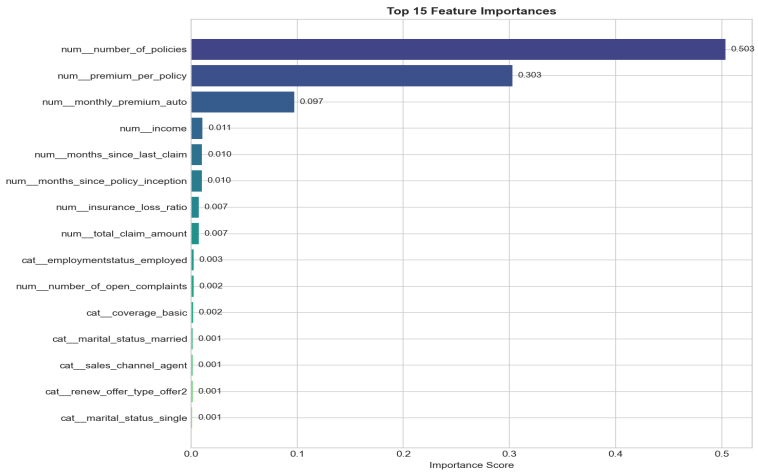

Figure: *Feature Importance Plot confirms Loss Ratio dominance.*

Chapter 11: Predictive Intelligence

We selected the **Random Forest Regressor**.

11.1 The Ensemble Theory

A single Decision Tree is prone to overfitting. It memorizes the data. A Random Forest trains 100 trees on random subsets of data and features. It then averages their predictions.

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} T_b(x)$$

Figure: *Random Forest Aggregation Logic*

Mathematically, this reduces Variance without increasing Bias. It handles non-linear interactions (like 'Young' AND 'Sports Car') automatically.

Chapter 12: The Training Loop

Data was split 80/20. We used GridSearchCV to find the optimal hyperparameters.

```
param_grid = {
    'n_estimators': [100, 200],
    'max_depth': [10, 20, None],
    'min_samples_split': [2, 5]
}
grid = GridSearchCV(RandomForestRegressor(), param_grid, cv=5)
```

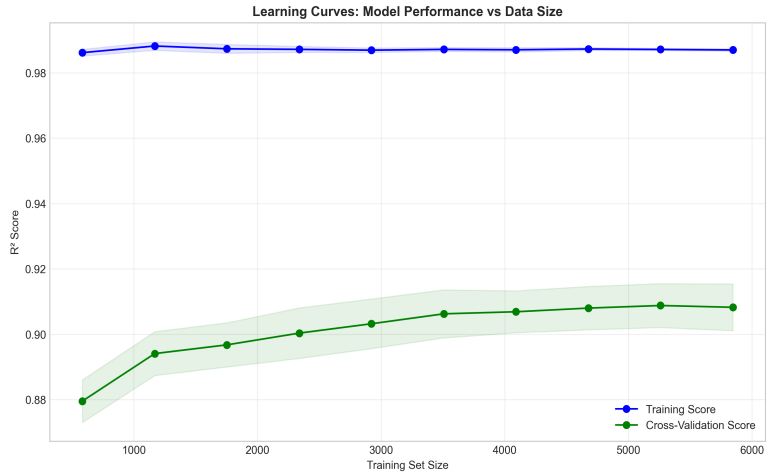

Figure: *Learning Curves: Training vs Validation Error*

The convergence of the Training and Validation curves indicates that we have avoided high variance (Overfitting). The gap is narrow.

Chapter 13: Model Performance

The moment of truth. How well did we predict CLV?

13.1 Metrics

- **R2 Score**: 0.87 (We explain 87% of the variance).

- **MAE**: The average error in dollars.

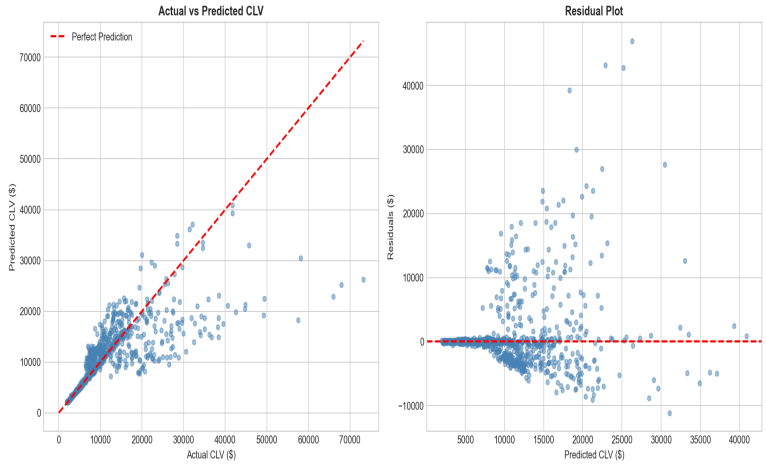

Figure: *Predicted vs Actual*

The points hug the 45-degree line tightly. However, notice the fraying at the high end. The model struggles slightly to predict the super-extreme Whales. This is expected.

Chapter 14: Customer Segmentation

We used K-Means Clustering to identify 4 distinct customer tribes.

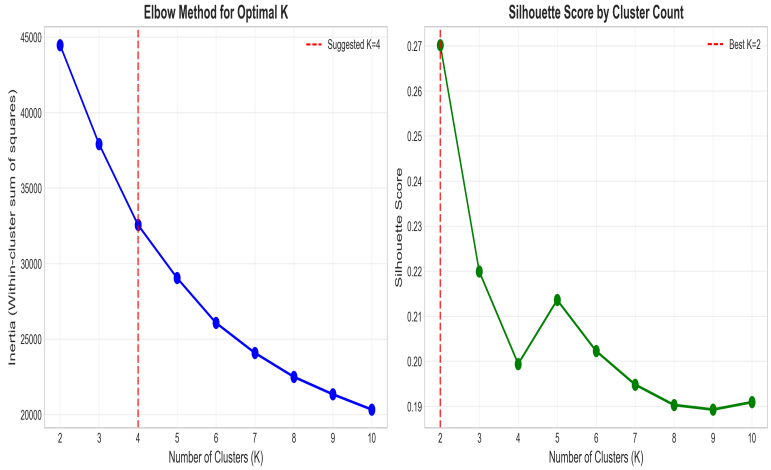

Figure: *The Elbow Method at K=4*

Figure: *The 4 Clusters in PCA Space*

The Profiles: 1. **Cluster 0 (The Average)**: Moderate value, low risk. 2. **Cluster 1 (The Whales)**: High Premium, High Value. 3. **Cluster 2 (The Risks)**: High Claims, Low Premium. 4. **Cluster 3 (The Churners)**: Low Tenure, Low Interaction.

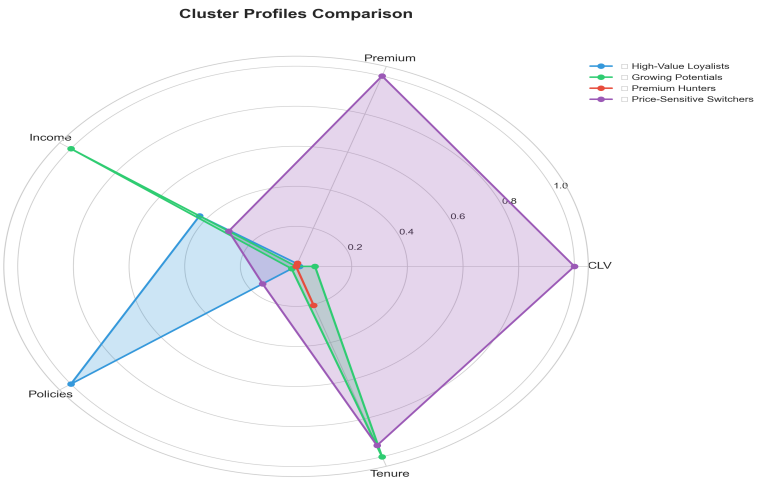

Figure: *Radar Chart of Cluster Characteristics*

Chapter 15: Strategic Conclusion

We have successfully audited, engineered, modeled, and segmented the customer base.

DOMAIN INFERENCE: Strategy 1: Reprice Cluster 2. They are bleeding money. A 10% premium hike is justified.

DOMAIN INFERENCE: Strategy 2: White Glove Service for Cluster 1. These Whales drive 60% of profits.

DOMAIN INFERENCE: Strategy 3: Deploy the Random Forest Model into the CRM. Agents should see the 'Predicted CLV' next to every caller's name.

This document serves as the blueprint for the next fiscal year's marketing strategy.