Project Documentation: Market Segmentation and Spending Prediction Using Neural Networks

Introduction

The objective of this project was to predict total spending in an online retail setting and derive insights from the dataset to understand customer spending behavior better.

Dataset Description

The dataset used is the Online Retail Data, which includes features such as Description, Quantity, Invoice Date, UnitPrice, CustomerID, and Country. The dataset comprises 541,908 records.

Data Preparation and Exploration

1. Data Cleaning:

   Removed all NaN values to ensure data integrity.

   Changed the data type of specific columns (Quantity and Invoice Date) for better analysis. The Invoice Date was further broken down into year, month, and day components.

   Converted Country values from categorical to dummy variables (0 and 1) for easier processing.

2. Exploratory Data Analysis (EDA):
   - Conducted EDA by plotting graphs between specific columns, leading to several key findings:
     - The United Kingdom has the highest total quantities.
     - Singapore has the highest average unit price.
     - November has the highest quantity sold, indicating peak season.
     - No significant correlation between Quantity and UnitPrice with the year.
     - "Item World War 2 Gliders ASSTD Designs" emerged as the best seller.
     - A negative correlation was observed between spending increases and the number of customers.

Feature Engineering

Developed new features to enhance model predictions and insights:

1. Country-based Features: Created features to reflect country-specific trends, especially for the UK and Singapore.
2. Time-based Features: Introduced binary features for November to capture seasonal trends.
3. Product Popularity: Identified the best-selling product for potential targeting.
4. Spending vs. Customer Behavior: Captured the relationship between increased spending and customer count.

5. Ratios and Relative Features: Introduced ratios to explore the relationship between spending and customer count more deeply.

Model Development

1. Neural Network Processing:
    - Prepared the dataset for modeling by defining independent variables (X) and the target variable (y), followed by splitting into training and testing sets.
    - Utilized TensorFlow for building the model, employing a Sequential model with two hidden layers utilizing Sigmoid and Linear functions.
    - Scaled the data to improve model efficiency.
    - Compiled and trained the model, observing the performance on training and testing sets.

Model Evaluation

The model was evaluated based on its ability to predict total spending accurately, with specific attention to the performance metrics during training and validation phases. Predictions (preds) were generated to assess the model's effectiveness.

Market Segmentation

1. Spending Level Segments: Utilized quantiles of TotalSpending to categorize transactions into high, medium, and low spending segments.
2. Geographic Segments: Analyzed spending trends by geographic location, focusing on the UK and Singapore, along with other countries.

Recommendations

The analysis revealed that average spending was at a medium level, with spending in Singapore almost twice that of the UK. A recommendation is to explore strategies for decreasing average spending in Singapore to possibly increase the customer base or transaction volume.