RETURN ANALYSIS :

Conclusions & points of attention

EDA (see file EDA.ipynb) - basics

- Data are comprised of 12 numerical data but 9 of them are in fact categorical
    - too be taken into account for ML
- and 1.75M lines covering a 2 month period
- no null values
- out of that duplicates represents a bit less than 50% of the total lines
    - further duplicate analysis show that it is difficult to exclude them without knowing more on the compant logging process
    - there is a possibilities that those duplicates (same shoes, same day same shop bought several times) are real sales (especially online) :
        - numerous customers
        - double size buying

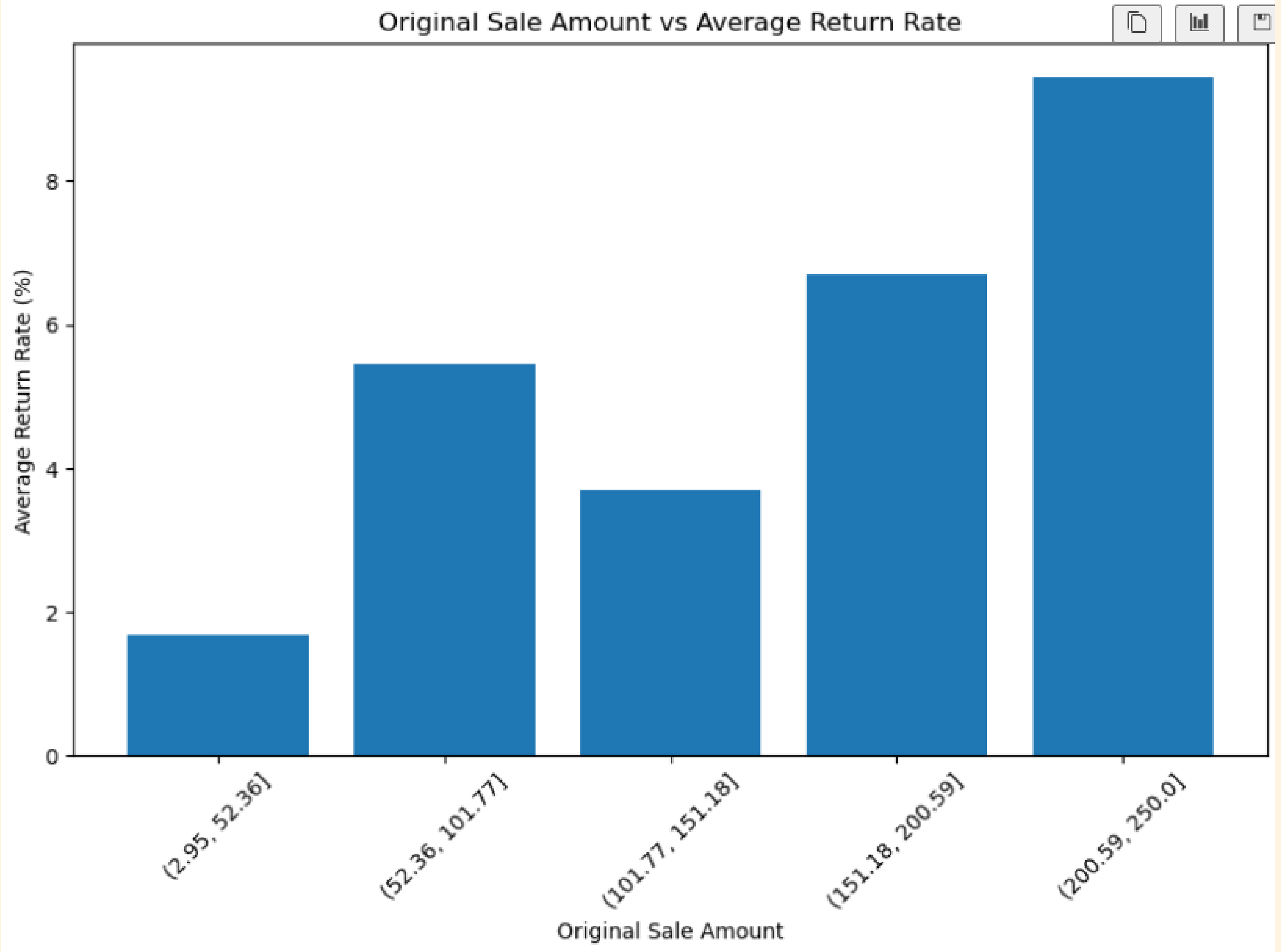
EDA (see file EDA.ipynb) - Business

- 100 MEUR in sales
- Return amount to 3.8% of the sales
- Average discount : 18%
- Average gross margin after discount and cost of good : 48%

EDA (see file EDA.ipynb) - Business: Price

- **Average return rate is increasing with price as can be seen on the below graph.**

Stronger communication or constraints strategy may be interesting to study for high price shoes to lower return rates.



EDA (see file EDA.ipynb) - Business: brands

- 649 brands
  - 20 of them have a return rate above 25%. Some do have return rate that are very high
  - maybe indicating quality issue? design issue ?

deeper analysis by brand could help to determined a strategy about underperforming brand regarding return (quality audit, stopped etc etc)

EDA (see file EDA.ipynb) - Business: brands

- 649 brands
  - 20 of them have a return rate above 25%. Some do have return rate that are very high
  - maybe indicating quality issue? design issue ?

deeper analysis by brand could help to determined a strategy about underperforming brand regarding return (quality audit, stopped etc etc)

EDA (see file EDA.ipynb) - Business: Products

- 25K products
- Product with return rate > 20% represents :
  - 10% of all the products
  - 16% of all return
  - for 2% of the sales

There is probably deeper analysis to do by product to determined which one could be stopped

EDA (see file EDA.ipynb) - Business: Shop

- The 2 onlines business :
    - represents 10% of total sales but 40% of the return
    - indeed the return rate is around 19% vs 4% average
- moreover :
    - their discount rate is higher by 7 points and thus their profitability

This needs a deeper analysis to set up a strategy that may correct those imbalanced element bith for return rate and profitability

EDA (see file EDA.ipynb) - Business: Clients

- In the top 20 clients (B2B?): 2 of them display a much higher return rate than other
- More than 100 clients have a return rate above 33% and total sales above 1KEUR

This kind of analysis could help to identofy clients that may abuse the system and elaborate a strategy to counter that

PREPROCESSING: main elements

- **Quantitative addition (new column) as new potential impactful information for ML Model :**
  - Discount Rate column
  - Profit Percentage column
  - Day of the Week
  - Number of Items per transaction
  - Number of Identical Items per Transaction
  - See preprocessing.ipynb file

- **Technical preprocessing to ensure higher usability by the ML Model :**
  - One hot encoding for Day of the week
  - Target encoding for all numerical columns that were in fact categoriel.
    - Replacing their numerical if by mean of the target for their category
  - Normalization with StandardScaler
  - See model.py file

Model : main elements

- **Target and performance criteria :**
  - **Sensitivity vs Precision**
  - it is important to identify a maximum number of positive case (high Sensitivity/recall)
  - but it is important to as well to ensure that we the model do not create to many false positive (high precison)
  - because any return prevention message towards a false positive may have negative impact
  - thus F1 that measure the best score taking into account the 2 parameters is our main evaluation criteria

- **Model choice : XGboost as best performance**
  - tested 3 ML model (XGboos, Random Forest , Decision tree)
  - 2 neural network model (Keria and) Sensitivity vs Precision
  - GridSearch and manual tuning, + reprocessing done to optimized paremeters
  - performance logging through MLflow (available on demand)
  -

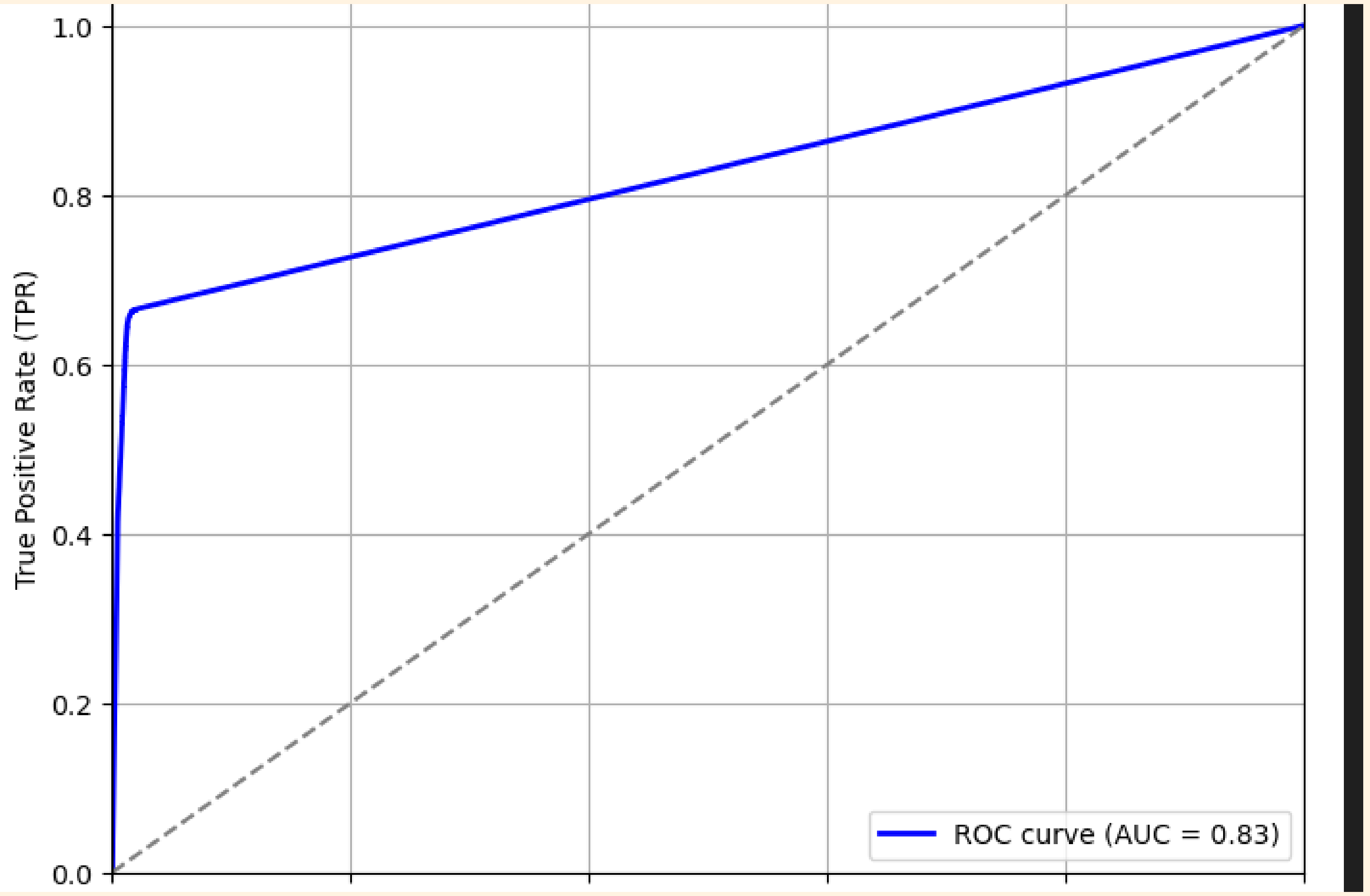
Model : Performance Metrics

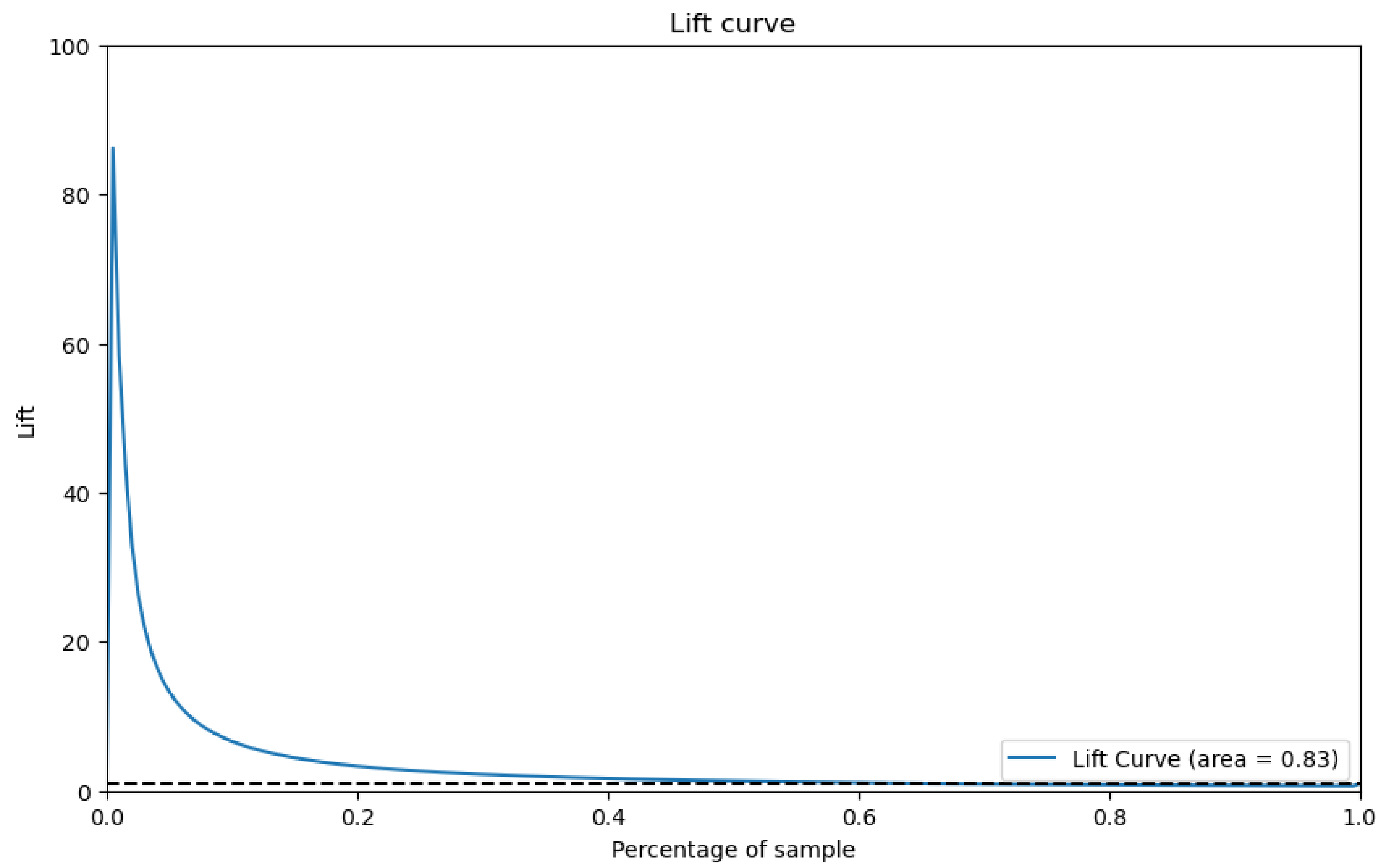
```
[[333102   5425]
 [  4556   8895]]
              precision    recall  f1-score   support

           0       0.99      0.98      0.99    338527
           1       0.62      0.66      0.64     13451

    accuracy                           0.97    351978
   macro avg       0.80      0.82      0.81    351978
weighted avg       0.97      0.97      0.97    351978

0.8226319069174524
```





Deployment:

Deployed through fast API locally : http://localhost:8000/docs#/

Repo : https://github.com/slvg01/8_Returns_management.git