DATA4381 Final Report

Project Title:

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

Full Name: Shashwat Dhayade
Domain and Data Science Advisor: Dr. Rostami
Report Submission Date: December 6, 2024

---

1. Introduction

Predicting Customer Lifetime Value (CLTV) is a vital challenge in industries like e-commerce and subscription services, where understanding customer behavior directly impacts business strategies and profitability. CLTV models estimate how much value a customer will bring to a company over time. This helps businesses make smarter decisions about marketing, resource allocation, and keeping customers loyal. This project sought to address the limitations of traditional CLTV prediction methods by leveraging advanced machine learning and deep learning techniques, incorporating robust feature engineering, and addressing the challenges posed by complex and highly variable datasets.

Traditional approaches, such as Recency-Frequency-Monetary (RFM) models, often fail to capture non-linear relationships or time-dependent purchasing behaviors. In contrast, ensemble methods like Random Forest and XGBoost, and neural architecture such as Feed-Forward Neural Networks (FFNN) and Long Short-Term Memory (LSTM) models, provide the tools to uncover deeper insights. This project aimed to create a hybrid approach by combining these methods, leading to improved predictive accuracy and enhanced interpretability in real-world applications.

---

2. Background

Traditional CLTV prediction methods, while interpretable and straightforward, have significant drawbacks when applied to dynamic consumer behavior. RFM models oversimplify customer interactions, relying on aggregated metrics that often fail to capture nuanced or evolving

patterns. Probabilistic models like Pareto/NBD perform well in stable environments but struggle with irregular purchasing patterns or noncontractual relationships.

Modern machine learning techniques address these challenges by leveraging computational power and advanced algorithms to model complex relationships. Ensemble models, such as Random Forest and XGBoost, combine the outputs of multiple weak learners to improve accuracy and generalizability. Neural networks further expand this capability by learning intricate patterns and dependencies within the data, particularly in sequential tasks where LSTMs excel.

This project's significance lies in its dual focus: advancing predictive capabilities through cutting-edge techniques and exploring the balance between interpretability and performance. By addressing the limitations of traditional methods, this work tries to contribute to the actionable insights for businesses navigating highly competitive markets.

---

3. Methodology

This project followed a structured methodology encompassing data understanding, preprocessing, feature engineering, modeling, and fine-tuning to tackle the challenges posed by a complex dataset. Each phase was essential to address issues such as skewed distributions, outliers, and feature dominance, while also ensuring scalability and interpretability.

3.1 Data Understanding

The dataset used in this project was obtained from the UCI public repository, a widely respected source for machine learning research. It comprised over one million transactions from an online retail platform, covering a two-year period and representing 5,853 unique customers. Key attributes included invoice numbers, product descriptions, purchase quantities, unit prices, and customer information such as country of origin.

Initial exploratory analysis revealed several issues with the dataset, including duplicate records, missing values, and extreme outliers. For instance, significant anomalies such as a $77,000 transaction that was later refunded skewed the dataset. The analysis also highlighted strong skewness in critical features, particularly monetary and frequency-related metrics, which required transformation and careful handling during feature engineering.

Further analysis using natural language processing (NLP) techniques focused on refund-related product descriptions, uncovering six distinct clusters, including kitchen items, decorative

products like Regency cake stands, and lighting items. These insights into refunded product types informed subsequent feature engineering steps.

3.2 Data Preprocessing

Data preprocessing focused on cleaning, standardizing, and refining the dataset to ensure it was suitable for modeling. Duplicate entries were removed, while irrelevant stock codes, such as those representing delivery charges or test products, were excluded. Missing values were addressed by eliminating transactions without essential information, such as customer IDs or valid monetary values.

Special attention was given to outliers. Customers exhibiting extreme or non-representative behaviors, such as high refunds with minimal purchasing activity, were flagged for removal. Additionally, the skewness in monetary and frequency metrics was mitigated through scaling and normalization techniques, enabling the models to better handle these distributions.

3.3 Modeling

The modeling phase employed both traditional machine learning techniques and deep learning architectures to predict CLTV. A Random Forest model using Recency-Frequency-Monetary (RFM) metrics and 23 other features was developed to establish feature importance. Monetary Weekdays emerged as the dominant predictor, accounting for the highest variance in outcomes (Figure 1).

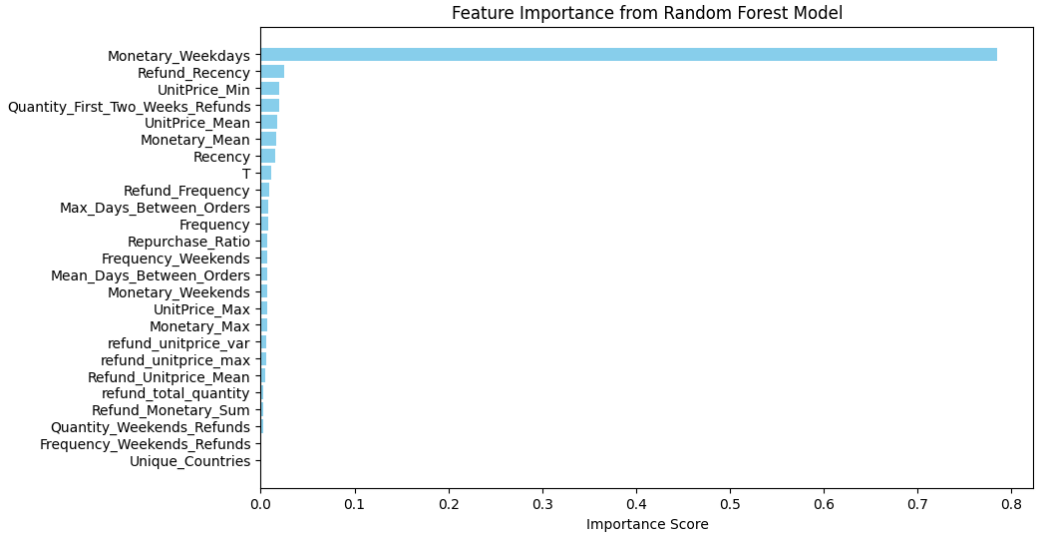

Figure 1: Feature Importance Plot

Deep learning approaches, including Sequential Neural Networks, were explored to capture non-linear relationships. However, these models struggled due to convergence issues, dataset imbalance, and over-reliance on dominant features like Monetary Weekdays.

Feature reduction using Principal Component Analysis (PCA) was introduced to address multicollinearity and computational complexity. While PCA retained 90% of the data's variance, its application did not significantly improve model performance, as evident from persistently low R-squared values. As seen from the scree plot (Figure 2) PCA is not able to figure the optimal number of components.

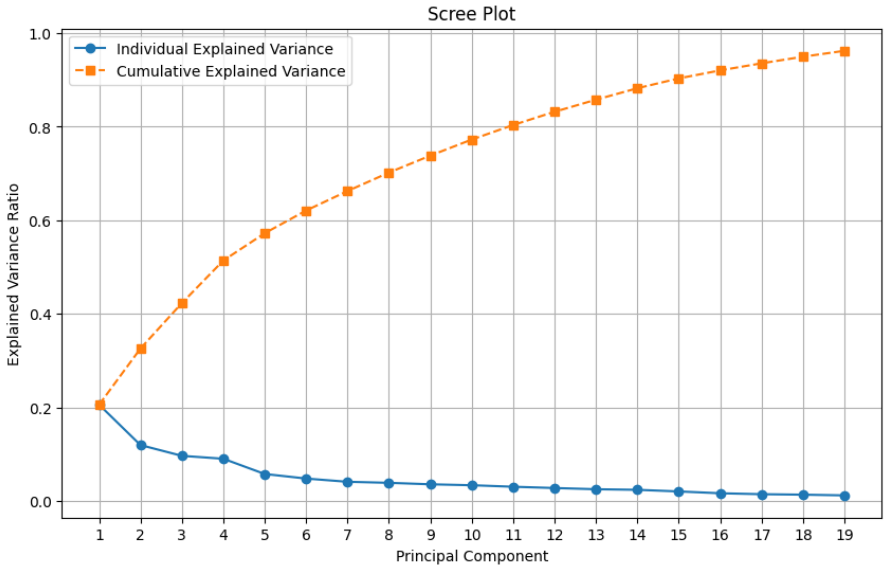

Figure 2: Scree Plot

3.4 Fine-Tuning

Fine-tuning involved iterative testing and refinement of features. Refund and discount metrics were incorporated into the dataset to capture customer behavior patterns more comprehensively. Multicollinearity was tackled through correlation analysis, leading to the removal of redundant features like "Discount Frequency" and "Refund Monetary Mean."

Separate Random Forest models were built with and without Monetary Weekdays to assess its influence. The inclusion of this feature improved performance, but its dominance masked the contributions of other variables. Neural networks underwent additional tuning but continued to struggle with capturing the dataset's complexity, highlighting the need for further architectural adjustments.

4. Challenges and Solutions

The project encountered significant challenges at multiple stages, particularly during data preprocessing and feature engineering. One of the most pressing issues was the presence of outliers, such as unusually high refunds or bulk purchases on single days, which distorted the dataset. For instance, customer 12346 made a single purchase worth $77,000, which was immediately refunded. While cases like this were removed, other high-value customers with consistent behavior were retained which made the data consistently imbalanced.

Feature engineering presented additional difficulties, particularly in managing multicollinearity. Refund and discount-related features often overlapped with monetary metrics, introducing redundancy that affected model performance. A correlation matrix was employed to systematically identify and remove redundant features, reducing the dataset to 26 final features (Figure 4).

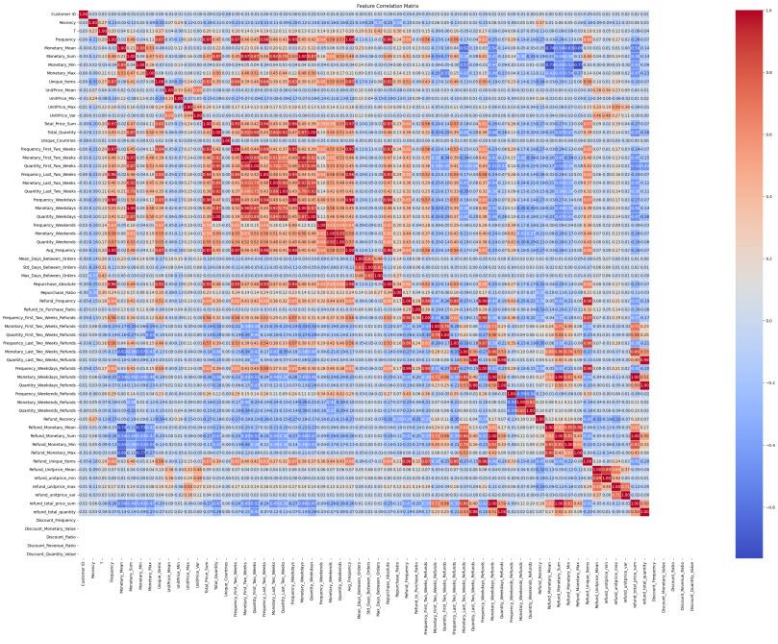

Figure 3: Correlation matrix with multi collinearity problem

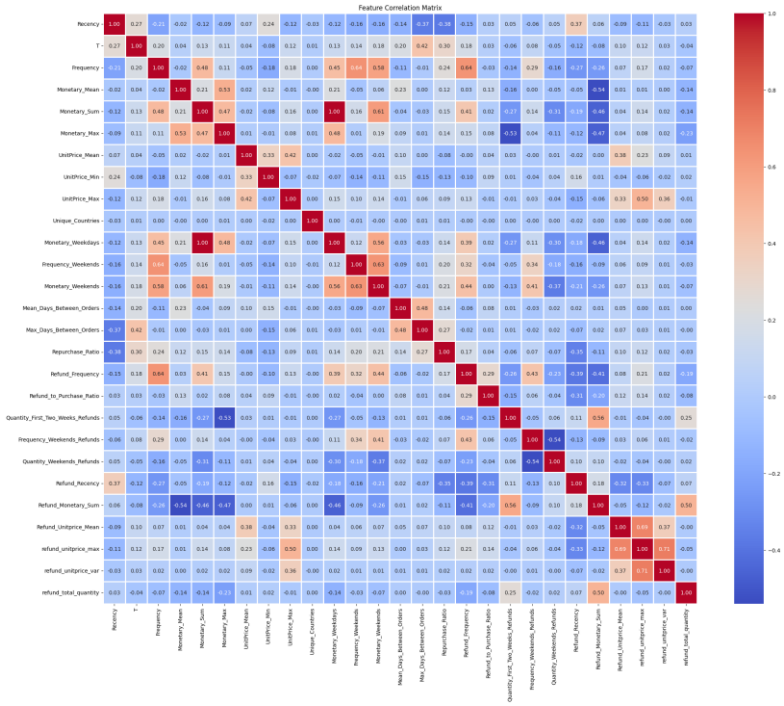

Figure 4: Correlation Matrix after removal of correlated features

Low R-squared values persisted across models, reflecting limited predictive accuracy. Advanced techniques, such as PCA and feature scaling, were applied to mitigate this issue, but the improvements were incremental. Deep learning models, despite their theoretical ability to capture complex patterns, struggled due to dataset imbalances and overfitting. Addressing these challenges required extensive experimentation with preprocessing methods and model architectures.

---

5. Key Results and Insights

The Random Forest model consistently outperformed other approaches, achieving the best R-squared value of 0.498 when Monetary Weekdays was included as a feature. Without this feature, the R-squared value dropped to 0.369, underscoring its influence on model performance. However, this dominance limited the exploration of other potentially informative variables.

Principal Component Analysis (PCA) was applied to reduce feature dimensionality, but its impact was limited. With PCA, the Random Forest model achieved an R-squared value of negative 0.093 when Monetary Weekdays was included and 0.134 when it was excluded. These

results indicated that PCA struggled to retain critical information necessary for accurate predictions.

Neural networks also exhibited limited success, with a test mean absolute error (MAE) of 250.77 when Monetary Weekdays was included and 277.40 when it was excluded. The training and validation loss curves revealed that the model failed to converge effectively, likely due to the dataset's imbalance and the dominance of a few features (Figure 5).

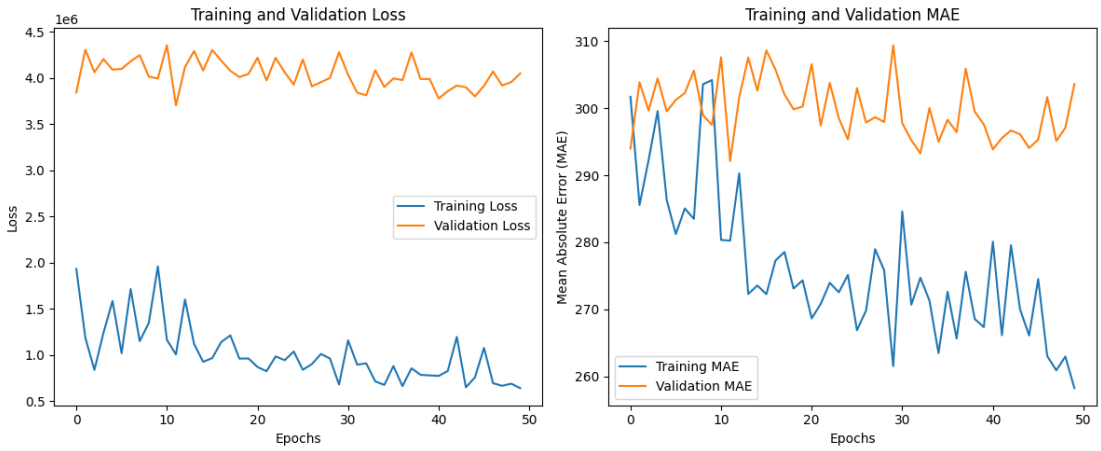

Figure 5: Training and Validation Loss Plot

NLP analysis of product descriptions provided valuable insights into refund patterns, identifying six clusters of frequently refunded items. These clusters, which included categories like kitchen and decorative items, can offer actionable insights for businesses seeking to minimize refund rates.

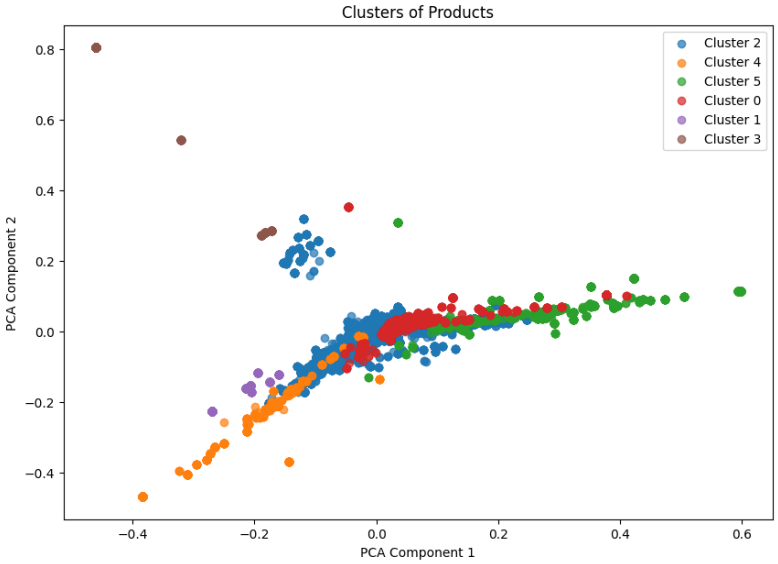

Figure 6: Clustering of Refunded Products

---

6. Conclusion

This project highlighted the complexities of predicting Customer Lifetime Value in noncontractual business environments. While Random Forest models demonstrated moderate success, the reliance on features like Monetary Weekdays limited their generalizability. PCA and neural networks failed to deliver significant improvements, underscoring the challenges of handling imbalanced datasets and dominant features.

The work underscored the importance of robust feature engineering and comprehensive preprocessing to address data irregularities. Although the project did not achieve optimal predictive accuracy, it provided valuable insights into customer behavior patterns, particularly through the analysis of refund trends. These findings lay the groundwork for future research into more effective and interpretable CLTV prediction models. It also helped in understanding the method of problem-solving for any other situations.

---

7. Plans for Capstone 2

The next phase of this research involves collaborating with Dr. Shen. I worked on a test project to predict chemical ecotoxicity using a dataset of 2,122 chemicals and 14 physiochemical features. This project involved employing machine learning techniques, including Random Forest, Support Vector Machines, and Multilayer Perceptron, to identify key predictors of toxicity.

Building on the lessons learned from this project, the focus will be on a similar topic. Different machine learning techniques will also be explored to enhance model interpretability and ensure the findings are actionable in real-world scenarios.