Application of Zero-Inflated Lognormal Loss Function for Predicting Cumulative Customer Lifetime Value

[Your Name]

August 16, 2024

Abstract

Predicting Customer Lifetime Value (CLV) is critical for customer segmentation, targeting, and overall business strategy. However, the zero-inflated and heavy-tailed nature of CLV data poses significant challenges to traditional regression models. In this report, we explore the application of the Zero-Inflated Lognormal (ZILN) loss function, a more tailored approach for CLV prediction. We describe the theoretical foundation of the ZILN loss, the architecture of the predictive model, and its implementation for forecasting the cumulative CLV over five years.

1 Introduction

Customer Lifetime Value (CLV) is a key metric in customer-centric marketing strategies, representing the total worth of a customer over their relationship with a business. Accurate CLV prediction allows companies to optimize marketing spend, improve customer segmentation, and increase overall profitability.

However, CLV prediction is challenging due to the nature of the data, which often includes:

- **Zero-Inflation**: A significant proportion of customers make only one purchase and never return, leading to many zero-value entries in the dataset.
- **Heavy-Tailed Distribution**: Among returning customers, a small fraction contribute disproportionately to the total CLV, resulting in a skewed distribution.

Traditional loss functions, such as Mean Squared Error (MSE), are not well-suited to handle these challenges. The Zero-Inflated Lognormal (ZILN) loss function provides a more appropriate framework by combining a Bernoulli distribution for zero vs. non-zero outcomes and a lognormal distribution for non-zero CLV values.

2 Theoretical Background

The ZILN loss function is designed to handle the specific characteristics of CLV data. It models the distribution of the target variable as a mixture of:

- **Bernoulli Distribution**: Captures the probability that a customer will have a non-zero CLV.
- **Lognormal Distribution**: Models the distribution of the log-transformed CLV values for customers with non-zero CLV.

The combined loss function is defined as:

$$L_{\text{ZILN}}(y, \hat{y}) = \begin{cases} -\log(1-p) & \text{if } y = 0 \\ -\log(p) + \log(\sigma\sqrt{2\pi}) + \frac{(\log(y)-\mu)^2}{2\sigma^2} & \text{if } y > 0 \end{cases}$$

where:

- p is the probability of non-zero CLV,
- μ and σ are the parameters of the lognormal distribution.

This formulation allows the model to learn both the likelihood that a customer will return (producing a non-zero CLV) and the distribution of their CLV if they do return.

3 Approach and Architecture

The implementation of the ZILN loss function within a predictive model involves several key components:

- **Model Architecture**: While various models can be used, including linear regression, gradient boosting, or deep neural networks, the architecture must be capable of handling both binary classification (for zero vs. non-zero CLV) and regression (for non-zero CLV values).
- **Loss Function Integration**: The ZILN loss function is integrated into the model training process. For complex models like deep neural networks, the ZILN loss can be used as the objective function to guide the optimization process.
- **Evaluation Metrics**: To evaluate the model's performance, traditional metrics like MSE may be insufficient. Instead, metrics such as the Normalized Gini Coefficient, Spearman's Rank Correlation, and Median Absolute Percentage Error (MdAPE) are more appropriate given the distributional characteristics of CLV data.

4 Case Study: Implementation for Cumulative CLV Prediction

In our case study, we focus on predicting the cumulative CLV over a five-year period. This is particularly challenging due to the long forecasting horizon and the potential for significant variance in customer behavior over time.

4.1 Data Preparation

The dataset used includes historical purchase data, customer demographics, and interaction metrics. The target variable is the cumulative CLV over the five-year period, calculated as the sum of all transactions within this timeframe.

4.2 Model Training

Given the complexity of the task, we chose a gradient boosting model (e.g., XGBoost) for its ability to handle non-linear relationships and interactions between features. The model was trained using the ZILN loss function to account for the zero-inflation and heavy-tailed nature of the data.

4.3 Evaluation

The model was evaluated using the following metrics:

- **Normalized Gini Coefficient**: To assess the model's ability to rank customers by their predicted CLV.
- **Spearman's Rank Correlation**: To evaluate how well the predicted ranks match the actual ranks of customers based on their cumulative CLV.
- **Median Absolute Percentage Error (MdAPE)**: To measure the accuracy of the model's predictions in percentage terms, providing a robust metric less sensitive to extreme values.

4.4 Results and Discussion

The model demonstrated a strong ability to discriminate between high-value and low-value customers, as indicated by a high Normalized Gini Coefficient and Spearman's Rank Correlation. The MdAPE also confirmed that the model's predictions were generally accurate, with lower sensitivity to outliers compared to MAPE.

5 Conclusion

The application of the ZILN loss function in CLV prediction addresses the specific challenges posed by zero-inflated and heavy-tailed data. Through our case study, we demonstrated that this approach significantly improves the model's ability to predict cumulative CLV over five years, making it a valuable tool for customer segmentation and targeted marketing.

Future work could explore the integration of additional customer attributes and the use of alternative models, such as deep neural networks, to further enhance prediction accuracy.