Title:

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

Full Name: Shashwat Dhayade

Report Submission Date: 11/22/2024

1. Introduction

Over the past two weeks, my focus was on data cleaning, feature engineering, and understanding outliers in the dataset to improve the performance of the Customer Lifetime Value (CLV) prediction model. Additionally, I used Random Forest, PCA, and Neural Networks for modeling. This report summarizes the tasks completed, challenges faced, and future plans to refine the project further.

---

2. Summary of Work Done

Data Cleaning:
Duplicates were removed, along with irrelevant stock codes like POSTAGE, identified as delivery charges rather than customer purchasing behavior. The resulting dataset comprised 794,242 records with 5,853 unique customers. I split the dataset into training (70%), validation (15%), and test (15%) subsets for model development.

Random Forest Modeling:
A Random Forest Regressor was implemented on the cleaned dataset, yielding an R-squared value of 0.519. Applying a Standard Scaler marginally reduced the R-squared to 0.516. Feature importance analysis revealed that *Monetary Weekdays* was the most influential feature, followed by *Unit Price Max* and *Recency*. However, the high correlation between *Monetary Weekdays* and *Monetary Sum* (revenue) likely overshadowed other features as seen in Figure 1 and 2.

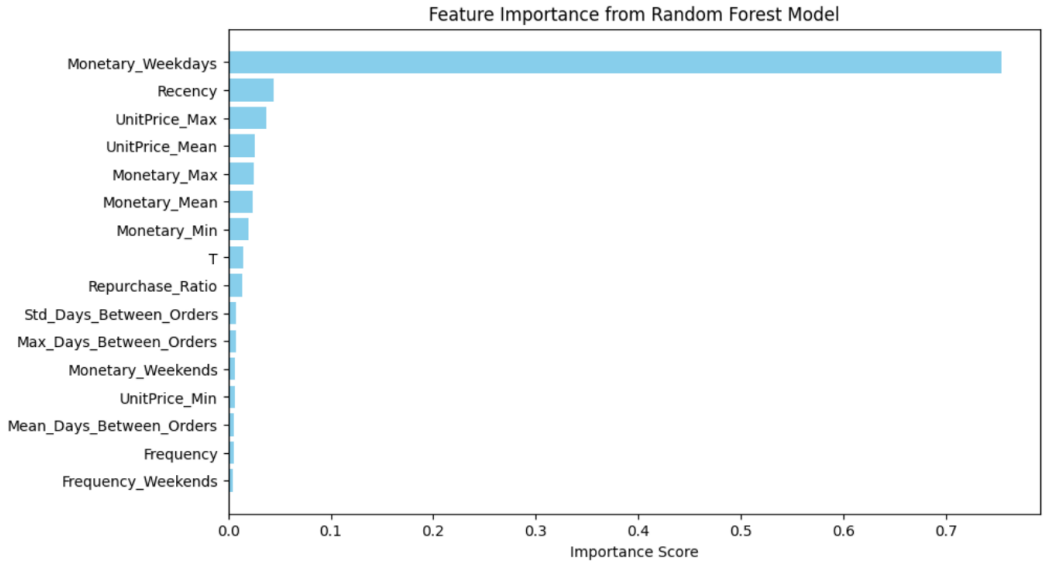

Figure 1: Feature importance plot for RandomForest

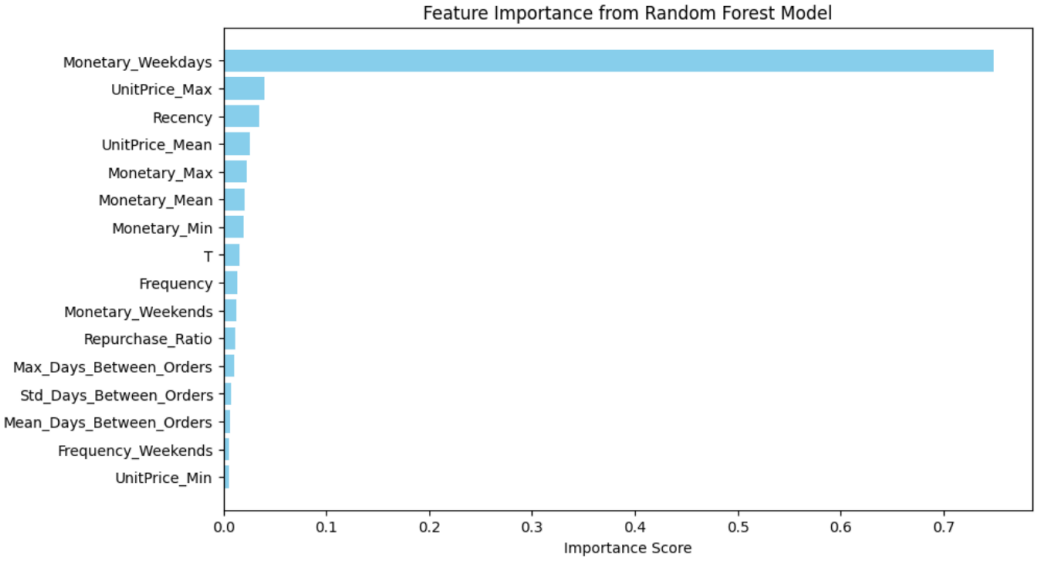

Figure 2: Feature Importance plot after Scaling

PCA Analysis:

Principal Component Analysis (PCA) showed that 90% of the variance was explained by 9-10 principal components seen in Figure 4. The PCA-based model achieved an R-squared value of 0.41, lower than the Random Forest model. Feature importance indicated that Principal Component 6 was the most predictive of *Monetary Sum (Figure 3)*.

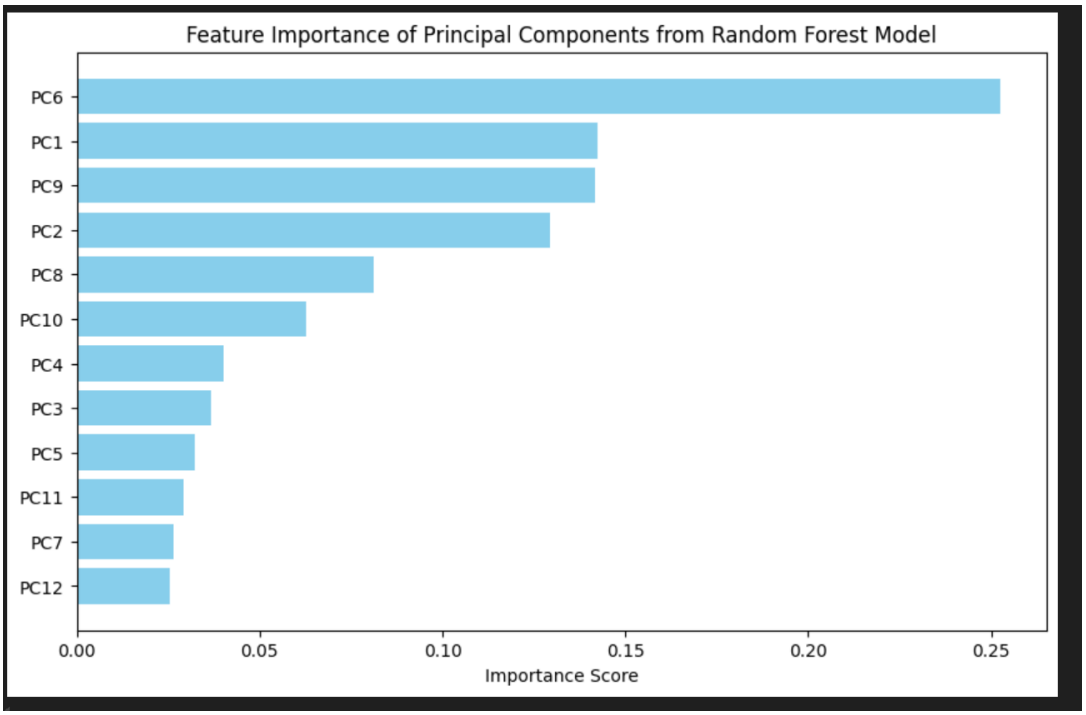

Figure 3: feature Importance plot for each Principal component

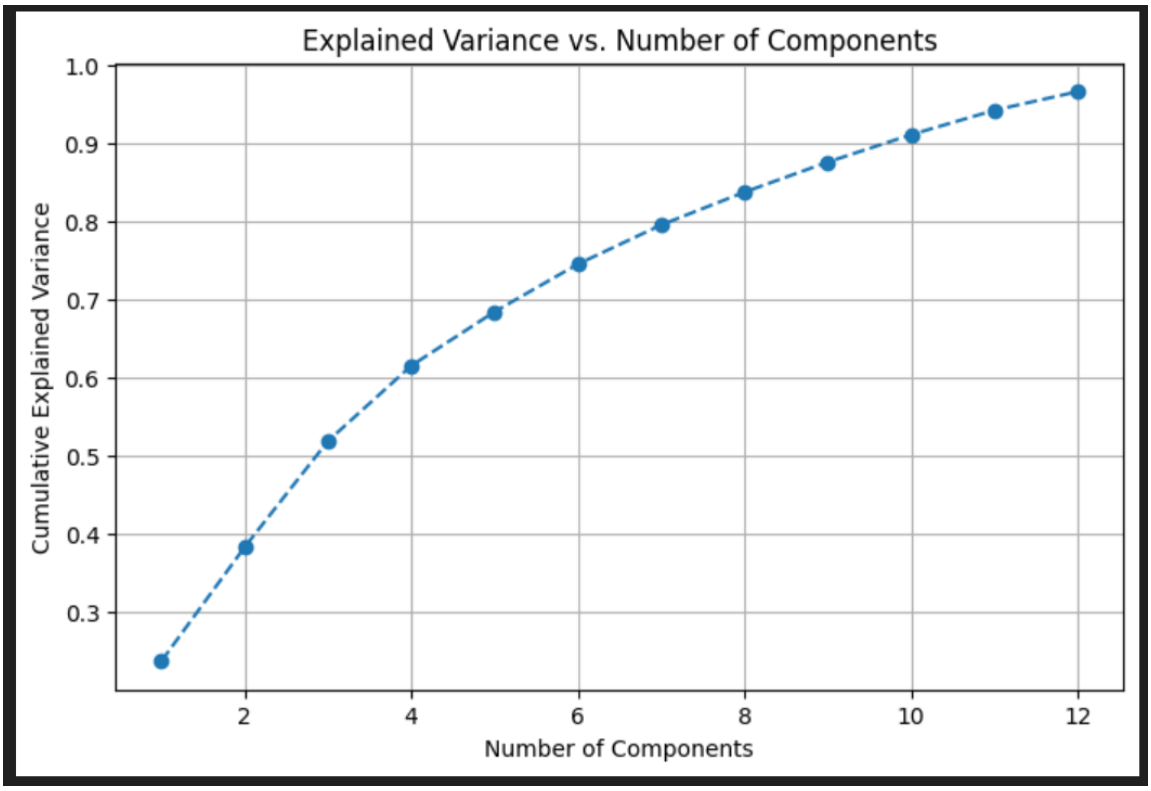

Figure 4: Variance plot for PCA

Neural Network Modeling:

I attempted to train a Neural Network model, but encountered significant challenges related to mismatched input shapes between the training, validation, and test datasets. For example, to explain, the approach I used for training involves leveraging data from the first quarter and using the corresponding labels or target variables from the second quarter to predict the labels for customers in the third quarter (Figure 5). This structure introduced missing values, as not every customer makes purchases in every quarter. Additionally, different product descriptions, numbers of purchases, and customer behaviors across quarters created inconsistencies in the dataset.

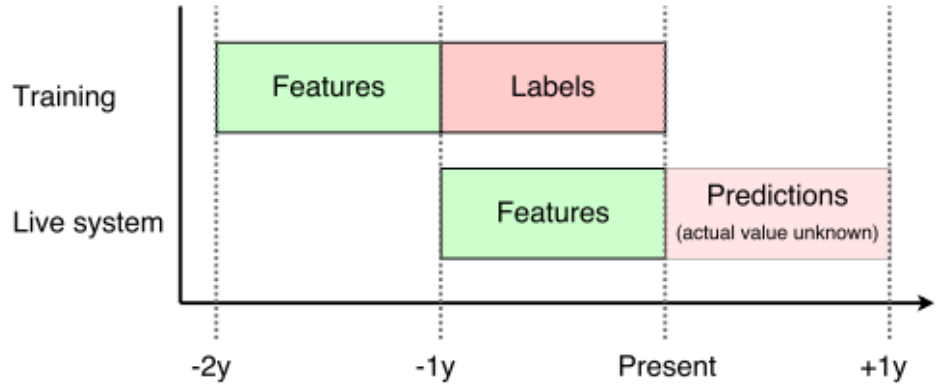

Figure 5: Training Split Method

To address these issues, I aggregated all customer data for 3 years and assigned a value of zero to customers who did not have certain purchase behavior like refund. This step ensured uniform shapes across the training, validation, and test datasets. However, this approach also necessitated losing details such as specific product information or precise timestamps (e.g., year, month, weekday, or hour). While splitting time-based data into these finer categories was suggested by Dr. Rostami, it proved unfeasible because customer data had to be aggregated, and combining all purchases meant losing the specific time data.

Despite ensuring consistent dataset shapes, the Neural Network model still failed to learn effectively. The validation loss and training loss, along with the mean absolute error (MAE) for both, showed no convergence and did not follow any discernible pattern (Figure 6). These results indicate that further refinements to the dataset or the model architecture are necessary to address these challenges effectively.

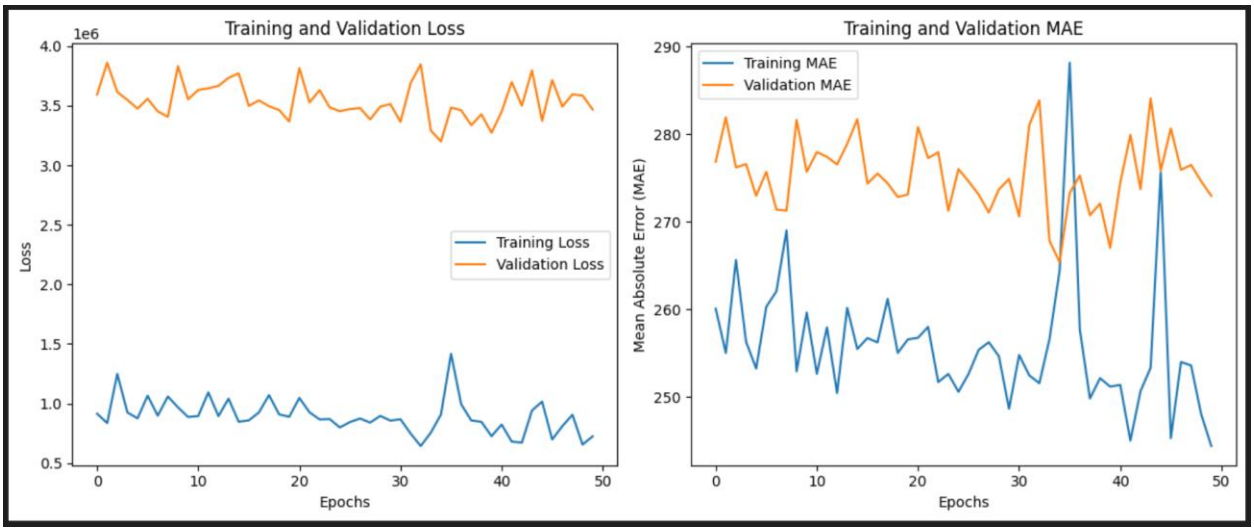

Figure 6: Performance Plot

Feature Engineering for Refunds and Discounts:

In this phase, I focused on incorporating chargeback fraud, and refund-related features into the dataset, as suggested by Dr. Rostami and Utkarsh, to capture customer behavior. A total of 63 new features were created, including:

- Refund Behavior Metrics:
  Features such as *Refund Frequency*, *Refund-to-Purchase Ratio*, *Refund Recency*, and various refund-related monetary and quantity metrics for specific time periods (e.g., weekdays, weekends, first two weeks, last two weeks).

  Random Forest Model Performance with Refund Features:

  With the new features, I trained a Random Forest regression model. The model's R-squared value improved slightly to 0.528, up from previous iterations, indicating that the additional features provided marginal predictive improvements.

  Monetary Weekdays remained the most influential feature, with a high importance score, but its dominance may have overshadowed other features. Recency ranked as the second most important feature, followed by Unit Price Max. Refund-related metrics, such as *Refund Recency* and *Refunds in the End of the Month*, also appeared in the top features, suggesting that refund patterns significantly impact customer behavior (Figure 7).

  Despite these improvements, *Monetary Weekdays* continued to dominate the feature importance rankings, highlighting potential redundancy due to its high correlation with *Monetary Sum*.

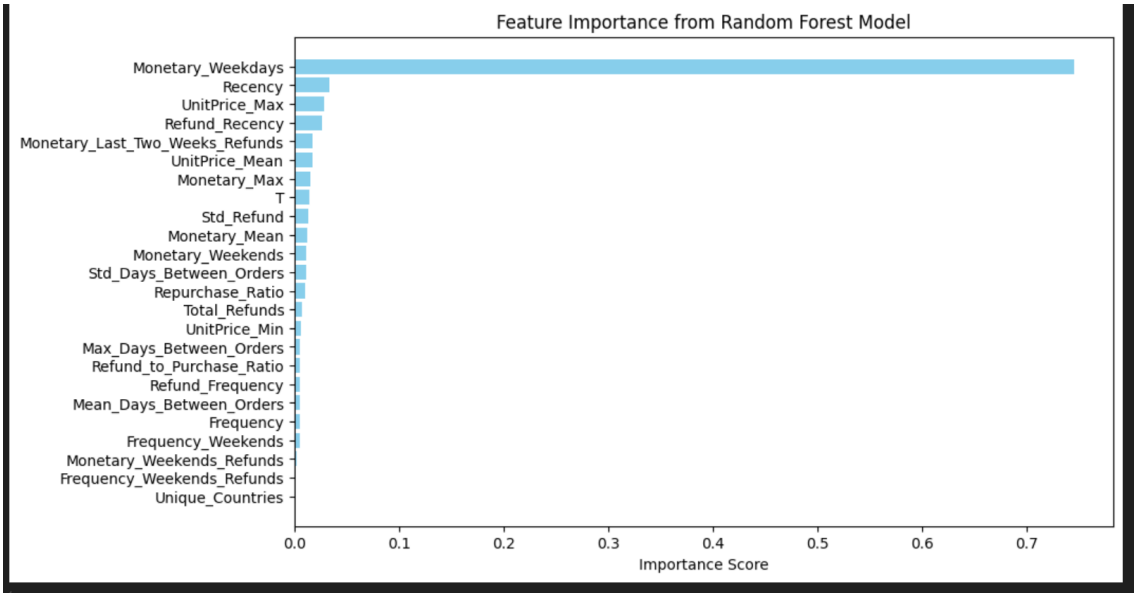

Figure 7: Feature Importance plot for RF with Refund Features

- Discount Behavior Metrics:
  Features including *Discount Frequency*, *Discount Monetary Value*, *Discount Ratio*, and *Discount Revenue Ratio* to understand customer response to discounts.
- Customer Purchase Metrics:
  Metrics such as *Total Price Sum*, *Total Quantity*, and the number of unique countries from which purchases were made.

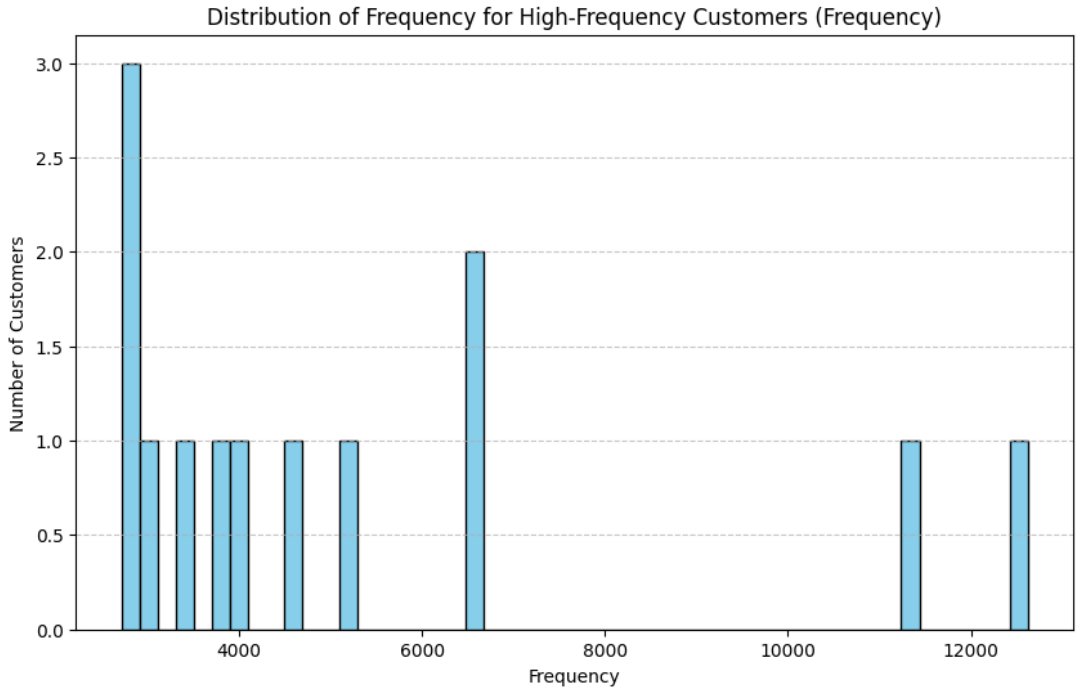

Figure 8: Outlier Distribution for Frequency

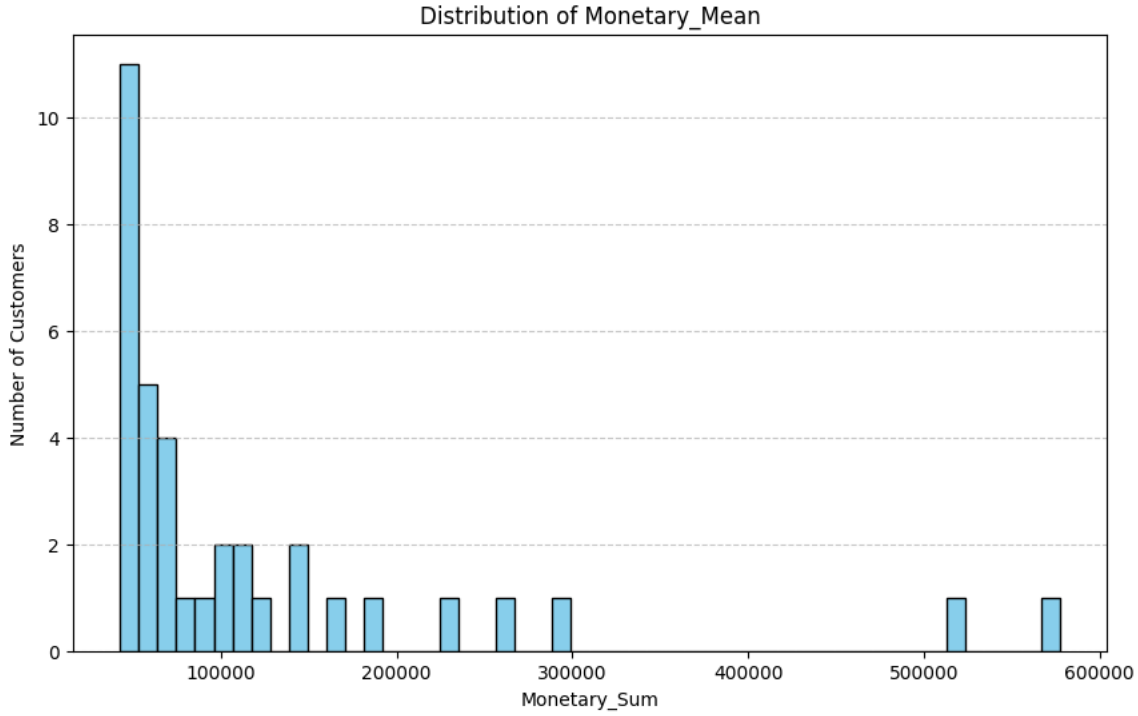

Figure 9: Outlier Distribution for Monetary

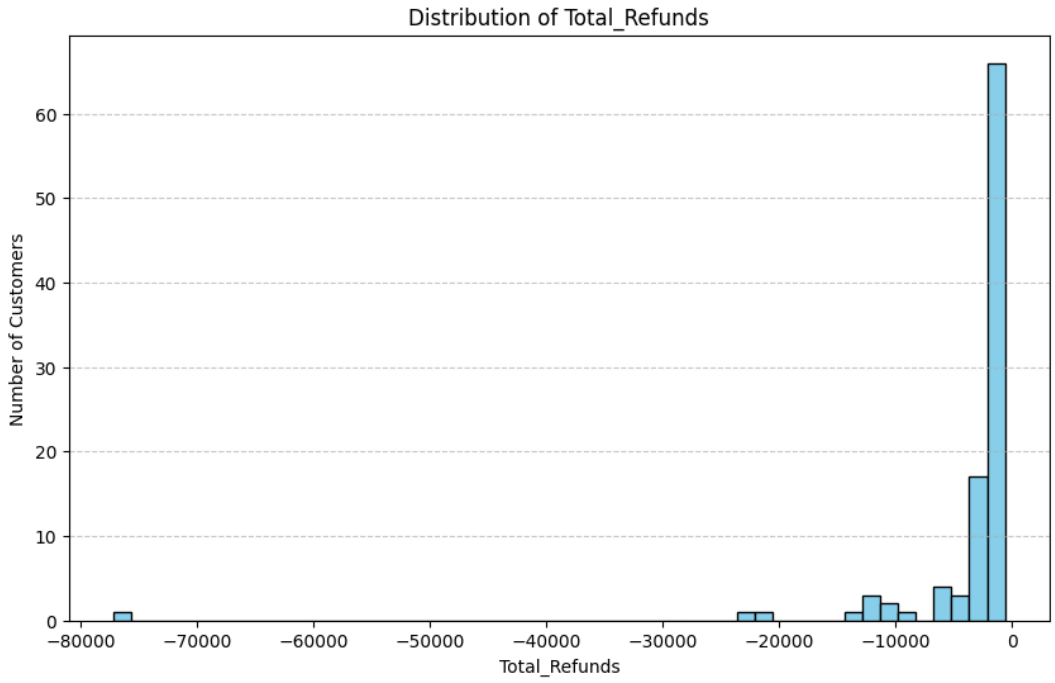

Figure 10: Outliers for Total Refund distribution

Upon analyzing these features, I identified extreme outliers and unusual patterns. For instance, one customer made purchases with quantities of 1 and revenues ranging from $3 to $8 but also had a single transaction worth $77,000 on January 18, 2011, which was immediately refunded (Figure 10 & 11). This significant refund skewed the data set, showing the importance of removing such anomalies.

	Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer ID	Country	Revenue
77939	499763	20682	RED SPOTTY CHILDS UMBRELLA	1	2010-03-02 13:08:00	3.25	12346.0	United Kingdom	3.25
77940	499763	20679	EDWARDIAN PARASOL RED	1	2010-03-02 13:08:00	5.95	12346.0	United Kingdom	5.95
77941	499763	15056N	EDWARDIAN PARASOL NATURAL	1	2010-03-02 13:08:00	5.95	12346.0	United Kingdom	5.95
77942	499763	15056BL	EDWARDIAN PARASOL BLACK	1	2010-03-02 13:08:00	5.95	12346.0	United Kingdom	5.95
77943	499763	15056P	EDWARDIAN PARASOL PINK	1	2010-03-02 13:08:00	5.95	12346.0	United Kingdom	5.95
193278	513774	21524	DOORMAT SPOTTY HOME SWEET HOME	1	2010-06-28 13:53:00	7.49	12346.0	United Kingdom	7.49
193279	513774	22692	DOORMAT WELCOME TO OUR HOME	1	2010-06-28 13:53:00	7.49	12346.0	United Kingdom	7.49
193280	513774	22660	DOORMAT I LOVE LONDON	1	2010-06-28 13:53:00	7.49	12346.0	United Kingdom	7.49
193281	513774	22687	DOORMAT CHRISTMAS VILLAGE	1	2010-06-28 13:53:00	7.49	12346.0	United Kingdom	7.49
193282	513774	48173C	DOORMAT BLACK FLOCK	1	2010-06-28 13:53:00	7.49	12346.0	United Kingdom	7.49
193283	513774	22691	DOORMAT WELCOME SUNRISE	1	2010-06-28 13:53:00	7.49	12346.0	United Kingdom	7.49
193284	513774	48111	DOORMAT 3 SMILEY CATS	1	2010-06-28 13:53:00	7.49	12346.0	United Kingdom	7.49
193285	513774	22690	DOORMAT HOME SWEET HOME BLUE	1	2010-06-28 13:53:00	7.49	12346.0	United Kingdom	7.49
193286	513774	21523	DOORMAT FANCY FONT HOME SWEET HOME	1	2010-06-28 13:53:00	7.49	12346.0	United Kingdom	7.49
193287	513774	48138	DOORMAT UNION FLAG	1	2010-06-28 13:53:00	7.49	12346.0	United Kingdom	7.49
193288	513774	22689	DOORMAT MERRY CHRISTMAS RED	1	2010-06-28 13:53:00	7.49	12346.0	United Kingdom	7.49
193289	513774	22365	DOORMAT RESPECTABLE HOUSE	1	2010-06-28 13:53:00	7.49	12346.0	United Kingdom	7.49
193290	513774	48185	DOORMAT FAIRY CAKE	1	2010-06-28 13:53:00	7.49	12346.0	United Kingdom	7.49
193291	513774	22688	DOORMAT PEACE ON EARTH BLUE	1	2010-06-28 13:53:00	7.49	12346.0	United Kingdom	7.49
193292	513774	48188	DOORMAT WELCOME PUPPIES	1	2010-06-28 13:53:00	7.49	12346.0	United Kingdom	7.49
193293	513774	48187	DOORMAT NEW ENGLAND	1	2010-06-28 13:53:00	7.49	12346.0	United Kingdom	7.49
193294	513774	22366	DOORMAT AIRMAIL	1	2010-06-28 13:53:00	7.49	12346.0	United Kingdom	7.49
193295	513774	20685	DOORMAT RED SPOT	1	2010-06-28 13:53:00	7.49	12346.0	United Kingdom	7.49
193296	513774	21955	DOORMAT UNION JACK GUNS AND ROSES	1	2010-06-28 13:53:00	7.49	12346.0	United Kingdom	7.49
288664	C525099	D	Discount	-1	2010-10-04 09:54:00	1.00	12346.0	United Kingdom	-1.00
431551	541431	23166	MEDIUM CERAMIC TOP STORAGE JAR	74215	2011-01-18 10:01:00	1.04	12346.0	United Kingdom	77183.60
431556	C541433	23166	MEDIUM CERAMIC TOP STORAGE JAR	-74215	2011-01-18 10:17:00	1.04	12346.0	United Kingdom	-77183.60

Figure 11: Unusual Purchase

Additionally, customers with very few purchases (1 to 4 transactions) or who only made purchases and then refunded them (netting zero purchases) were flagged for removal (Figure 12). Similarly, bulk purchases made on a single day with no other activity across three years were also deemed unrepresentative of general purchasing behavior and excluded.

```
data1[data1['Customer ID']==13829]
```

	Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer ID	Country	Revenue
307922	527119	22890	NOVELTY BISCUITS CAKE STAND 3 TIER	12	2010-10-14 18:19:00	8.5	13829.0	United Kingdom	102.0
415974	C539055	22890	NOVELTY BISCUITS CAKE STAND 3 TIER	-12	2010-12-15 16:36:00	8.5	13829.0	United Kingdom	-102.0

```
data1[data1['Customer ID']==15940]
```

	Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer ID	Country	Revenue
327883	C529176	35400	WOODEN BOX ADVENT CALENDAR	-4	2010-10-26 18:17:00	7.95	15940.0	United Kingdom	-31.8
443607	542915	35400	WOODEN BOX ADVENT CALENDAR	4	2011-02-01 16:18:00	8.95	15940.0	United Kingdom	35.8

+ Code + Markdown

```
data1[data1['Customer ID']==16073]
```

	Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer ID	Country	Revenue
452065	544078	22360	GLASS JAR ENGLISH CONFECTIONERY	48	2011-02-15 15:41:00	2.55	16073.0	United Kingdom	122.40
456080	C544558	22360	GLASS JAR ENGLISH CONFECTIONERY	-11	2011-02-21 12:33:00	2.55	16073.0	United Kingdom	-28.05

Figure 12: Few Purchases made throughout the dataset

After addressing these outliers, the refined dataset consisted of 793,647 records and 5,596 unique customers.

Correlation Matrix Analysis and Feature Reduction:

Next, I analyzed a correlation matrix for the 63 features to identify multicollinearity.

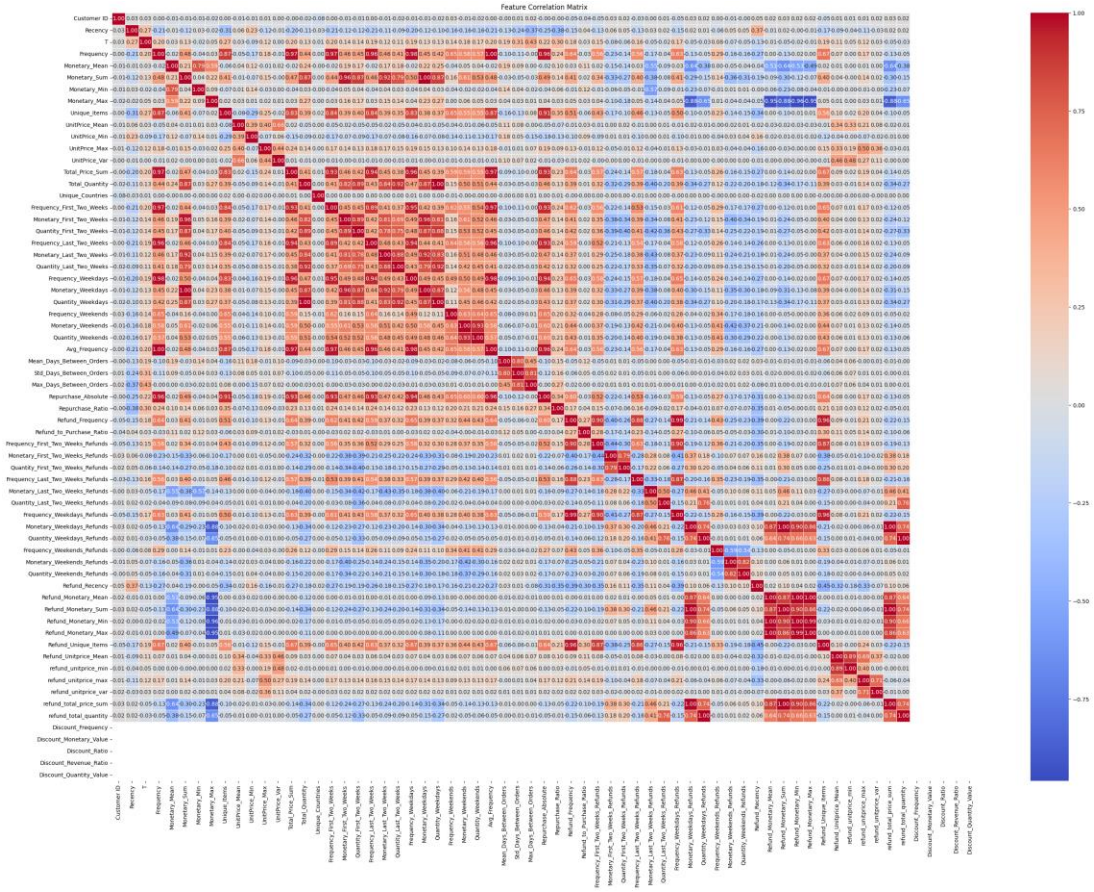

Figure 13: Correlation Matrix with multi-collinearity problem

Refund and discount features showed significant correlation with existing metrics, introducing redundancy. To address this, I removed several features, including:

- Refund-related features like *Monetary Weekends Refunds*, *Refund Monetary Max*, and *Frequency Weekdays Refunds*.
- General purchase metrics such as *Quantity Weekdays*, *Unique Items*, and *Total Price Sum*.
- Features with overlapping information, like *Frequency First Two Weeks Refunds* and *Monetary First Two Weeks Refunds*.

The features were reduced to 35 features for training.

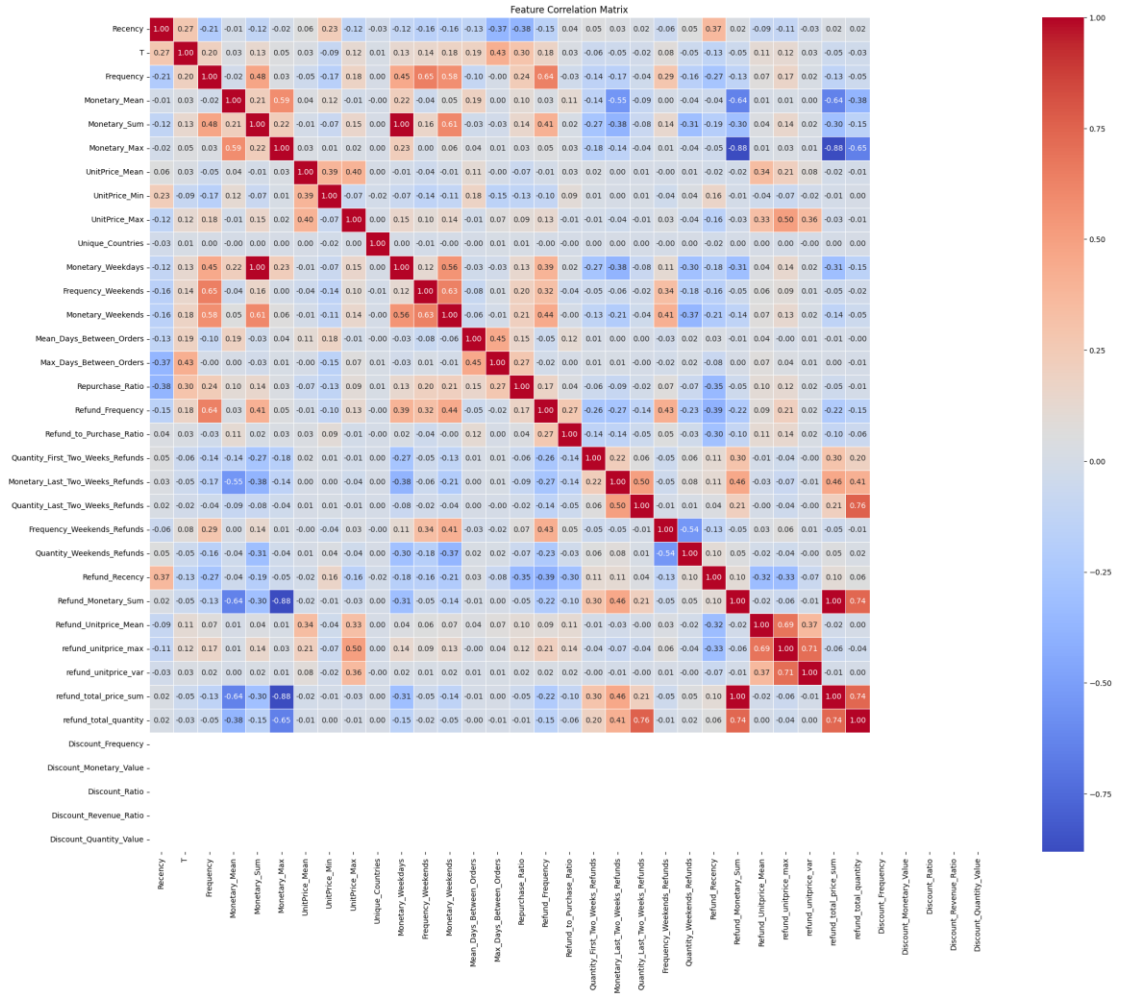

Figure 14: Correlation matrix after feature reduction

---

3. Progress and Milestones

Completed Tasks

Data cleaning was finalized by removing duplicates and irrelevant stock codes, such as POSTAGE, which were identified as non-representative of customer purchasing behavior. Feature engineering was a critical focus, resulting in 63 new metrics that captured nuanced aspects of refunds, discounts, and customer behavior, such as refund frequency, discount monetary value, and refund-to-purchase ratios.

A correlation matrix analysis was conducted to identify and address multicollinearity issues, leading to the removal of redundant or highly correlated features. Random Forest regression models were built and evaluated using the refined datasets, achieving improved R-squared values compared to previous attempts, indicating incremental improvements in model

performance. Additionally, PCA was explored to reduce dimensionality, and Neural Networks were tested for prediction.

Pending Tasks

First, the monetary variable needs to be categorized into high, medium, and low-value groups to explore classification modeling as an alternative approach. Next, a deeper investigation of outliers and unusual purchasing patterns is required to address data irregularities and ensure the dataset accurately represents customer behavior. Finally, integrating categorical variables such as product descriptions, refund details, and specific product information into future models will help capture more granular aspects of customer behavior and enhance prediction accuracy.

---

4. Problem-Solving and Challenges

Outliers posed a challenge to accurately modeling customer behavior, particularly cases with unusually high refunds or bulk purchases made on single days. These anomalies skewed the dataset and affected model predictions. While flagged for removal, further investigation is needed to understand their full impact on model performance.

Training the Neural Network model was difficult due to mismatched dataset shapes and insufficient learning. Aggregating customer data and assigning zero values for missing purchases ensured consistency, but the model still failed to converge effectively, indicating a need for further refinement.

Correlation among refund-related and monetary features introduced multicollinearity, requiring the removal of redundant features. While this improved the dataset's performance, dominant features like Monetary Weekdays continued to overshadow others, limiting overall model accuracy.

---

5. Technical Depth and Accuracy

The Random Forest model achieved its highest R-squared value of 0.528 after integrating refund and discount-related features, suggesting a slight improvement in predictive performance. However, PCA and Neural Networks did not perform as expected, indicating that further refinements in data preprocessing and feature selection are needed. The presence of outliers and the dominance of highly correlated features like *Monetary Weekdays* remain critical issues to address.

While modeling after extracting all features from the dataset, I encountered a training error indicating that input data contained values too large for a float variable or infinity. This issue likely arose due to the presence of extreme values in specific features. Further investigation is required to identify and resolve these problematic entries to ensure smooth model training.

Table1. Accuracy for each Model

Model	**R-squared Value**	**MAE**
Base RandomForest	0.519	-
Random Forest after Scaling data	0.516	-
Random Forest after using PCA	0.41	-
Neural Network	-	228.83
Random Forest after adding Refund Features	0.528	-

---

6. Future Plans and Goals

Future plans involve further investigating outliers and unusual purchasing patterns, focusing on removing customers with extreme or invalid behaviors that may skew the model. To understand the influence of other features, Monetary_Weekdays will be removed from the dataset in subsequent analyses. Additionally, the inclusion of categorical variables, such as product descriptions and refund-related details, will be explored to enhance the model's accuracy. Finally, Neural Network modeling and RandomForest will be revisited with refined datasets and additional preprocessing to address previous issues and improve predictive accuracy.