Product Data Challenge - Thanksgiving sales trends

Introduction to the Problem

The Product Data Challenge presented several analysis questions that require comprehensive product data. The sales data contains complete item names, and the Google Product Taxonomy data maps onto the sales data. However, approximately 68% of the sales data still lacks a category after mapping. The challenge is to find a data science approach to classify these items into categories effectively.

Matching Items to Categories

To determine which category each item belongs to, especially in cases with multiple assigned categories, we used a fuzzy matching process with the `fuzzywuzzy` Python library. This approach worked well to select a definitive category per item.

Framing the Classification Problem

We approached this problem as a hierarchical classification task, using item names to predict product categories. The target variables are hierarchical, such as "Apparel > Loose > Bottoms > Jeans." We trained a local classifier per node in the target variable hierarchy using the `hiclass` library in Python.

Exploratory Data Analysis

We conducted exploratory data analysis to understand the dataset better. For example, the "PageOne Lighting Skylight LED Round Flushmount" had a sale amount of $33.5k and a commission of $6,380. It could be a legitimate transaction, but we assumed that multiple items might take the form of multiple transactions, therefore it is likely anomalous and should be removed.

Feature Engineering

We decided to use only the item name as a feature since other sales information might not be informative. We subsetted the sales data to use completed categories (labels) for training and testing. Text processing involved removing special characters, numbers, punctuation, tokenizing the text, removing stop words, and rejoining words into a cleaned sentence.

Target Variable

The target variables initially looked like "Apparel > Loose > Bottoms > Jeans" and needed to be processed into a list form, like ['Apparel', 'Loose', 'Bottoms', 'Jeans'].

Vectorizing Text Data

Text classification models require converting text data into numerical form. We used a CountVectorizer to achieve this, converting the text data into an array/matrix of token counts.

TF-IDF

We applied Term-frequency times inverse document-frequency (TF-IDF) post the CountVectorizer. TF-IDF works well in text classification, determining token relevance based on frequency and document occurrence.

Model Training

We performed a train/test split of 80/20% and trained the hierarchical classification model using the `hiclass` approach, with a base classifier (RandomForestClassifier).

Model Evaluation

The model was evaluated using adjusted classification metrics suitable for hierarchical problems, including F1-score (85.7%), Precision (79.4%), and Recall (93.2%).

Model Inference

Predictions were generated for all items in the original sales data, forming the basis for answering the posed product data questions.

Questions and Findings

Q1: Thanksgiving Sales Trends

* Transactions: Black Friday (26) > Cyber Monday (29) > Thanksgiving (25).
* Country: US had the largest sales, followed by CA and GB.
* Thanksgiving saw cheaper items ($56) and larger basket sizes (~3), while Cyber Monday had smaller basket sizes (~2).

Q2: Top Selling Products

* Overall: "Oversized Vintage Tunic Sweatshirt for Women" was the top seller.
* By Product Category: "Apparel & Accessories > Clothing > Pants > Jeans" was the top seller.

Q3: Highest Commissioned Products

* With Cards: "The Business Platinum Card® from American Express" had the highest commission.
* Without Cards: "Theragun Elite Black - Smart Percussive Therapy Massager" had the highest commission.

Q4: Frequently Purchased Together

Frequently purchased item pairs included "High-Waisted Pixie Ankle Pants for Women" and "Oversized Vintage Tunic Sweatshirt for Women."

Q5: Product Categories Frequently Purchased Together

"Apparel & Accessories > Clothing > Pants > Jeans" and "Apparel & Accessories > Clothing > Shirts & Tops > T-Shirts" were frequently purchased together.

Q6: Sales Trends Comparison (Black Friday vs. Cyber Monday)

* Overall, "Theragun" was the top seller in terms of revenue for both days.
* Units sold varied between Black Friday (leggings and T-shirts) and Cyber Monday (backpacks and jeans).
* Revenue sources differed, with more electronics sold on Cyber Monday.

Q7: Business Recommendations

* Consider encouraging purchases of items adjacent to top sellers (add-ons), especially in the fitness product category.
* Investigate psychological perceptions of spending on Black Friday vs. Cyber Monday.
* Adjust pricing strategies for gaming consoles based on Cyber Monday trends.
* Explore strategies to spread spending over different sales days.

Deployment Solution

For model deployment, we envision two potential solutions:

1. Classification Model Deployment: Deploy the category classification model to provide clients with predicted product categories for their items.
2. Recommendations Deployment: Develop an optimizer that assists companies in allocating items to sales days and provides recommendations for discounts. This solution would leverage the categories predicted by the classification model to optimize sales strategies.

Ideas for Further Improvement

* Remove outliers and credit card items for better data quality.
* Explore advanced NLP techniques, such as pre-trained word embeddings.
* Use deep learning frameworks like Keras, TensorFlow, or PyTorch for improved model performance and faster training.
* Consider incorporating additional features to enhance item name-based classification.
* Perform model interpretation and feature importance analysis (e.g., SHAP).
* Research how other companies have solved similar problems for further insights.