Machine Learning Project – Online Retail – Stage 2

Mathieu Cowan, Adrien de Mailly Nesle, Alaric de Bastard

GENERAL INTRODUCTION

On the last step, we pre-processed our data and we analysed it, visualized it. We attempted to predict the revenue from metadata features on a first approach and then on and then using the embedding of the description of the product. Our main goal is to be able to predict the patterns on the client to be able to make the best business decisions for our online retail. We will now make a further approach on this matter by on a first part predict the cumulative revenue on a certain period for the products in the United Kingdom, and then by implementing a process to be able to categorize the clients based on their patterns/purchases

Part 1 - Predict the cumulative revenue on a certain period for the products in the United Kingdom

1. Previous methods and limitations

On the precedent model, we embedded the product description into vector representation using SentenceTransformer which transforms the textual data into numerical embeddings, and then we did a 80-20 split to then train Linear regression, Random Forest Regressor, SVR and K-NeighborsRegressor. However, we had limitation on this data because it considered each transaction. However, there are many transactions during the year for the same product. Also, we can assume that for such an analysis, we may need to focus this task on the analysis of the customer patterns during a high-sales period. We will see further on the new task we did to make a further analysis on that matter.

2. Improvement Assumptions

Now we will use a different way of attempting to solve the **problem of prediction of the revenue with the description of the products**. Indeed, here we had an idea that was good to perform the embedding on an important country in our dataset such as Germany. However, in this specific case the hyperparameters are not optimized enough, and the embedding is not optimized either since it was done on each transaction and not on each product.

In addition, it would be more judicious and more precise to analyze the trends of a country that represents the majority of our dataset, namely the United Kingdom and to take a specific period which corresponds to a lot of sales as we have seen in previous steps, and the particularity of this period makes the analysis of the customer patterns quite interesting.

Also, we will use a parameters grid which will optimize our search for the best hyperparameters for each model, which would mean that we should find better prediction.

3. Problem Formalization and method

3.1. New Algorithm description

Basically, when it comes to the data, we will filter it on the United Kingdom country, take only rows from 26/11 to 26/12. Also, we will to a group by on the method and then aggregate the revenue to make sure we will have the cumulated sales of each product. We will therefore take a period that seems to us the most suitable to understand customer behavior, namely the period from November 26 to December 26.
We will consider the description of the product and the total revenue of this product in the said period and country. We will then do an embedding on the description and perform in a better way multiple models including advanced ones such as GradientBoostingRegressor and RandomForestRegressor.

3.2. Methodology - New Algorithm implementation

For the embedding, we will use a parameters grid to make sure to optimize the number of hyperparameter combinations for the models we will use, and we will put all our results in a csv we will print so that we can make an analysis out of it.
We also use a function that for each model generates different hyperparameters from the grid so that all combinations are evaluated with the model evaluation function that will calculate the training time, inference time, rmse train and test, MSE train and test.
This happens for each model/hyperparameters combination and then at the end we have the CSV that we print to make our analysis and interpretations of the results.

We have the following models with their hyperparameters :
Linear Regression : We use this model to assess the relationship between features and the target variable.

- Fit intercept : permit an adaptation of the model in case that the data is centered or not

KNeighborRegressor : model based on the closest neighbours. It uses the nearest observation in the space and calculates the mean of those neighbours.

- N_neighbours : number of neighbours taken into account to make the prediction
- Weights: mesures the impact of the neighbours
- P : distance used to mesure the proximity

Random Forest Regressor : An ensemble model that builds multiple decision trees for hopefully a more precise prediction

- N_estimators : number of trees
- Max_depth : maximum depth of the trees
- Min_sample_split : minimum number necessary observations so that we regulate the nods.

Gradient boosing regressor : ensemble model based on the boosting. It constructs the treese and each tree corrects the errors of the precedent one.

- N_estimators : number of trees
- Learning_rate : rate of learning of the trees
- Max_depth : maximum depth of the trees

Limitations

Obviously, not all the products will be considered because some products are not sold at all during these periods. Also, we are aware that it could be considered as a risk to use only the embedding of the description of the product to predict the revenue, however we believe that we need to explore all the possibilities concerning embeddings using the dataset considering the limitations of our dataset.

4. Discussion on our results

Metrics Used :

- **Training time**: Time required to train a model on a training dataset.
- **Inference time**: The time taken by the model to make predictions on new data after training.
- **RMSE (Root Mean Squared Error) train and test**: The root mean squared error, calculated on the training and test sets respectively, which measures the difference between predicted and actual values.
- **MAE (Mean Absolute Error) train and test**: The mean absolute error, calculated on the training and test sets respectively, which measures the average of the absolute differences between predicted and actual values.

Interpretation of the results :

A. Linear Regression

Whether fit_intercept is true or not the results are pretty much the same, we have a relatively modest overfitting (1481-926) for the RMSE and the difference between MAE tests and MAE test isn't that big (685-485)

B. KNeighborsRegression

We notice that there is a big gap between performances with the uniform weights and with the distance weights.

Indeed, they all have pretty similar RMSE test values (same for MAE). However the gap between the train and test value varies a lot. Indeed, we see that the train values for the RMSE is certainly high for the 'uniform' weights but they are still inferior than the one of the RMSE test and the difference between the two is relatively modest. However this is still better than having a model where the gap is way higher, which would implicate a huge overfitting.

Regarding that, we can consider that the best model here is the model 12, even if we have the following values:

- RMSE train : 1040.63
- RMSE test : 1327.13
- MAE train : 244.52
- MAE test : 331.6

C. Random Forest Regressor

For the Random Forest Regressor, the train time and the inference time is very long, however it resulted to very high overfitting each time - lowest gap between RMSE test and train : 2025 - 556.

D. Gradient Boosting Regressor

Finally for Gradient Boosting Regressor, a new Advanced model we used, the training time wasn't that long compared to random forest regressor. for this model the overfitting was very high (example,RMSE test of 405 and RMSE train of 2047).

Generally, in this case the MAE test is twice as big as the MAE of test. Which wouldn't make this model adapted for our problem unfortunately.

CONCLUSION – PART 1

With all of this, we realize that Linear Regression and and KNeighbour regression with the right parameters, even if they are not the most optimal models, they are the most adapted model with the goal to predict the cumulated revenue from the embedding of the description of the product on a particular year.

However, we realized that with the difference of the object of the embedding, the random forest regressor did not perform as well. We also saw that the values of the RMSE diminued for the Linear Regression compared to the precedent step. Even if there are still errors, we have noted the true difference when we are more precise on what is embedded for the model and when we adapt better what we apply the models to.

We could have tested even more combinaisons of hyperparameters, but the problem is the more we do that, the more complex the algorithm is and then the combinated training time of the algorithm is way too long.

Furthermore, the predictions are just a little better than the previous ones, we found a little bit less overfitting for certain models and more for others. We could have had better prediction on another period for example.

Part 2 – Categorization of the clients based on their patterns

In the previous step we used random forests to try and predict the total value of an order based of multiple spatial, temporal and other parameters. This approach did not succeed very well as those parameters don't seem to correlate with the goal we were aiming for. In this step we will try to label our clients to better know their behavior.

1. Problem

The goal of this step is to categorize customers based on their purchasing behavior to create targeted marketing strategies. This is achieved using an **RFM analysis**, which is based on the following metrics:

- **Recency (R)**: How recently a customer made a purchase.
- **Frequency (F)**: How often a customer makes purchases.
- **Monetary (M)**: The total monetary value of purchases made by the customer.

RFM analysis is a proven approach for identifying homogenous customer groups that share similar purchase behaviors.

2. Formalization and Methods

RFM analysis is a widely used method for segmenting customers based on transactional behavior. It addresses the following objectives:

- Identify **loyal customers** or "champions."
- Detect **lost customers** or customers at risk of churning.
- Recognize customers who need marketing intervention.
- Categorize new or high-value customers.

To segment customers into distinct groups, clustering algorithms are employed. The methodology involves:

1. **Data Preparation**: Cleaning and preprocessing the raw dataset.
2. **Feature Engineering**: Calculating RFM metrics.
3. **Data Normalization**: Standardizing data to prepare it for clustering algorithms.
4. **Clustering**: Using algorithms like **K-Means** and **Agglomerative Clustering**.
5. **Evaluation of Results**: Validating clusters using metrics like silhouette scores and visualizations.

3. Algorithm Description

K-Means Clustering

- Partitions the data into k clusters based on distances to centroids.
- **Advantages**: Efficient and scalable for large datasets.

- **Limitations**: Sensitive to centroid initialization and requires predefining the number of clusters (k).

Agglomerative Clustering

- Hierarchical clustering method that groups data points based on inter-cluster distances.
- **Advantages**: Provides a dendrogram to visualize hierarchical relationships.
- **Limitations**: Computationally intensive for large datasets.

These algorithms are suitable for identifying natural groupings within the data without prior labels.

4. Limitations

Several limitations of the approach are noted:

1. **Scaling Sensitivity**: Algorithms like K-Means require normalization to prevent biases due to differences in feature scales.
2. **Optimal Number of Clusters (k)**: Selecting the optimal number of clusters can be subjective and impact the results.
3. **Cluster Overlap**: Customer groups may overlap, leading to ambiguity in segment definitions.
4. **Data Bias**: The quality and representation of the initial dataset strongly influence the outcomes.

5. Methodology

1. **Data Import and Cleaning**:
    a. Removed invalid transactions (e.g., negative prices or quantities).
    b. Filtered out customers with missing IDs and eliminated duplicates.
2. **RFM Metrics Calculation**:
    a. **Recency**: Calculated as the number of days since the last purchase.
    b. **Frequency**: The total number of transactions for each customer.
    c. **Monetary**: The total spending by each customer.
3. **Data Normalization**:
    a. Applied **StandardScaler** to ensure equal weighting of RFM metrics during distance-based calculations.
4. **Clustering**:
    a. **K-Means** clustering with the optimal number of clusters determined using the elbow method and silhouette scores.
    b. **Agglomerative Clustering** to observe hierarchical relationships between customer groups.
5. **Validation**:
    a. Visualized clusters in reduced dimensions (e.g., PCA).
    b. Evaluated cluster quality using metrics like silhouette scores.

6. Algorithm Implementation and Hyperparameters

K-Means:

- **Hyperparameter**: `n_clusters` (number of clusters).
- **Elbow Method**:
    - The inertia plot indicates a notable decrease at 4 or 5 clusters, suggesting a good balance between simplicity and accuracy.
- **Silhouette Scores**:
    - Achieved maximum scores at 2 clusters (~0.43) but acceptable results for 4 clusters (~0.35).

Agglomerative Clustering:

- **Hyperparameter**: Linkage method (e.g., average, complete).
- A dendrogram suggests natural groupings around 4 clusters.

7. Results

Cluster Visualization:

- Customers were segmented into distinct groups such as:
    - **Champions**: High-value, frequent, and recent buyers.
    - **At-Risk Customers**: Customers who have reduced purchasing frequency.
    - **Lost Customers**: Customers who haven't made a purchase for a long time.
    - **Recent Customers**: Newly acquired customers with recent transactions.

Customer Distribution:

- Histograms revealed that "Lost Customers" and "Champions" represent significant portions of the customer base, highlighting key segments to target or re-engage.

8. Metrics

The following metrics were used to evaluate clustering quality:

1. **Silhouette Score**:
    a. Measures cluster separation and compactness.
    b. Observed scores: ~0.43 for 2 clusters, ~0.35 for 4 clusters.
2. **Inertia**:
    a. Used in the elbow method to find the optimal number of clusters.
3. **Cluster Visualization**:
    a. Reduced-dimensional scatter plots confirm cluster separation and grouping quality.

9. Overfitting

- **Risk**: While clustering is unsupervised, overfitting may occur if too many clusters are created, leading to non-meaningful groupings.
- **Mitigation**:
    - Validated clusters using multiple algorithms (K-Means and Agglomerative).
    - Examined cluster centroids to ensure interpretability and meaningful segmentation.

CONCLUSION – PART 2

This part effectively applied RFM analysis to segment customers into actionable groups. By leveraging clustering algorithms like K-Means and Agglomerative Clustering, we identified distinct customer behaviors that can inform targeted marketing campaigns.

Future Improvements:

1. Incorporate additional features such as embedding of Invoice for richer segmentation.
2. Automate the selection of the optimal number of clusters using advanced methods like the Gap Statistic.

GENERAL CONCLUSION

This concludes Stage 2 with the further analysis of the behavior of the client through two very important aspects of our online retail : the impact of the revenue which we attempted to predict – even if the prediction could have been better, we have seen better prediction with relatively modest results than the previous step and we have done interesting analysis on the differences between the performances of multiple models. We also applied RFM analysis to segment customers and to use clustering to make a better identification of the patterns of the clients. For the next steps, we could typically use Neural Networks which could very well be more adapted for time series, we could also use the suggestions at the end of each step.