In-depth Analysis and Replication Guide for Customer Lifetime Value Prediction Approaches

Generated by ChatGPT

August 20, 2024

1 Introduction

Customer Lifetime Value (CLV) is a pivotal metric in business strategy, quantifying the total revenue expected from a customer over the duration of their relationship with a company. Accurate CLV predictions enable businesses to optimize marketing, sales, and customer service investments. This report delves into the methodologies presented in three significant papers on CLV prediction, offering detailed explanations, equations, model transformations, and implementation guidelines for replicating the approaches.

2 Paper Summaries

2.1 Billion-user Customer Lifetime Value Prediction: An Industrial-scale Solution from Kuaishou

2.1.1 Objective and Overview

This paper tackles the challenge of predicting CLV for a billion-user base, specifically within Kuaishou, a major short-video platform. The goal is to develop a scalable, accurate model that can operate in real-time and handle the vast, complex data generated by users.

2.1.2 Methodology

The approach combines traditional statistical models with advanced machine learning techniques to create a robust prediction model. The following methodologies and transformations are central to the model:

Gradient Boosted Decision Trees (GBDT) GBDTs form the core of the prediction model. GBDT is an ensemble learning method that builds multiple decision trees sequentially, with each new tree correcting the errors of the previous ones. The model minimizes the loss function, typically the mean squared error for regression problems, using gradient descent.

The mathematical representation of the model is:

$$f(x) = \sum_{m=1}^{M} \gamma_m h_m(x) + \epsilon \tag{1}$$

where $h_m(x)$ represents the m^{th} decision tree, γ_m is the learning rate, and ϵ is the residual error.

Embedding Layers for Categorical Features Given the large number of categorical variables (e.g., user actions, content types), the model uses embedding layers to convert high-dimensional categorical data into dense vectors. This reduces the dimensionality and captures relationships between categories.

The embedding process can be mathematically described as:

$$\text{Embedding}(\mathbf{c}_i) = \mathbf{V}_e \cdot \mathbf{c}_i \tag{2}$$

where $\mathbf{c}_i$ is the one-hot encoded vector of the categorical variable, and $\mathbf{V}_e$ is the embedding matrix.

Sequence Modeling User behavior over time is crucial for accurate CLV prediction. The model uses simple sequence models that can handle large-scale data efficiently. LSTM networks were considered, but due to computational constraints, simpler models such as Exponential Smoothing were used.

$$S_t = \alpha X_t + (1 - \alpha) S_{t-1} \tag{3}$$

where S_t is the smoothed statistic at time t, X_t is the actual value, and α is the smoothing parameter.

Scalability and Implementation The model is implemented on a distributed computing framework to handle the massive dataset. Apache Spark was utilized for its ability to process large-scale data in parallel. The GBDT model was distributed across multiple nodes, ensuring efficient processing.

2.1.3 Results and Application

The implementation of this model led to significant improvements in CLV prediction accuracy. The model's scalability allows it to be applied in any large-scale environment, particularly in digital platforms with high user activity.

2.2 Estimating Customer Lifetime Value Using Machine Learning Techniques (IntechOpen)

2.2.1 Objective and Overview

This paper explores various machine learning techniques for CLV estimation within the civil aviation industry, comparing traditional methods like RFM and NBD-Pareto with modern machine learning models.

2.2.2 Methodology

The paper details several approaches, focusing on both traditional and machine learning-based models:

RFM Model (Recency, Frequency, Monetary) The RFM model is a traditional statistical approach where the CLV is predicted based on:

- **Recency (R)**: Time since the last purchase.
- **Frequency (F)**: Total number of purchases made.
- **Monetary (M)**: Total amount spent by the customer.

The CLV is calculated as a weighted sum:

$$\text{CLV} = w_R \times R + w_F \times F + w_M \times M \tag{4}$$

where w_R, w_F, and w_M are the weights assigned to each component.

NBD-Pareto Model The NBD-Pareto model assumes that the number of transactions follows a Negative Binomial Distribution (NBD) and the timing between transactions follows a Pareto distribution.

The probability that a customer with x transactions will make y additional transactions in the future is:

$$P(Y = y|X = x) = \frac{\Gamma(r+x)}{\Gamma(r)x!} \left(\frac{\alpha}{\alpha+t}\right)^r \left(\frac{t}{\alpha+t}\right)^x \tag{5}$$

where Γ is the gamma function, r is the shape parameter, α is the scale parameter, and t is the time since the last purchase.

XGBoost XGBoost, an advanced machine learning model, is used to enhance predictive accuracy. The model uses gradient boosting to optimize the loss function, which is typically the logarithmic loss for binary classification or mean squared error for regression.

The objective function for XGBoost is:

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{6}$$

where ℓ is the loss function, $\hat{y}_i$ is the predicted value, and $\Omega(f_k)$ is the regularization term to prevent overfitting.

2.2.3 Comparison and Insights

The paper's comparison shows that while RFM and NBD-Pareto models provide easy interpretability, they often lack predictive power. XGBoost outperforms these traditional models, especially in handling non-linear relationships and interactions between features.

2.2.4 Results and Application

The XGBoost model demonstrated superior performance in predicting CLV for the civil aviation industry, making it applicable to other industries with complex customer behaviors. The implementation focuses on feature engineering, where categorical variables are encoded and missing data is imputed to ensure robustness.

2.3 Customer Lifetime Value Prediction: A Multi-relational Evaluation Model (Santos, 2018)

2.3.1 Objective and Overview

This paper introduces a Multi-Relational Evaluation (MRE) model for CLV prediction, extending traditional models by incorporating social relationships and interactions between customers, particularly in the context of civil aviation.

2.3.2 Methodology

RFMc Model The RFMc model is an enhancement of the traditional RFM model, incorporating a new variable, class of service (c), to account for the different values contributed by economy, business, and first-class passengers.

The formula is extended to:

$$\text{CLV} = w_R \times R + w_F \times F + w_M \times M + w_c \times c \tag{7}$$

where w_c accounts for the class of service.

Multi-Relational Evaluation (MRE) Model The MRE model evaluates the relationships between passengers based on their co-occurrence on flights, the frequency of flying together, and the time intervals between shared flights.

The MRE model uses a scoring mechanism:

$$\text{MRE}(i,j) = \sum_{k=1}^{K} \omega_k \times \text{rel}_k(i,j) \tag{8}$$

where $\text{rel}_k(i,j)$ represents the k^{th} relational factor between passengers i and j, and ω_k is the weight assigned to that factor.

Time Decay Factor A time decay factor τ is introduced to weigh recent interactions more heavily than older ones:

$$\tau(t) = e^{-\lambda(T-t)} \tag{9}$$

where T is the current time, t is the time of interaction, and λ is the decay rate.

Integration with XGBoost The MRE model is integrated with XGBoost to predict CLV, with the relational scores added as features in the model. This hybrid approach leverages both relational data and the predictive power of gradient boosting.

2.3.3 Key Insights and Results

The MRE model, combined with XGBoost, showed a substantial improvement in predictive accuracy, particularly in capturing the effects of social interactions on customer behavior. The model's performance was validated using real-world data from a major airline.

2.3.4 Results and Application

This approach is particularly effective in industries where social interactions play a crucial role, such as aviation, hospitality, and social networking platforms. The model can be adapted by incorporating relevant relational data and tuning the decay factor based on industry-specific timelines.

3 Discussion

The methodologies discussed demonstrate the evolution of CLV prediction models from simple statistical techniques to complex machine learning models. Each approach offers unique advantages, from the scalability of GBDT in large datasets to the relational insights provided by the MRE model. Businesses should consider the nature of their data and customer interactions when selecting a model.

4 Conclusion

In conclusion, these papers provide comprehensive methodologies for CLV prediction, each suitable for different business contexts. Whether dealing with large-scale platforms or industries with complex customer relationships, the models discussed offer robust frameworks for accurate CLV prediction. Future research should focus on enhancing these models' scalability and applicability across various sectors.

5 References

- Paper 1: Billion-user Customer Lifetime Value Prediction: An Industrial-scale Solution from Kuaishou.
- Paper 2: Estimating Customer Lifetime Value Using Machine Learning Techniques, IntechOpen.

- Paper 3: Customer Lifetime Value Prediction: A Multi-relational Evaluation Model (Santos, 2018).