Video script Machine Learning Project (Final Stage)

Alaric de Bastard, Adrien de Mailly Nesle, Mathieu Cowan

Hello everyone, today we will introduce our Machine Learning study on online retail.

Our primary objective is to predict the revenue and profit from transactions, enabling better stock management and optimized pricing strategies. By analyzing sales trends, especially during peak periods like Christmas, we aim to understand customer behaviors and improve decision-making.

We worked with the UCI Online Retail dataset, which includes transaction details such as product descriptions, prices, quantities, and customer purchase over a year.

Our preprocessing focused on data cleaning, we firstly remove discounts, restock and refund to be more easily analyzed. We created features like revenue per transaction and remove outliers and missing values.

For modeling (**Stage 1**), we implemented regression techniques like Linear Regression and Random Forest, comparing their performance using metrics like Mean Squared Error. Random Forest showed the best results, with features like time and customer IDs proving most influential.

We also use embedding function to transform product descriptions

Embedding refers to a method where high-dimensional data, such as text, picture or video is transformed into lower-dimensional vector representations more useful. The goal of embeddings is to capture the semantics or relationships between data points.

In **the second approach of stage 1**, we transformed textual data into numerical embeddings using advanced sentence transformers to capture the semantic meaning of the descriptions. These embeddings were used as inputs for several regression models, including Linear Regression, Random Forest, Support Vector Regression, and K-Nearest Neighbors. Among these, Linear Regression delivered the most reliable predictions, indicating a modest relationship between product descriptions and revenue. Despite challenges like overfitting and embedding limitations, this approach highlighted the potential of leveraging textual data for revenue predictions.

In **Stage 2**, we analyzed customer behavior using RFM segmentation. RFM stands for Recency, Frequency, and Monetary value, which measures how recently a customer made a purchase, how often they purchase, and how much they spend. By applying clustering algorithms such as K-Means and Agglomerative Clustering, we identified meaningful customer groups. These included 'Champions'—our most loyal and high-value customers—and 'At-Risk Customers,' who might need targeted re-engagement. This study offers actionable insights for creating focused marketing strategies and improving customer retention.

In the **second part of stage 2**, we used the SentenceTransformer embedding to predict the cumulative revenue of the products of UK from November 26th to december 26th . We used notably random forest regressor and Gradient boosting Regressor. We had slightly better results, but the overfitting and data limitations were still issues

In **Stage 3**, we focused on Germany's 2010 transactions. We used the MLP Regressor, a neural network model, optimizing hyperparameters to capture complex relationships between the embedding of the description and the revenue. This resulted in better control over overfitting and more reliable predictions compared to Stage 2.

Additionally, we classified revenues into high and low categories by two different ways, first in a manual way with the quantiles (median), which resulted in pretty good results for most of the classifying models including MLP Classifier for this part. Then we used DBSCAN, we see here the desequilibred distribution between the categories leading to biased classifications.

Overall, throughout many ways and many insights, from stage 1 to stage 3, we successfully made sure we had a significant evolution of our understanding of client behavior and revenue dynamics, offering valuable insights for strategic decision-making.