Stage 1 - Project Machine Learning

Mathieu Cowan, Adrien de Mailly Nesle, Alaric De Bastard DIA 3

Business scope :

The principal goal here is to predict the price of an order and its benefice. The purpose of this whole study is to make sure that we can have a better understanding of the behavior of the profits/sales and so that afterwards we will be able to target optimization of benefice for our business. We will then be able to manage better our stocks and adapt ourselves in terms of pricing business and to identify correctly the needs of the customers.

Problem : We want to predict the income of the online retail with the different features of our dataset.

Algorithm used: in this study, multiple regression models will be used in order to do such predictions. It is adapted for modeling and quantifying the relationship between the explanatory variables and the target (revenue).

Limitations : The relationship between the features of our dataset can be quite complex, so we need to make the dataset more usable for such models. It is possible that there can be non-linear patterns with our features, which can complexify the task

Explanation of the data :

Our dataset contains data of transactions during a calendar year from an online retail store with different features :

- InvoiceNo: a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation

- Stock: a 5-digit integral number uniquely assigned to each distinct product

- Description: product name

- Quantity: the quantities of each product (item) per transaction

- InvoiceDate: the day and time when each transaction was generated

- UnitPrice: product price per unit

- CustomerID: 5-digit integral number uniquely assigned to customer

- Country: the name of the country where each customer resides

Data Pre-processing :

To facilitate the analysis of the data, we will put multiple analysis/decisions in place to facilitate our work on the dataset:

Work on the features :

- **Feature Treatment:**
  - We created columns for Day, Hour, Week, and Month by extracting them from the InvoiceDate column so we can more easily make a deep temporal analysis of the data.
  - The UnitPrice column was cast to float so that we can use it properly in a numerical way.
  - A new column Total was created to calculate the total value of each transaction, obtained by multiplying Quantity by UnitPrice.
- **Data Cleaning:**
  - Rows with missing product descriptions were dropped.
  - When a transaction has its feature InvoiceNo which starts by "C", it means that it is canceled. Therefore, such rows were dropped.
  - Transactions which didn't have a strictly positive unitary price or a price exceeding 1000 (extreme outlier) were taken out from our dataset to ensure data quality.
  - When a transaction has a negative quantity, this transaction needs to be dropped because they certainly meant errors or returns.
  - When a product only had one transaction during the whole calendar year where the study is made, the row is removed because it then does not contribute significantly to the sales of the online retail.
  - To make sure we were accurate in our analysis, we took out duplicate rows.
  - To optimize the text processing and analysis we took out the commas and replaced them by spaces
- **Feature Creation:**
  a. To make sure that we could identify missing customers, all missing CustomerD values was replaced by -1.
  b. To facilitate the analysis, a new feature BaseStockCode was created from the StockCode feature where we take out the letters so that we can create a more general identifier for the stock groups.
  c. We put a new feature CommonDescription which allows us to assemble similar products so that eventually it will facilitate the categorization of the data.

General Methodology

To gain insights and prepare the dataset for modeling, the following methodology was followed:

1. Data Exploration and Feature Understanding

Through a pushed analysis of our dataset, we identify the main features of the trends of the sales weekly and monthly. We notably see that during the holiday season (especially Christmas and New Year) the tendencies are that the sales will be higher, sometimes three times higher than other weeks. We see that with the visualisation of our data; therefore, we understand better the behavior of customers.

2. Handling Missing Values

We observed missing values with the CustomerID, we handled it by assigning -1 to null values to better identify unknown customers and assuring consistency for our analysis. Missing product descriptions forced us to drop rows where such a case was met because it results to not be able to identify products.

3. Imbalanced Data

We saw first that most of the products are sold in the UK, which can create unbalance in some analysis. Furthermore, some items were heavily purchased, and others were purchased only once during the whole year, so we took the decision to delete of the dataset the cases where a product was purchased only once to focus on meaningful transactions for the whole dataset.

4. Outliers

Outliers were identified in the UnitPrice and Quantity columns. Transactions with unit prices exceeding 1000 or negative values were removed, as they represented either input errors or unrealistic values. Similarly, transactions with negative quantities (likely returns) were discarded, leaving only valid purchase data.

Visual Analysis:

To understand transaction trends:

- Weekly Transactions: We saw with the plot of transactions per week that there are a lot of variation of tendencies during the whole year depending on the season, the period (example, pic when it is the Christmas/New Year period)
- Weekly and Monthly Sales Value: We could identify better the spending habits of the clients with similar tendencies by making a plot of the profit per week (*quantity × unit price)*
- Monthly Transactions: We did a similar study for the number of transactions and similar tendencies were shown.
- Customer Analysis: we analyzed the number of transactions of customers who have more than 20 transactions so that we can see the distributions between the customers of all of the transactions during the calendar year.

Utilisation of models

Approach 1: Revenue Prediction from Metadata Features

Steps:

- Data Cleaning and Feature Engineering:
  - Timestamps were converted into numerical data (e.g., day, hour, week, month) for easier manipulation.
  - We used SimpleImputer to impute missing values with means
  - The target variable was the total value of an order, calculated as the product of Quantity and UnitPrice.
- Model Implementation:
  - Linear Regression was used as the baseline model.
  - Random Forest Regressor was further optimized using GridSearchCV to test various hyperparameters.
- Metrics for Evaluation:
  - We used Mean Squared Error (MSE) and Mean Absolute Error (MAE)

Results:

- Linear Regression:
  - MSE: 1650.32
  - MAE: 35.67
- Optimized Random Forest:
  - MSE: 1480.76
  - MAE : 30.12

Analysis of Results:

- The optimized Random Forest model outperformed Linear Regression, achieving lower MSE and MAE values.
- Feature Importance: Random Forest will identify the features which contribute the most to the predictions. Temporal features (exemple : Week) and customer-related attributes (example : e.g., CustomerID) will likely play a more significant role in the prediction of the revenue

Approach 2: Revenue Prediction from Product Descriptions

Steps:

1. Data Preparation:
    - We embedded the products descriptions into vector representation using SentenceTransformer, which transforms the textual data into numerical embeddings that capture semantic meaning.
    - Split of the embeddings : 80% train, 20% test using train_test_split.
    - Features were normalized using StandardScaler so that all variables have a mean of 0 and a standard deviation of 1.
2. Model Selection and Implementation: Four regression models were selected:
    - Linear Regression: We use this model to assess the relationship between features and the target variable.
    - Random Forest Regressor: An ensemble model that builds multiple decision trees for hopefully a better performance.
    - Support Vector Regression (SVR): We use a kernel which is linear to predict the target variable.
    - K-Nearest Neighbors (KNN): A non-parametric model that predicts based on the closest neighbors in the feature space.

Each model was trained on the scaled embeddings and evaluated on the test set.

3. Metrics for Evaluation:
    - Mean Squared Error (MSE)
    - Mean Absolute Error (MAE)

Analysis of Results:

- Linear Regression performed best on training and test datasets with one of the lowest test MSE (1405.49). This implies that the relationship between embeddings and revenue is relatively modest
- Random Forest had comparable score the MSE_test, maybe it is due to the number of estimators.
- SVR and KNN the test MSEs were high, especially for SVR. This suggests these models are less suitable for this dataset or require further analysis on it.
- Overfitting: The overfitting was significant because of the notable difference between test and train for MSE.

Conclusion and Insights

We did a developed analysis of our dataset with embedding-based model which had a modest results. The linear regression and the Random Forest Regressor were the best-performing algorithms overall although the prediction is yet to be optimized. The overfitting was still notable, in the future we can work on more advanced ways of using regression, of using embeddings, and to categorize the products in a more optimized and advanced way.