# Test from UZUM Market! 🛒

## Task

Product returns pose a serious problem for online marketplaces. Statistics show that approximately 15% to 40% of all online purchases are returned, which not only creates financial difficulties for businesses, but also poses serious environmental problems due to the huge amount of goods ending up in landfills. To reduce the number of returns, it is critical for sellers to be able to identify the reasons for returns as early as possible. However, this information often becomes available only after a significant number of product returns. Understanding the reasons for product returns and predicting them can go a long way in reducing this problem.

## Target

Develop a machine learning model that will solve a multi-class classification problem and predict the probability distribution of reasons for returning a product based on many factors, including textual customer reviews.

### Input format

The link to [google drive](https://drive.google.com/drive/folders/1c9ABGWtH5xgJFIPSANEJusIxTMuwIuFD?usp=sharing) contains 5 files:

`return\_reasons.parquet` - a dictionary file with unique reasons for returning a product. Each reason has an id and description. Total unique 5 -

[DEFECTED, WRONG\_ITEM, BAD\_QUALITY, PHOTO\_MISMATCH, WRONG\_SIZE]

`reviews.parquet` - a file with reviews of purchases on the marketplace.

Each review has:

- order\_item\_id - unique order identifier

- product\_id - unique product identifier

- customer\_id - unique customer identifier

- review\_text - client’s text review about the product

- shop\_id - unique store identifier

- rating - rating on a scale from 1 to 5

- date\_created - timestamp of review creation

`returns.parquet` - file with returns of goods on the marketplace.

- id - unique identifier for returning goods

- product\_id - unique product identifier

- cause - one of 5 reasons for return

- comment - response to return

- date\_created - timestamp of return processing

- order\_item\_id - unique identifier of the product order

- customer\_id - unique customer identifier

- purchase\_price - product order price

products.parquet - file with product descriptions in Russian and Uzbek languages

- product\_id - unique product identifier

- category\_id \*\*\*\*- id of the product category

- category\_title \*\*\*\*- title of the product category

- product\_description \*\*\*\* - product description in Russian and Uzbek languages

\*\*\*\*

`test.parquet` is a file containing the same as `returns.parquet`, except the reason for the return. This is exactly what you have to predict.

❗️❗️❗️ All files are in \*parquet\* format, which is the format we expect to receive your output file in. You can read more and get acquainted with \*parquet\* [here](https://pandas.pydata.org/pandas-docs/version/1.1/reference/api/pandas.DataFrame.to\_parquet.html)

### Metrics

To evaluate the classification model model, we chose Macro Average Precision:

- Average Precision is a metric that takes into account the precision for each class separately and then averages them. It is suitable for cases where both precision and recall for each class are important.

- Macro Average Precision - calculates precision for each class separately and then averages them, treating all classes equally regardless of their frequency or imbalance in the data set.

### Solution requirements

❗️❗️❗️ The solution should be downloaded from [link](https://forms.gle/H3efhqv177Fyw1tp6) as a zip archive, which contains in the root:

- \*\*solution\*\* folder with all \*.py and \*.ipynb files

- file \*\*result.parquet\*\*

The \*\*result.parquet\*\* file must contain two columns:- product\_id \*\*\*\*- product identifier \*\*\*\*from \*\*test.parquet\*\*

- order\_item\_id - order identifier for product \*\*\*\* from \*\*test.parquet\*\*

- prob\_return\_reason\_DEFECTED - probability of return due to \*\*\*\*DEFECTED

- prob\_return\_reason\_WRONG\_ITEM - probability of return due to WRONG\_ITEM

- `prob\_return\_reason\_\*\*BAD\_QUALITY` - probability of return due to reason `BAD\_QUALITY`

- `prob\_return\_reason\_PHOTO\_MISMATCH` - probability of return due to reason `PHOTO\_MISMATCH`

- `prob\_return\_reason\_WRONG\_SIZE` - probability of return due to reason \*\*\*\*`WRONG\_SIZE`

❗️❗️❗️\*\*The order of columns and reasons in them is strictly as follows. Changing it will affect the final metric!

The solution should reflect the research and all the code needed to reproduce the solution. We expect code to be clean, formatted, and consistent. It will be cool if the root of the solution is a small README.md

Strict zip archive name format: \*\*LastName\_FirstName.zip\*\* (example \*\*Ivanov\_Ivan.zip\*\*)

### Criteria for evaluation

1. To evaluate the final solution, Macro Average Precision will be used, calculated on data from \*\*test.parquet\*\*

2. Not only the results of the model will be evaluated, but also the quality of the code, its cleanliness and design.

### Note:

- You can use any text processing methods and machine learning models to achieve your goal.

- Try to make maximum use of information from text reviews to improve the quality of predictions.

Good luck!