Title:

Enhancing Customer Lifetime Value Predictions with Advanced Machine Learning and Deep Learning Models

Shashwat Dhayade

1. Introduction:

1.1 Background Information:

Since the 1980s, customer relationship management (CRM) has increasingly become the focus of academia and industry. Facts have proven that maintaining good relationships with specific customers can increase corporate profits and increase the significant advantages of enterprises in market competition (Sun et al. 2023). Customer Lifetime Value (CLTV) is a critical metric for businesses, especially in industries like e-commerce and subscription-based services, where understanding customer behavior and long-term engagement is essential for strategic decision-making. CLTV helps businesses optimize their customer retention strategies, marketing investments, and overall profitability by predicting the potential value each customer will bring over their lifecycle (Sun et al. 2023). It serves as a cornerstone for customer relationship management (CRM), allowing businesses to tailor their approach toward high-value customers, improve customer loyalty, and reduce churn rates.

Traditional methods used for predicting Customer Lifetime Value (CLTV), such as Recency-Frequency-Monetary (RFM) models and Pareto/NBD probability models, often struggle to capture the complexities of modern consumer behavior. These techniques, which predict future transactions and dropout probabilities using a negative binomial distribution and Pareto model, respectively, rely heavily on predefined assumptions and limited customer data. This can lead to suboptimal predictions, particularly when applied to large datasets or noncontractual relationships where customer churn is less predictable (Vanderveld et al., 2016). While these traditional methods offer simplicity and ease of interpretation, they fall short in today's dynamic marketplace, where more robust and adaptive approaches are required to understand customer behavior.

RFM analysis and Pareto/NBD models, which rely on historical data to estimate CLTV, are well-suited for stable customer behavior patterns. However, RFM models tend to oversimplify customer dynamics, failing to account for a broader range of purchasing behaviors. Another traditional approach, the two-stage model, separates the CLTV prediction process into two parts: churn prediction and revenue estimation. While this method provides a structured approach to modeling customer behavior, it often oversimplifies the relationship between churn and revenue, missing key interactions (Weng et al., 2024).

Modern machine learning techniques offer a significant advancement over traditional methods, particularly when handling large and complex datasets. Deep learning models excel in capturing nonlinear relationships between features (Weng et al., 2024). Feed Forward Neural Network (FFNN) models can capture intricate feature interactions, while Long Short-Term Memory (LSTM) networks are

specifically designed to handle sequential data, making them ideal for businesses with irregular purchasing patterns. These models are well-suited for the unpredictable nature of noncontractual relationships, where customer retention and purchasing behaviors can vary widely.

Probabilistic models, such as the Zero-Inflated Lognormal (ZILN) model, provide another modern alternative by better capturing the mutable and complex nature of CLTV distributions. These models offer detailed insights into the likelihood of different customer outcomes based on historical behavior, which makes them particularly effective in environments with high customer variability (Weng et al., 2024). Ensemble methods, including Gradient Boosting Machines (GBM), XGBoost, and Random Forest, combine the predictions of multiple weaker models to improve overall performance. These techniques are scalable and provide enhanced prediction accuracy, particularly when dealing with large datasets and complex customer behavior patterns (Tsai et al., 2013).

While modern machine learning techniques provide superior accuracy and scalability, they also come with challenges. These methods are computationally intensive and require more resources to implement. Additionally, deep learning models are prone to overfitting if not properly regularized, and they are generally less interpretable compared to traditional methods, which can be a barrier for business decision-makers (Rajeshwari et al., 2024). Despite these limitations, the flexibility of modern techniques in capturing complex, nonlinear relationships and their ability to incorporate diverse user attributes makes them a powerful tool for CLTV prediction.

This study aims to advance the current understanding and application of CLTV prediction by addressing the limitations of both traditional and modern methods. The key contribution of this project is to improve the prediction accuracy and scalability of CLTV models through a combination of deep learning models and machine learning techniques. The study proposes a hybrid approach that leverages advanced models like FFNN and LSTM for capturing complex, non-linear interactions, as well as ensemble methods such as RandomForest and XGBoost to aggregate predictions and boost overall performance.

While prior research has demonstrated the superiority of modern techniques over traditional methods, there remains a gap in integrating these models to maximize predictive power while maintaining interpretability. This project intends to fill that gap by incorporating both the flexibility of deep learning and the scalability of ensemble methods. Furthermore, by introducing advanced techniques like automatic feature discovery through deep learning, the study will identify hidden patterns in customer behavior that traditional models might overlook, offering a more refined approach to CLTV prediction in noncontractual business environments.

By integrating cutting-edge models and addressing their potential limitations, such as overfitting and interpretability challenges, this study will contribute to the ongoing development of more accurate and actionable CLTV prediction models. This, in turn, will enable businesses to better target high-value customers, optimize marketing strategies, and improve overall profitability in increasingly competitive markets.

1.2 Previous Research

I contributed to this project by applying several machine learning models to predict Customer Lifetime Value (CLTV) and segment customers using an extensive dataset from an online retail platform. The dataset comprised over 1 million transactions, including detailed customer information such as invoice numbers, stock codes, product descriptions, quantities purchased, unit prices, and customer IDs. The project followed a systematic approach that involved key steps and methodologies aimed at maximizing the predictive accuracy of CLTV models.

One of the critical stages in the project was feature engineering. I conducted extensive feature extraction to enhance the predictive power of the models. This process included generating variables such as recency, frequency, and T (time since the first purchase). Additionally, I calculated the monetary value of customer purchases using metrics such as the mean, sum, minimum, and maximum purchase values. Other key features included the number of unique products purchased, unit price statistics (minimum, maximum, mean, and variance), and transaction metrics. The transaction metrics captured the frequency and monetary value of purchases made on weekdays, weekends, and specific periods, such as the first and last two weeks of each quarter. I also derived features related to customer behavior, such as the repurchase count and ratio, and the average time between orders.

To predict CLTV, I applied several machine learning models. One of the models used was the RandomForestRegressor, which I trained quarterly to predict future monetary value by progressively adding new quarters of data, following the approach suggested by Tietz (2018). Through feature importance analysis, I found that "monetary_weekdays" was the most significant predictor. However, despite the insights gained, the RandomForestRegressor exhibited a high root mean squared error (RMSE) of around 1000, suggesting that while the model provided some value, there was considerable room for improvement in predictive performance.

In addition to the RandomForestRegressor, I employed the Pareto/NBD model, a probabilistic approach, to predict customer lifetime probability and CLTV. This model operated on the assumption that customer transactions followed a Poisson process while customer dropout (churn) followed an exponential distribution. The RMSE for this model was significantly lower at 1.2, indicating that it was more effective for predicting CLTV in this specific context.

Both models provided valuable insights into customer behavior, but each had its limitations. The RandomForest model was useful for identifying important features, yet its high prediction error indicated that more advanced techniques, such as deep learning, could yield deeper insights and improved accuracy.

1.3 Objective

The primary objective of this proposal is to enhance the accuracy and efficiency of Customer Lifetime Value (CLTV) prediction by leveraging advanced machine learning and deep learning techniques. Specifically, the project aims to improve the current RandomForestRegressor model by incorporating additional algorithms, optimizing the feature set, and experimenting with neural networks. This initiative seeks to address the limitations of traditional models, particularly in noncontractual business environments where customer purchase patterns are often irregular.

One of the main objectives is to improve CLTV prediction by integrating advanced machine learning and deep learning models. Two specific models are proposed for this purpose: Feed-Forward Neural Networks (FFNN) and Long Short-Term Memory (LSTM) networks. FFNNs are particularly suited for capturing non-linear relationships between customer features that traditional models, such as linear regression, might overlook. FFNNs can automatically learn and model complex interactions among customer behaviors and transaction histories, making them crucial for accurately predicting CLTV in today's dynamic markets. In contrast, LSTM networks are designed to handle sequential data, which makes them ideal for businesses where purchase patterns are irregular. Unlike traditional models like RandomForest or Recency-Frequency-Monetary (RFM), which struggle to account for time-based dependencies, LSTM networks excel at learning from sequences of purchases. This allows LSTMs to capture long-term dependencies and provide more accurate predictions for businesses with unpredictable buying cycles. The goal of using these deep learning techniques is to reduce errors, such as Root Mean Squared Error (RMSE), and enhance the predictive power of CLTV models, particularly in scenarios with variable and complex customer behavior.

In addition to improving prediction accuracy, the proposal also explores the potential of hybrid models that combine machine learning and deep learning techniques. Specifically, ensemble methods such as RandomForest, Gradient Boosting Machines (GBM), and XGBoost will be evaluated. These models aggregate predictions from multiple weaker models, which helps to mitigate overfitting and improve generalization to new data. For example, RandomForest and XGBoost are capable of modeling complex customer behaviors by combining the outputs of multiple decision trees. While RandomForest is known for its robustness and interpretability, XGBoost, with its built-in regularization features, is particularly effective at preventing overfitting in large datasets. Similarly, GBM builds models sequentially, each one correcting the errors of its predecessor, allowing it to capture subtle nuances in customer behavior that simpler models may overlook. These ensemble methods are expected to enhance overall model performance by aggregating diverse predictions, thus improving accuracy and effectively modeling the variability in noncontractual customer relationships.

Another critical objective of this proposal is to optimize feature engineering using deep learning techniques. Traditional machine learning models often rely on manual feature engineering, which is time-consuming and prone to human error. In contrast, deep learning models, such as FFNNs and LSTMs, can automatically discover complex relationships between variables. For instance, FFNNs can uncover latent patterns and relationships among customer behaviors and attributes that may not be immediately visible through traditional feature engineering techniques. LSTMs, with their ability to analyze temporal data, can identify important patterns related to the timing of purchases and their influence on future value. By refining and automating feature interactions, these deep learning models will be able to model the intricate, non-linear relationships that traditional manual methods might miss. This capability is especially important in predicting CLTV, where customer behavior is influenced by numerous interrelated factors, such as purchase history, frequency, monetary value, and engagement metrics.

Through these approaches, the proposal seeks to significantly improve the accuracy and efficiency of CLTV prediction, providing businesses with better tools for decision-making and customer management.

2. Dataset

The dataset used for this project is tabular in nature and was derived from an online retail platform. It contains detailed transactional data about customer purchases. Some of the key features included in the dataset are invoice numbers, stock codes, product descriptions, the quantity of items purchased, the price of individual items, customer IDs, purchase dates and times, and the country of purchase.

This dataset originates from the UCI public repository, a well-known source of publicly available datasets for machine learning research. It consists of 1,067,371 instances, where each instance represents a single purchase transaction. In total, the dataset covers transactions for 5,880 unique customers, providing a rich foundation for understanding customer behavior and predicting Customer Lifetime Value (CLTV).

To ensure high data quality and enhance model performance, several pre-processing techniques will be applied to the dataset. This included handling missing values by removing records that lacked essential transaction data, such as customer IDs. Transactions with negative quantities will be removed to maintain data integrity. Additionally, key customer behavior features will be engineered from the raw transactional data to improve the predictive accuracy of the models. These features included Recency, Frequency, and Monetary value (with summary statistics like sum, mean, minimum, and maximum), the number of unique products purchased, and unit price statistics, such as mean and variance. Moreover, purchase frequency during weekdays, weekends, and holidays was also extracted to provide insights into customer purchasing patterns.

3. Methodology

The methodological approach will focus on refining the choice of models, optimizing feature selection, and employing deep learning architectures. A comparative analysis of various models will be undertaken to assess their effectiveness in improving CLTV prediction.

3.1 Choice of Models:

The RandomForestRegressor model will serve as the baseline for this project due to its capability to handle complex datasets and provide interpretable feature importance. Although RandomForest is adept at modeling non-linear relationships, its effectiveness is limited when dealing with sequential or highly variable data. To overcome these limitations, the project will focus on expanding the feature set and fine-tuning the hyperparameters to enhance its predictive performance in estimating future customer monetary value.

Gradient Boosting Machines (GBM) and XGBoost will be employed as ensemble models to aggregate the predictions from weaker models, thereby improving overall accuracy. These models are particularly effective in addressing the shortcomings of traditional methods, such as Recency-Frequency-Monetary (RFM) analysis, by modeling complex customer behaviors through sequential boosting. XGBoost, in

particular, will help mitigate overfitting—a common issue with deep learning models—through its emphasis on regularization, which enhances its robustness and generalizability.

Feed-Forward Neural Networks (FFNN) will be used to capture non-linear relationships among features that traditional models, like RandomForest or RFM, might overlook. The deep architecture of FFNN allows it to learn from intricate interactions within the dataset, making it highly effective for predicting Customer Lifetime Value (CLTV) in environments where customer behavior is complex and noncontractual.

Long Short-Term Memory (LSTM) Networks will be leveraged for their ability to handle sequential data and time-based dependencies. Given that CLTV prediction involves forecasting future behavior based on historical transactions, LSTM's capacity to learn from long-term dependencies will significantly enhance predictions, especially for customers with irregular purchasing patterns. This approach aims to address the sequential nature of transaction data, improving the accuracy of CLTV predictions by better modeling time-based dependencies.

3.2 Model Evaluation Metrics:

Root Mean Squared Error (RMSE) will be the primary metric for evaluating CLTV prediction accuracy, with lower values indicating better performance. The R^2 score will measure how well the model explains the variance in CLTV, with higher values signifying a better fit. Mean Absolute Error (MAE) will provide an average of the absolute differences between predicted and actual values, offering a straightforward assessment of prediction accuracy.

3.3 Model Interpretability:

For models like RandomForest and Gradient Boosting, feature importance analysis will be performed to identify which features most significantly influence CLTV prediction. This analysis is crucial for enhancing business decision-making, as it helps pinpoint the key drivers of customer value.

In the case of deep learning models such as Feed-Forward Neural Networks (FFNN) and Long Short-Term Memory (LSTM) networks, saliency maps and Grad-CAM (Gradient-weighted Class Activation Mapping) will be utilized. These techniques will visualize the impact of input features on the model’s predictions, offering insights into how these models arrive at their decisions and highlighting which transaction features are most influential in determining CLTV.

4. Outcome

By accurately predicting Customer Lifetime Value (CLTV), businesses can leverage these insights to make more informed strategic decisions across various areas beyond just marketing investments. One major area impacted by improved CLTV predictions is marketing. With a clearer understanding of which customers are likely to generate the most revenue over time, businesses can tailor their marketing strategies more effectively. High-CLTV customers can be targeted with personalized marketing campaigns, which maximizes customer engagement and retention. Additionally, these insights can lead

to cost-effective customer acquisition strategies by focusing on bringing in high-value customers, reducing customer acquisition costs (CAC) while increasing profitability. This enables businesses to allocate resources more strategically, improving return on investment (ROI) for marketing initiatives.

CLTV predictions also play a crucial role in enhancing customer retention efforts. By accurately identifying high-value customers, businesses can design tailored retention strategies such as loyalty programs, personalized offers, and rewards, ensuring long-term engagement and satisfaction. With more precise segmentation, companies can provide personalized interactions that not only improve customer loyalty but also boost customer lifetime profitability. Moreover, businesses can use these predictions to proactively engage customers who are at risk of churn. Early identification of these customers allows for personalized interventions, such as retention offers or enhanced services, which can significantly reduce churn rates, particularly for valuable customers.

Beyond marketing and retention, CLTV predictions can inform inventory management decisions. By forecasting demand based on the purchasing behavior of high-CLTV customers, businesses can optimize inventory levels, ensuring that key products are adequately stocked. This reduces the risk of stockouts for popular items while minimizing excess inventory of less valuable products. In turn, businesses can streamline supply chain operations, leading to more efficient inventory management. Additionally, CLTV data can guide product prioritization, enabling companies to focus on items that generate higher revenue from valuable customer segments.

Customer service is another critical area where CLTV predictions can have a significant impact. Businesses can prioritize high-CLTV customers by offering them more personalized and premium support services. This not only increases customer satisfaction but also fosters deeper loyalty among the most valuable segments. Customer service teams can also use CLTV data to offer tailored solutions based on the specific needs and history of each customer, improving the overall service experience and ensuring that high-value customers receive the attention they require.

CLTV predictions can also influence product development and customization strategies. Insights from CLTV data help companies optimize their product offerings by focusing on the features and services that are most valued by high-CLTV customers. This ensures that product development aligns with customer preferences, enhancing user satisfaction and driving long-term loyalty. Furthermore, businesses can develop data-driven product roadmaps that target their most profitable segments, ensuring that future product features or lines contribute directly to maximizing customer lifetime value.

Another important application of CLTV predictions is in pricing strategies. By integrating CLTV data into dynamic pricing models, businesses can offer tailored pricing based on customer value. For example, businesses may offer discounts to customers with lower predicted CLTVs to incentivize more spending, while maintaining premium pricing for higher-value customers. This approach balances profitability with customer retention and helps businesses maximize revenue across different customer segments.

Plan for Next Semester (Capstone 2):

In Capstone 2, the focus will shift more towards deep learning techniques to further enhance prediction accuracy and model performance. The plan includes refining Feed-Forward Neural

Networks (FFNN) by exploring more complex architectures to better capture non-linear relationships in the data, aiming to improve the model's predictive capabilities. Additionally, the use of Long Short-Term Memory (LSTM) networks will be expanded to analyze time-sequential customer transactions, offering deeper insights into temporal patterns in purchasing behavior.

Moreover, hybrid models combining traditional machine learning techniques, such as RandomForest, with deep learning models like LSTM or FFNN will be explored to develop solutions that leverage the strengths of both approaches. The objective for the next semester is to delve deeper into neural networks, further refine feature engineering through automated methods, and optimize the overall architecture to create scalable and deployable solutions for real-world applications.

5. Expected Outcomes:

Deployment in Real-Time Environments

Firstly, the trained models will be integrated into existing Customer Relationship Management (CRM) or Enterprise Resource Planning (ERP) systems to leverage real-time transactional data. This integration will be facilitated through API deployment, which will allow the CRM system to send customer data and receive predictions instantly. Tools such as Streamlit will be utilized to ensure seamless integration. Additionally, to support scalability and ensure real-time responsiveness, the models will be deployed on cloud platforms such as AWS (Amazon SageMaker), Google Cloud, or Microsoft Azure. These platforms offer auto-scaling capabilities and are well-suited for handling real-time data.

For real-time data feedback and continuous learning, monitoring systems will be put in place to track the accuracy and performance of the CLTV predictions. Metrics like Root Mean Squared Error (RMSE) and precision-recall will be continuously monitored to maintain prediction accuracy in evolving business environments. To address potential prediction drift caused by changes in customer behavior, the system will periodically collect new data and retrain the models. This retraining process can be triggered automatically based on performance metrics or manually based on new business insights, ensuring that the models remain adaptive to changing patterns.

Scalability and performance will be addressed through batch processing for large datasets, allowing for simultaneous predictions across large customer segments. This approach is particularly useful for managing the computational demands of deep learning models such as Feed-Forward Neural Networks (FFNN) and Long Short-Term Memory (LSTM) networks.

Finally, post-deployment analysis will involve comparing the models' predictions to actual customer lifetime values over time to identify any performance gaps. Based on this analysis, fine-tuning of hyperparameters or model architectures, such as increasing the number of LSTM layers or adjusting regularization parameters in FFNN, will be conducted to further enhance the models' accuracy and performance.

References

Chamberlain, Benjamin & Cardoso, Ângelo & Liu, C.H. & Pagliari, Roberto & Deisenroth, Marc. (2017). Customer Lifetime Value Prediction Using Embeddings. 1753-1762. 10.1145/3097983.3098123.

Sun Y, Liu H, Gao Y. Research on customer lifetime value based on machine learning algorithms and customer relationship management analysis model. Heliyon. 2023 Feb 13;9(2):e13384. doi: 10.1016/j.heliyon.2023.e13384. PMID: 36852044; PMCID: PMC9958434.

Vanderveld, A., Pandey, A., Han, A., & Parekh, R. (2016). *An Engagement-Based Customer Lifetime value system for e-commerce*. https://doi.org/10.1145/2939672.2939693

Harrison Tietz. (2018). Case Study in Customer Lifetime Value. *https://harrison4192.github.io/Simulacra/zodiac.html#*

Arefin, S., Parvez, R., Ahmed, T., Ahsan, M., Sumaiya, F., Jahin, F., & Hasan, M. (2024). *Retail Industry Analytics: Unraveling consumer behavior through RFM segmentation and Machine learning*. https://doi.org/10.1109/eit60633.2024.10609927

Tsai, C., Hu, Y., Hung, C., & Hsu, Y. (2013). A comparative study of hybrid machine learning techniques for customer lifetime value prediction. *Kybernetes*, *42*(3), 357–370. https://doi.org/10.1108/03684921311323626

B. Rajeshwari, Mallikarjuna Reddy Doodipala, K.Kiran Kumar Varma, P. Pattabhi Ram, Janardhana Rao N. (2024). Understanding Consumer Behavior in the Retail Sector Using RFM Segmentation and Machine Learning: An Analysis. *European Economic Letters (EEL)*, *14*(3), 412–420. https://doi.org/10.52783/eel.v14i3.1783

Weng, Y., Tang, X., Xu, Z., Lyu, F., Liu, D., Sun, Z., & He, X. (2024). OptDist: Learning Optimal Distribution for Customer Lifetime Value Prediction.