Title:

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

Full Name: Shashwat Dhayade

Report Submission Date: 10/17/2024

1. Introduction

During the past two weeks, significant progress was made on data understanding and preprocessing. The dataset, which comprises invoices, quantities, prices, and customer information, was thoroughly examined to remove inconsistencies, clean the data, and prepare it for modeling. The overall goal is to predict the CLTV using this refined dataset. This report details the steps taken, challenges encountered, and plans for the next phase.

---

2. Summary of Work Done

During this reporting period, the primary focus was on data cleaning, a critical step in preparing the dataset for analysis. Initially, 34,335 duplicate records were identified, and overlaps in the data from 2010-2012 were resolved. This resulted in a refined dataset with 11,676 duplicates remaining.

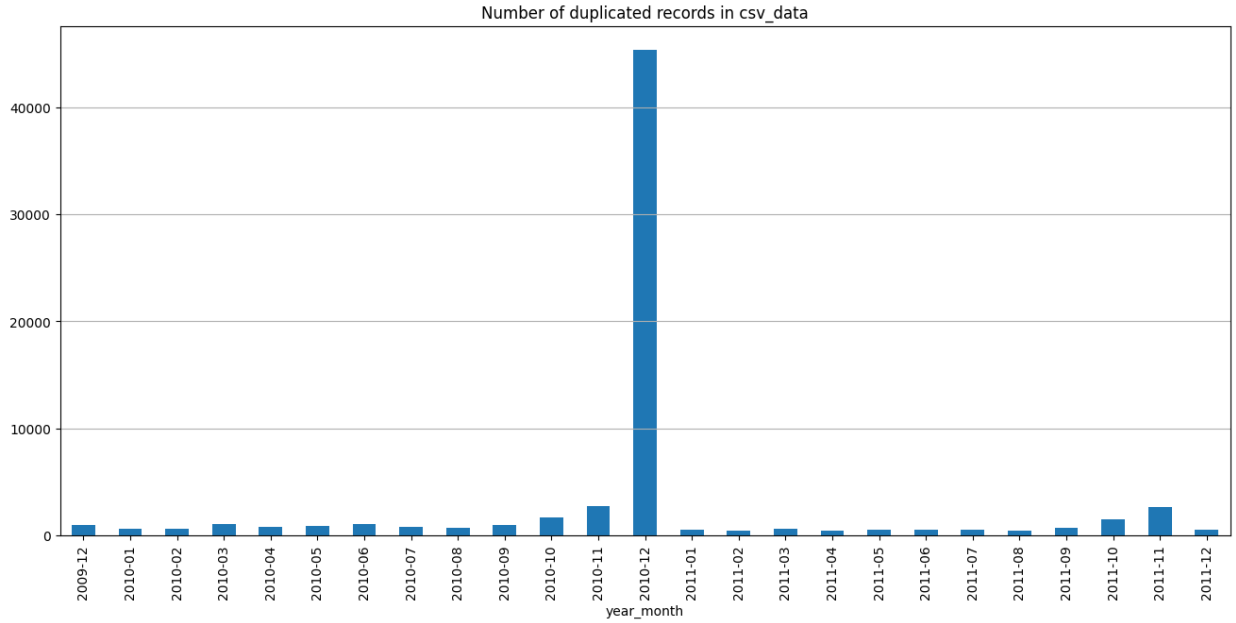

Figure 1. Number of duplicate records in retail data.

Additionally, we addressed missing values, including 235,287 instances of missing Customer IDs and 4,275 missing product descriptions. Customers with no positive purchase history, including 128 customers who only had refund transactions, were excluded from the dataset as they did not contribute to understanding customer purchasing behavior.

Each column attribute was carefully analyzed for consistency and accuracy. Notably, invoices containing refunds were identified by the presence of a “C” at the start of the invoice number. Multiple purchases made at the same time and date shared the same invoice number, while refunds had distinct invoice numbers. The Quantity attribute showed negative values for returned items and positive values for completed purchases. InvoiceDate included both the date and time of purchase, while Price represented the price per unit of a product. All Customer ID values were positive and unique. The dataset spanned transactions from 43 countries, with the vast majority originating from the UK.

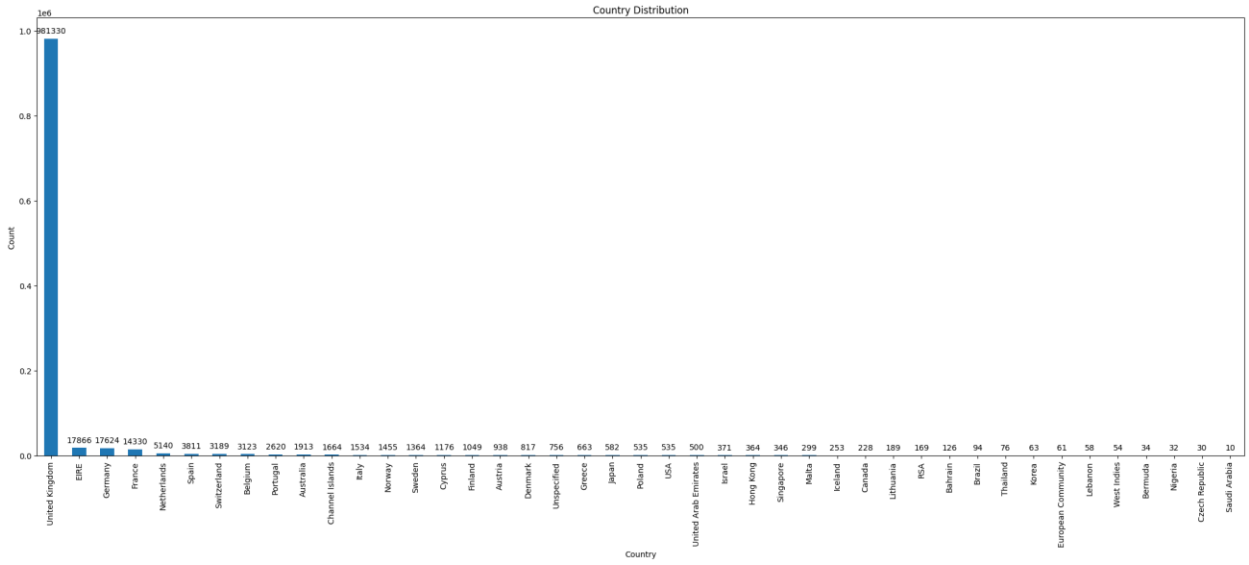

Figure 2. Distribution of Country and the number of purchases

Further analysis revealed 5,296 unique product descriptions and 4,643 unique stock codes. Upon review, several product descriptions were deemed unrelated to customer purchases and were removed from the dataset. These included:

- **Discounts (170 instances)**: Most of these were refunds, though five were incorrectly recorded as purchases, which were subsequently removed.
- **Manual (1,066 instances)**: Likely representing labor costs, these were excluded as they do not reflect customer purchasing patterns.

- **Bank Charges (35 instances)**, **Test Products (16 instances)**, **Adjustments (55 instances)**, and **CRUK Commission** (fees paid by Customer 14096 to the Cancer Research UK organization) were similarly removed as they did not represent genuine product purchases.

Some product descriptions such as **POSTAGE**, **DOTCOM POSTAGE**, and **Next Day Carriage** were retained for now and will be evaluated in the model to determine whether they should remain in the final dataset or be removed based on their impact on the model's predictive performance.

Most of the irrelevant stock codes corresponded to these non-purchase-related descriptions, with the exception of the **C2** stock code, which consistently had a price of $50. This stock code likely represents either a delivery charge or a product, and its role will be further examined during the model-building phase.

Following the cleaning process, the dataset was reduced to 808,240 records, with 11,665 remaining duplicate entries. These duplicates will also be evaluated during model testing to assess their influence on prediction accuracy. This cleaned and structured dataset now forms the foundation for the next phase of the project, which involves feature engineering and model development.

---

3. Progress and Milestones

Key milestones achieved in this period include the completion of the data cleaning and preprocessing phases. The analysis and correction of special entries like test products, manual adjustments, and discounts have provided a clearer and more manageable dataset. However, some tasks remain ongoing, including feature engineering, where we will derive important customer-level variables such as total purchase value, recency, and frequency. The next step will be to build a base model based on the Recency-Frequency-Monetary (RFM) strategy, which will help in the prediction of CLV. This model will serve as a benchmark for further iterations and improvements.

---

4. Problem-Solving and Challenges

Several challenges arose during this phase of the project. One of the main issues was identifying meaningful variables from a complex dataset. For example, many stock codes, such as manual adjustments or administrative entries, did not directly relate to customer purchasing behavior and had to be addressed. This required detailed assumptions about the relevance of each variable, such as removing codes related to comissions, bank charges, or test products.

Solutions were implemented to clean and streamline the data, such as removing irrelevant codes and adjusting for positive quantities that represented refunds. These actions have resulted in a cleaner dataset that can more effectively contribute to model accuracy.

---

5. Technical Depth and Accuracy

From a technical perspective, the most significant work revolved around ensuring the accuracy and relevance of the data. The cleaning process involved eliminating redundant and erroneous data, which posed the risk of introducing noise into the predictive model. By applying logical filters and assumptions, such as excluding stock codes that represented administrative fees or charges, we now have a dataset that is not only cleaner but also more suitable for predictive modeling. This technical groundwork is essential for building a robust CLV prediction model using machine learning techniques. The dataset of 808,240 instances is now well-prepared, setting the stage for the next steps in model development.

---

6. Future Plans and Goals

Looking ahead, the next two weeks will focus on feature engineering and model building. The aim is to derive key customer-level features that will be fed into the predictive model. Following this, a base model using the Recency-Frequency-Monetary (RFM) strategy will be developed to provide an initial prediction of Customer Lifetime Value. The subsequent plan involves experimenting with more advanced machine learning models, such as decision trees and gradient boosting, to enhance prediction accuracy. Additionally, we will evaluate the model's performance using standard metrics.