Customer Lifetime Value Prediction Model Focus on High-Value Customer Reliability Analysis

Bowen Xu

{BowenXu-2@student.uts.edu.au}

Information Technology, University of Technology Sydney, Sydney, New South Wales, Australia

Abstract. Customer Lifetime Value (CLTV) prediction is a fundamental task in customer analytics. It is a challenging problem due to its highly skewed, long-tailed and unbalanced distribution, especially for high-value customers. This research uses three machine learning models: Random Forest, eXtreme Gradient Boosting (XGBoost) and Gradient Boosting to generate the future six-month CLTV and test the stability of these models in the high-value customer dataset. During feature engineering, a logarithmic transformation and inverse restoration strategy are used to reduce distributional skewness and improve model performance. Eventually, Gradient Boosting delivers the strongest overall performance, achieving R2 = 0.9613, RMSE = 459.7925 and MAE = 101.8458 on the full dataset. Gradient Boosting still maintains strong performance on high-value segments with R2 = 0.9406 on top25% dataset and R2 = 0.9161 on top10% dataset.

Keywords: Customer Lifetime Value, machine learning, logarithmic transformation, customer analytics.

1 Introduction

Customer Lifetime Value (CLTV) denotes the total worth of a customer that can be expected in the future as revenue. It is widely used for customer segmentation, marketing strategy and resource allocation. In real business environments, a large proportion of total revenue usually comes from a small group of high-value customers.

Customer Lifetime Value is widely used in marketing analytics to quantify the expected future economic value of a customer. Since most real-world transaction datasets do not directly provide CLTV labels, prior research commonly constructs CLTV targets from historical purchase records using probabilistic customer-base models. Among these approaches, the Beta-Geometric/Negative Binomial Distribution (BG/NBD) model estimates future purchase frequency by modeling heterogeneity in individual buying rates and dropout behavior and has become a standard and interpretable baseline for customer-level forecasting [1]. When combined with monetary models, this framework enables practical estimation of future customer value and provides a principled way of generating labels for supervised learning.

ICIAAI 2026, May 22-24, Singapore, Singapore

DOI 10.4108/eai.22-5-2026.2365222

Meanwhile, it is fundamental in customer analytics to transform original data into representative features. The Recency, Frequency, and Monetary value (RFM) framework summarizes purchase behavior in a compact way. It has been widely adopted for segmentation and value modeling due to its simplicity and reasonable performance [2]. RFM-style features are often used as inputs for predictive models that learn mappings from historical behavior to future value outcomes.

From a modeling perspective, ensemble learning methods such as Gradient Boosting have shown strong performance in capturing nonlinear interactions and complex feature dependencies. Gradient Boosting doesn't build a massive model at once. It constructs additive models through stage-wise optimization, enabling flexible approximation of complex functions while maintaining strong generalization capability [3]. In this study, the CLTV dataset is found to be highly skewed, nonlinear and unbalanced. These properties make boosting-based models suitable for CLTV prediction.

This study uses the Online Retail dataset from Kaggle. In the original dataset, every row represents information for a single order. Since CLTV prediction focuses on customers, the dataset is transformed so that every row represents a customer's RFM features. BG/NBD and Gamma-Gamma models are used to generate the value of CLTV in the future six months (6_month_clv) based on RFM features. Notably, this study significantly improved model performance by predicting the log value of CLTV and then applying a reverse method to reduce skewness. Finally, the three optimized models were applied to predict the high-value customer dataset (top 25% and top 10%). Their performances are compared and evaluated.

2. Methodology

2.1 Data preparation

The experiments are conducted on the Online Retail dataset obtained from Kaggle, which contains transactional records of an online retail business. The original dataset has 541909 rows and 8 columns. Each row represents a single item-level transaction associated with an invoice, including product quantity, unit price, transaction time, and customer identifier.

Before modeling, an exploratory data analysis (EDA) is performed to understand data quality, feature distributions, and potential data issues. The dataset contains both numerical and categorical attributes. The explanation of all features is shown in Table 1.

Table 1. Feature explanation

InvoiceNo	Unique identifier of a transaction.
InvoiceDate	Timestamp of the purchase.
Quantity	Number of units purchased in a transaction.
UnitPrice	Price per unit of the product.
CustomerID	Unique customer identifier.
StockCode / Description / Country	Product and geographic information.

Basic descriptive statistics reveal strong skewness and heavy-tailed distributions in transaction-level monetary values and quantities, indicating the presence of extreme purchases and a large number of small-value transactions. Such distributions are typical in real-world retail datasets and motivate careful feature aggregation and transformation in subsequent steps.
The analysis also confirms that the dataset is event-driven and ordered by time, which enables the construction of behavioral features based on temporal purchasing patterns.

2.2 Feature engineering

Since CLTV is defined at the customer level rather than the transaction level, it is necessary to aggregate raw transaction records into meaningful behavioral summaries. This work focuses on four core behavioral features derived from the classical RFM framework and its extensions. The four core features are demonstrated in Table 2.

Table 2. RFM features

Recency (R)	Time since the last purchase.
Frequency (F)	Number of transactions in the historical period.
T (Tenure)	Length of customer observation period.
Monetary (M)	Total historical spending.

These features capture complementary aspects of customer behavior: how recently the customer was active, how often they purchase, how much they spend, and how long they have been observed [2].

2.3 Transaction-Level to customer-level aggregation

The original dataset records purchases at the order-item level, where a single invoice may contain multiple product entries. To construct customer-level features, the data is first aggregated at the invoice level by summing quantities and transaction values for each invoice. Subsequently, invoice-level records are grouped by CustomerID to generate customer-level summaries.

Formally, for each customer i:

Frequency is computed as the number of unique invoices:

$$F_i = \text{count}(\text{InvoiceNo}_i) \tag{1}$$

Monetary Value is computed as the total transaction amount:

$$M_i = \sum_{j=1}^{N_i} \text{OrderValue}_{ij} \tag{2}$$

Recency is computed as the time difference between the reference date T_{ref} and the customer's last purchase date:

$$R_i = T_{ref} - \text{max}(\text{InvoiceDate}_i) \tag{3}$$

Customer Age (T) is computed as:

$$T_i = T_{ref} - \text{min}(\text{InvoiceDate}_i) \tag{4}$$

This aggregation transforms the raw transactional data into a compact customer-level dataset suitable for predictive modeling. After transformation, the dataset has 1916 columns for machine learning.

2.4 CLTV target construction using bg/nbd and gamma-gamma

CLTV is not a labeled variable in a transactional dataset. In the real world CLTV should be determined from historical purchasing behavior and projected into the future instead. To develop a sound supervised learning target, this research leverages the classical probabilistic modeling framework based on BG/NBD and Gamma-Gamma models.

The Beta-Geometric / Negative Binomial Distribution model uses two latent processes – the individual purchase rate and the probability of giving up (churn) – to estimate how many future transactions any customer is expected to undertake. Due to the historical recency and frequency of this transaction, the expected number of transactions over a future horizon can be expressed as:

$$E[N_i(t)] = f\left(R_i, F_i, T_i \mid \theta_{BG/NBD}\right) \tag{5}$$

where R_i , F_i , and T_i denote recency, frequency, and customer age, respectively, and $\theta_{BG/NBD}$represents the learned model parameters.

To estimate the expected monetary value per transaction, the Gamma-Gamma model assumes that transaction values follow a Gamma distribution and that customer spending heterogeneity can be captured probabilistically. The expected average transaction value for customer $i$is given by:

$$E[V_i] = g(M_i, F_i \mid \theta_{GG}) \tag{6}$$

where $M_i$is historical monetary value and θ_{GG}denotes model parameters.

By combining these two models, the predicted future CLTV over a time horizon $t$is computed as:

$$CLTV_i(t) = E[N_i(t)] \times E[V_i] \tag{7}$$

In this study, a 6-month forecasting horizon is selected to balance practical relevance and data stability. The resulting estimated CLTV values serve as pseudo-ground-truth labels for supervised machine learning models. This probabilistic CLTV construction follows the methodology proposed by Fader et al. [1], which has become a standard baseline in customer analytics.

2.5 Core point: log-transformation

An exploratory analysis of the estimated distribution of CLTV shows a strong long-tail pattern: most customers have relatively small future value while a tiny fraction of customers have very large CLTV values. The presence of heavy tails introduces challenges for regression modeling, as large values dominate the loss function and create instability in training.

Logarithmic transformation is widely used to stabilize variance and reduce skewness in economic prediction tasks [4]. So, to solve this problem, a logarithmic transformation is applied to the target variable:

$$y' = \log(1 + y) \tag{8}$$

The estimated CLTV distribution exhibits a strong long tail: most customers have relatively small future value, while a small number of customers have extremely large CLTV values. The heavy tails of the distribution complicate the regression modeling as the loss function will be dominated by the larger values causing instability in training.

After obtaining the prediction results of log (CLTV), the inverse transformation is applied to recover the real CLTV values:

$$y = \exp(y') - 1 \tag{9}$$

2.6 Modeling and evaluation

In this study, three models, which are Random Forest, Gradient Boosting, and XGBoost, are evaluated to identify a machine learning model for CLTV prediction.

By unifying all preprocessing operations into a single machine learning pipeline, I ensured reproducibility and prevention of data leakage. The pipeline performs type conversion to numeric, uses median for missing value imputation and scaling by standardization sequentially. The use of encapsulation of preprocessing steps in pipelines, ensures that the same transformations are done using train, validation and test sets. This is useful to ensure fairness in modelling and evaluation.

In addition, the pipeline works excellently with cross-validation and hyperparameter tuning, allowing for preprocessing and model fitting to be done together without any manual effort. This design will enhance the robustness of experimentation and ensure that model performance reflects true generalization and not artifacts of data processing steps.

Random Forest's pipeline is demonstrated in Figure 1 as sample.

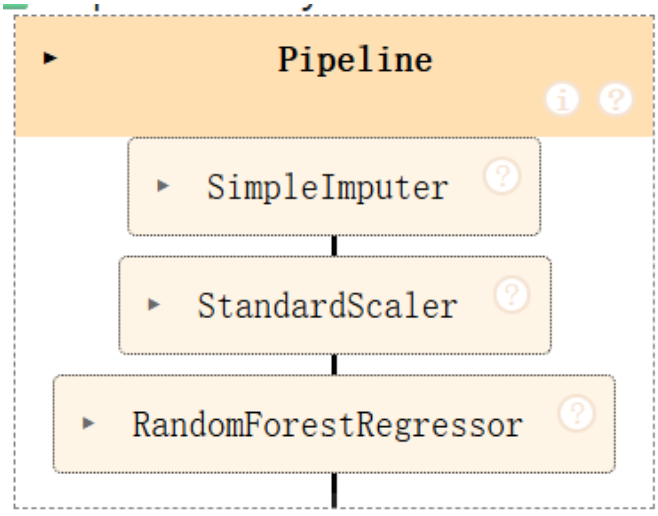

Fig. 1. Model pipeline (Picture credit: Original)

Random Forest is an ensemble learning model that builds a multitude of decision trees based on bootstrap sampling and random selection of features [3]. The predictions from every tree are

independent. Hence, the final output that is obtained is through averaging all tree predictions. Thus, it greatly reduces the variance and improves robustness against overfitting. The Random Forest predictor can be more formally written as.

$$\hat{f}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b\,(x) \tag{10}$$

where $T_b(x)$denotes the prediction of the b-th decision tree and $B$is the total number of trees.

The methodology called Gradient Boosting [5] generates an additive model in a stage-wise fashion and constructs it by sequentially using weak learners that fit the prior model residuals. Beginning with the new learner, which focuses on making amendments to the wrongs committed till now, the ensemble begins minimizing the loss function. The model update rule can be expressed as follows.

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{11}$$

where $h_m(x)$represents the newly fitted weak learner and $\gamma_m$is its corresponding step size.

XGBoost [6] is a major upgrade to GBM which makes the algorithm better and faster. The objective function minimizes training loss and model complexity together.

$$Obj = \sum_i l(\,y_i, \hat{y}_i) + \sum_k \Omega(\,f_k) \tag{12}$$

where $l(\cdot)$denotes the loss function and $\Omega(\cdot)$penalizes model complexity to reduce overfitting.

Due to the strong right-skewed nature of CLTV, all models are trained on its logarithm. Predictions, however, will be inverse-transformed back for evaluation. Models' performance is evaluated using R ², Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) on the original CLTV scale.

Five-fold cross-validation with inverse-log corrected scoring investigates reliable generalization assessment. Besides global performance measurement, the models are also examined on the top 25% high-value customers to see robustness under value concentration & long-tail effects. This evaluation setting is motivated by recent CLTV prediction studies. These studies emphasize highly skewed (long-tailed) customer value distributions and the importance of accurately modeling high-value segments [7].

3. Results

This part presents the empirical performance of three regression models, which are Random Forest [3], Gradient Boosting [5], and XGBoost [6] to predict six-month customer lifetime value. All models were trained on log-transformed CLTV targets and inverse log correction was used to evaluate the models on the original CLTV scale.

3.1 Model prediction performance

Table 3 summarizes the predictive performance across models using R ², MSE, RMSE, and MAE. Among the three models, Gradient Boosting achieves the best overall performance, followed by XGBoost, while Random Forest yields comparatively weaker results.

Table 3. Models' general performance

	RF	XGB	GB
R2	0.8591	0.8995	0.9613
MSE	770792.1474	549651.69	211409.1291
RMSE	877.9477	741.38	459.7925
MAE	185.5696	160.88	101.8458

The superiority of Gradient Boosting is particularly evident in error-based metrics, where it achieves the lowest RMSE and MAE. This indicates stronger robustness against extreme-value errors, which is critical under the heavy-tailed CLTV distribution. Prior work shows that many real-world prediction targets follow long-tailed distributions, where conventional evaluation metrics can be dominated by extreme samples and instability [8]. Logarithmic transformation has therefore been widely adopted to stabilize variance and improve learning robustness under such conditions [9].

XGBoost also demonstrates strong predictive accuracy and scalability advantages consistent with prior work [5]. In comparison, Random Forest performs less well since it is less effective at capturing complex nonlinear relationships [3]. These results further support the empirical observation that boosting-based models are often more suitable than bagging-based ensembles for difficult regression tasks such as CLTV prediction [5].

3.2 Prediction distribution analysis

The distribution of predicted values versus actual values for log(CLTV) and CLTV across three models is shown in Figures 2, 3, and 4.

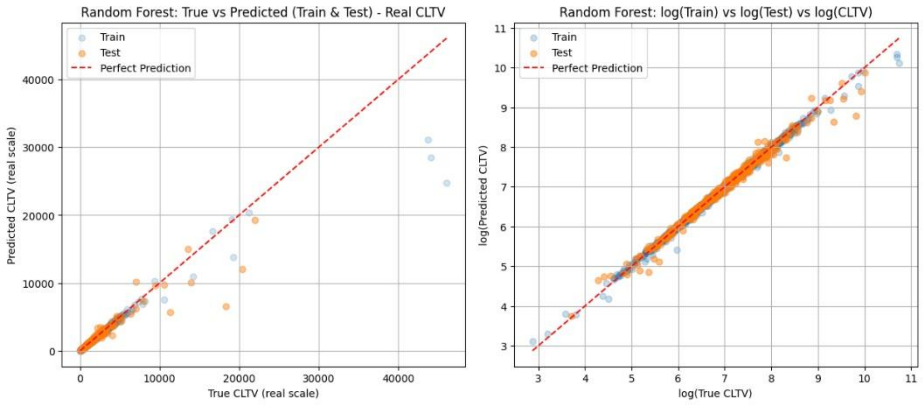

Fig.2. Random Forest performance (Picture credit: Original)

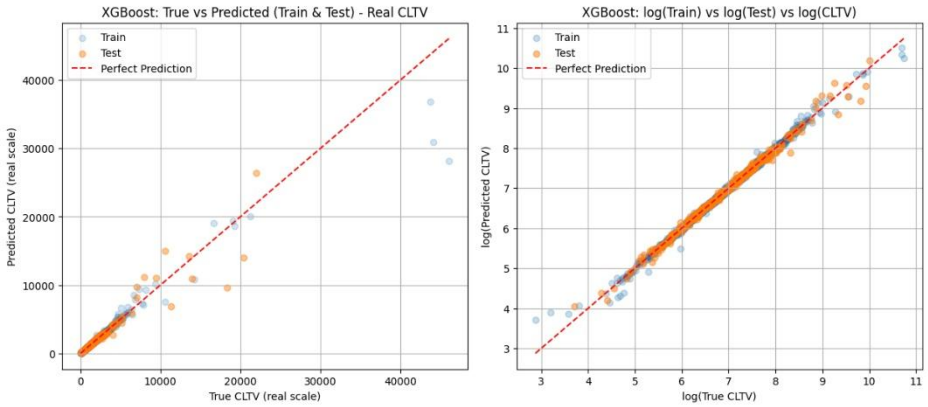

Fig. 3. XGBoost performance (Picture credit: Original)

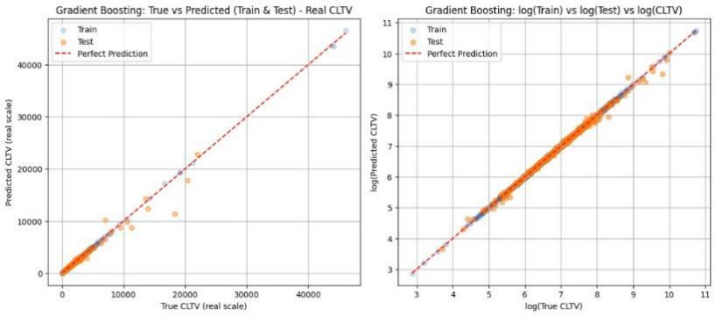

Fig. 4. Gradient Boosting performance (Picture credit: Original)

The predictions on the log scale are close to the reference diagonal for most samples, indicating stable learning behavior. After the inverse transformation, the spread of predictions increases with higher CLTV values, especially in the high-value customers range. Gradient boosting has a considerably tighter fit on the large-value samples than random forest and XGBoost. This behavior is consistent with past results that boosting models sequentially endeavor to fit hard-to-fit residuals and extreme values through stage-wise optimization [10]. This study shows that sequential residual learning helps model rare customers that generate 80% of revenues.

3.3 Feature importance analysis

The results of the feature importance of all three models are shown in Figure 5. All the models use monetary value and purchase frequency as similar contributors to prediction. While recency is the least contributor.

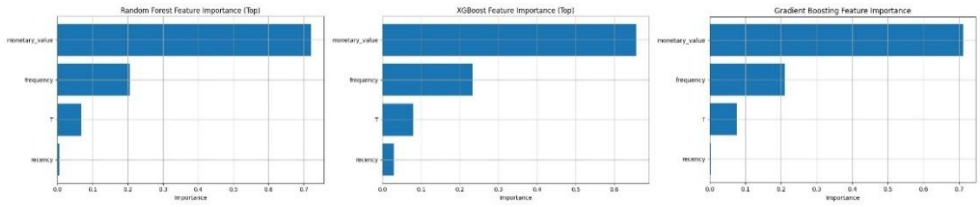

Fig. 5. Feature importance (Picture credit: Original)

This trend is concurred by the classical RFM theory, which states that the monetary and the frequency dimensions are the main drivers of future customer value [2]. The fact that recency does not contribute much to the prediction indicates that in a transactional retail setting, long-term spending intensity and engagement depth matter more than recent inactivity.

The findings from the Random Forest[3], Gradient Boosting [5] and XGBoost [6] show consistent importance rankings across these algorithms, indicating that the feature engineering design is robust and the architecture is not important.

3.4 Cross-Validation and overfitting analysis

The training and cross-validation performance differ substantially for Gradient Boosting as shown by the learning curve in Figure 6. According to the author of the paper [5], the observed behavior is consistent with moderate overfitting. This was expected as boosting ensembles are highly flexible, and regularization variants (e.g., stochastic boosting) are often used to reduce overfitting [10].

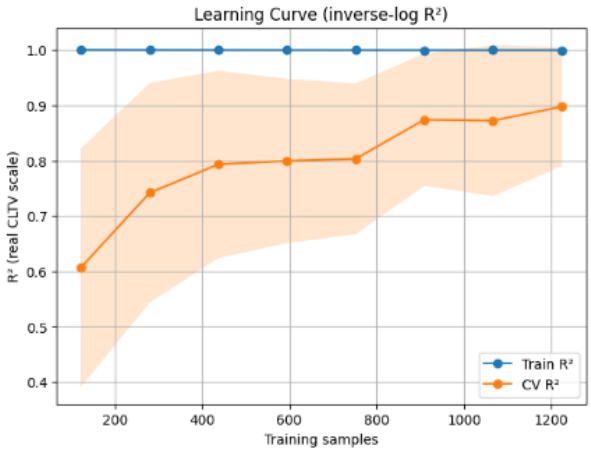

Fig. 6. Gradient Boosting learning curve (Picture credit: Original)

Nevertheless, the R^2 value from Gradient Boosting after inverse-log correction is highest through cross validation. Although the model may have a higher variance, it evidently generalizes better under controlled validation settings [4].

3.5 High-Value customer subset evaluation

To determine robustness that is commercially significant, the models are additionally assessed on a high-value customer subset defined by the top 25% and top 10% of the CLTV distribution. The results are demonstrated in Tables 4 and 5.

As the evaluation subset is made more selective, the relative advantage of gradient boosting increases, the results indicate. This demonstrates that optimal models behave differently under extreme-value regimes and that an explicit evaluation on high value segments is required to deploy reliably.

Table 4. Models' performance in the top 25% data

Model	R2	MSE	RMSE	MAE
Gradient Boosting	0.940569	842297.6	917.767721	338.817931
XGBoost	0.845181	2194211	1481.286781	564.700656
Random Forest	0.794855	2907473	1705.131431	621.511828

Table 5. Models' performance in the top 10% data

Model	R2	MSE	RMSE	MAE
Gradient Boosting	0.916112	2038440	1427.739595	680.328909
XGBoost	0.779048	5369015	2317.113426	1227.645090
Random Forest	0.710417	7036700	2652.677816	1243.775896

4. Conclusion

This study shows that logarithmic transformation is essential for improving CLTV prediction since it has highly skewed and long-tailed distributions. By modeling log-transformed CLTV and inverting the prediction results before evaluation, extreme values are stabilized, and model learning becomes more reliable and stable.

Among the evaluated models, Gradient Boosting has consistently given the best performance, though, overfitting was noticed in the CV evaluation. One possible reason could be that the model gradually corrects its previous mistake in a stage-wise manner, thus improving its ability to detect complex nonlinear relationships. This seems to be particularly useful for the CLTV dataset, where the high-value customers are sparse and the target distribution is highly skewed. Both Gradient Boosting and the logarithm are impressive implementations. Gradient boosting likely beats random forest and XGBoost so significantly on the high-value customers' dataset.

In summary, the results point out the appropriateness of logarithmic transformation in a highly skewed and heavy-tailed dataset. The implications of the results are that model evaluation should not just focus on overall accuracy, but also on how well it performs for high-value customers. This is often more appropriate in practical customer value management settings.

References

[1] Fader, P.S., Hardie, B.G.S., & Lee, K.L.: Counting your customers the easy way: An alternative to the Pareto/NBD model. Marketing Science. 24(2), pp. 275-284 (2005) doi: https://doi.org/10.1287/mksc.1040.0098

[2] Hughes, A.M.: Strategic Database Marketing: The Masterplan for Starting and Managing a Profitable, Customer-Based Marketing Program. 2nd ed. McGraw-Hill (2000). ISBN-10: 0071351825; ISBN-13: 978-0071351829.

[3] Breiman, L.: Random forests. Machine Learning. 45(1), pp. 5-32 (2001) doi: https://doi.org/10.1023/A:1010933404324

[4] Hastie, T., Tibshirani, R., & Friedman, J.H.: The elements of statistical learning: Data mining, Inference, and Prediction, Second Edition. Springer. (2009) ISBN-10: 0387848576. ISBN-13: 978-0387848570.

[5] Friedman, J.H.: Greedy function approximation: A gradient boosting machine. The Annals of Statistics. 29(5), pp. 1189-1232 (2001) doi: https://doi.org/10.1214/aos/1013203451

[6] Chen, T., & Guestrin, C.: XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 785-794. Association for Computing Machinery (2016) doi: https://doi.org/10.1145/2939672.2939785

[7] Curiskis, S., Dong, X., Jiang, F., & Scarr, M.: A novel approach to predicting customer lifetime value in B2B SaaS companies. Journal of Marketing Analytics. 11(4), pp. 587–601 (2023). doi: https://doi.org/10.1057/s41270-023-00234-6

[8] Zhou, H., Sun, K., Li, S., Fan, Y., Jiang, G., Zheng, J., & Li, T.: STATE: A robust ATE estimator of heavy-tailed metrics for variance reduction in online controlled experiments. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24). Association for Computing Machinery (2024) doi: https://doi.org/10.1145/3637528.3672352

[9] Yan, Y., & Resnick, N.: A high-performance turnkey system for customer lifetime value prediction in retail brands. Quantitative Marketing and Economics. 22(2), pp. 169-192 (2024) doi: https://doi.org/10.1007/s11129-023-09272-x

[10] Friedman, J.H.: Stochastic gradient boosting. Computational Statistics & Data Analysis. 38(4), pp. 367-378 (2002) doi: https://doi.org/10.1016/S0167-9473(01)00065-2