Customer Lifetime Value (CLV) Prediction and Analysis

Introduction:

Customer Lifetime Value (CLV) is a critical metric to figure out which customers are more important and needed to give more attention to. CLV analysis also helps to provide customized experience for different category of customers based on their purchasing habit. In this assignment, I have used Generalized Additive Models (GAM) to predict customer CLV based on the available historical data. Here, I have also evaluated the model's performance based on LIME explanation to figure out if the model is performing optimally or not.

Data Preparation:

Data Overview: This Online Retail II data set contains all the transactions occurring for a UK-based and registered, non-store online retail between 01/12/2009 and 09/12/2011.

The dataset contains InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, Country information in separate columns.

Feature Engineering: I have derived the following features from the dataset to predict the CLV.

Recency: The number of days since the customer's last purchase.
Frequency: The total number of transactions made by the customer.
Monetary Value: The total value of purchases made by the customer.

Data Cleaning:

- Negative quantities were removed to ensure data integrity.
- Monetary values were clipped at the 95th percentile to remove extreme outliers that could skew the model's predictions.

Methodology:

CLV Prediction Using GAM: I used GAM to predict CLV in different parts. First, I have predicted the customers monetary value and customer churn and then used these two prediction to calculate the CLV.

Monetary Value Prediction: A GAM was trained to predict the total monetary value a customer generates based on their Recency and Frequency.

Churn Prediction: A separate GAM was used to estimate the probability that a customer would churn based on Recency and Frequency.

CLV Calculation: The CLV was computed by multiplying the predicted monetary value by the churn probability.

Explain Model With LIME: To understand how the model is calculating the predictions, I have applied LIME to explain hoe recency and frequency is contributing to the predicted CLV for a customer. This will enable me to check if the models decision making process is in alignment with business intuition and will help to improve transparency of the predictions.

CLV Prediction Analysis:

Lets check the below examples for analysis.

CustomerID	Recency	Frequency	MonetaryValue	PredictedMonetaryValue	ChurnProbability	CLV
12346	325	34	9530.08	725.8561185	1.007735324	731.4708505
12347	2	253	5633.32	4416.597226	0.994470576	4392.175987
12348	75	51	2019.4	1337.361026	1.0082631	1348.411774
12349	18	175	4428.69	3041.448325	1.004079961	3053.857316
12350	310	17	334.4	459.8581977	1.020396906	469.2378822
12351	375	21	300.93	315.1033965	1.024089918	322.6942116
12352	36	103	2849.84	1883.969095	0.997037855	1878.388505

Let's take customer with CustomerID 12346, this customer has high recency indicating that the customer didn't purchase for a long time. Even though the customer spent a significant amount of money previously, the predicted monetary value is low due to high churn probability. This indicates that the business should take measures to re-engage with this customer immediately.

Now if we take a look at customer with CustomerID 12347, this customer has made many purchases recently (low recency) and has a high frequency of transactions, which indicates strong engagement. Their predicted CLV is one of the highest, which suggests they are a high-value customer. The business should prioritize retaining this customer by offering personalized discounts, loyalty rewards, or exclusive deals to maintain their loyalty.

Data Plotting:

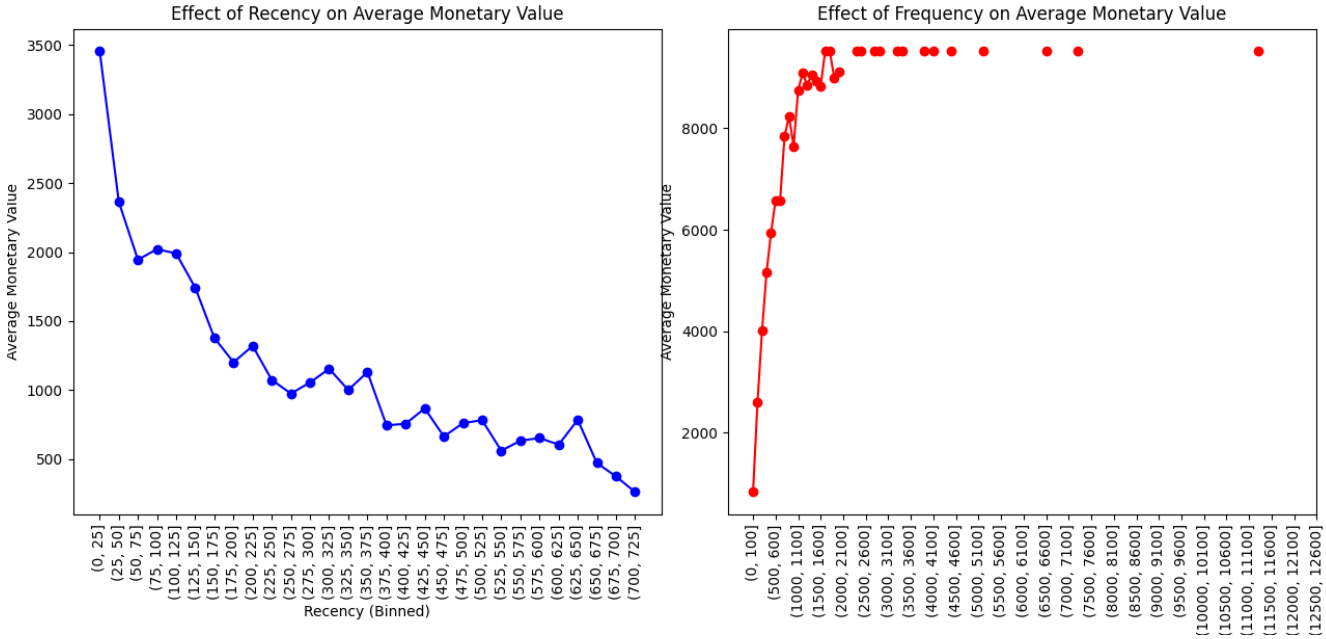

Fig 1: Effect of recency and frequency on average monetary value

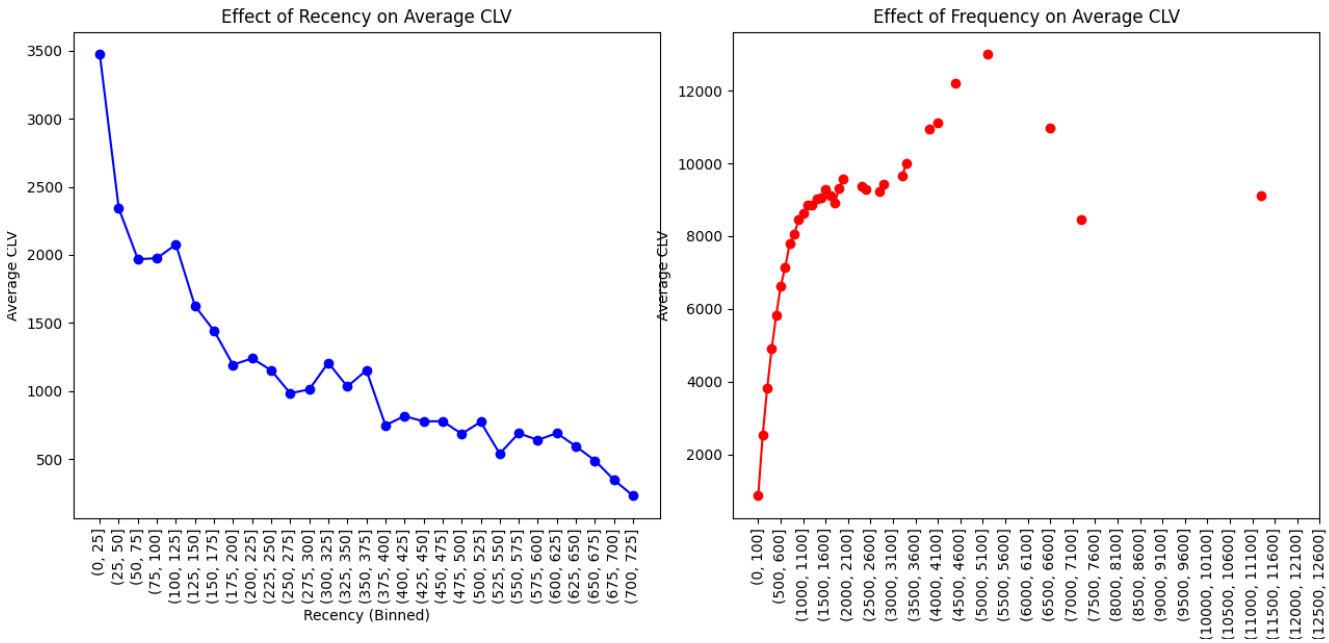

Fig 2: Effect of recency and frequency on CLV

```
LIME Explanation for Predicted Monetary Value:
21.00 < Frequency <= 53.00: -2011.229269135749
95.00 < Recency <= 379.00: -209.52614482068572

LIME Explanation for Predicted CLV:
21.00 < Frequency <= 53.00: -2094.291448775186
95.00 < Recency <= 379.00: -211.27260395397232
```

Fig 3: LIME Explanation

LIME Explanations:

The LIME explanations were generated for both Monetary Value and CLV predictions. Below is a summary of the LIME explanations.

Monetary Value Prediction:

- The feature 21.00 < Frequency <= 53.00 has a significant negative impact of -2011.23, which means that for customers whose frequency is within this range, the predicted monetary value is reduced considerably.
- Similarly, 95.00 < Recency <= 379.00 also has a negative impact of -209.53 on the predicted monetary value, indicating that higher recency (customers purchasing less recently) reduces the predicted value.

CLV Prediction:

- The same features (21.00 < Frequency <= 53.00 and 95.00 < Recency <= 379.00) also have negative effects on CLV prediction, suggesting that when frequency is moderate, and recency is high, the CLV is predicted to decrease.

Interpreting Results:

- Both LIME explanations show that frequency and recency have significant impacts on the model's predictions, which aligns with expectations for customer lifetime value (CLV) models—higher frequency should generally correlate with higher CLV, and higher recency (i.e., the longer since the last purchase) should negatively impact CLV.
- However, the large negative impact of 21.00 < Frequency <= 53.00 suggests that the model is contributing negatively to the predictions of moderate-frequency customers. This suggests that the model may be underestimating the value of customers who make frequent but lower-value purchases.
- Similarly, the impact of 95.00 < Recency <= 379.00 on CLV suggests that the model expects customers who haven't purchased recently to have lower CLV, which is reasonable as high recency is expected to have some impact on CLV.

Conclusion:

In this assignment I have demonstrated the use of Generalized Additive Models (GAM) for predicting Customer Lifetime Value (CLV) using transactional data. I have analyzed the performance of the model through LIME explanations and visualizations. While the model performed generally well, adjustments to feature engineering and model tuning could further enhance the accuracy of CLV predictions, particularly for high-frequency customers.

References:

https://archive.ics.uci.edu/dataset/502/online+retail+ii
https://www.shopify.com/ca/blog/what-is-customer-lifetime-value
https://pygam.readthedocs.io/en/latest/notebooks/quick_start.html#Fit-a-Model
https://lifelines.readthedocs.io/en/latest/Survival%20Regression.html
https://chatgpt.com/
https://stackoverflow.com/questions/64019645/plotting-binned-correlation-of-two-variables-using-common-axis
https://www.geeksforgeeks.org/binning-data-in-python-with-scipy-numpy/
https://www.w3schools.com/python/matplotlib_subplot.asp
https://marcotcr.github.io/lime/tutorials/Tutorial%20-%20continuous%20and%20categorical%20features.html