Quantitative Marketing / Marketing Econometrics

TSE – M2 EA/ES – 2025-2026 – Anna BLANPIED

Project n°2

Predict who is going to repurchase on the next month!

Context

You have recently joined the **Data Science Department** of a major supermarket chain. After only one week on the job, you are assigned to assist the **Marketing Department** in preparing for its upcoming promotional email campaign.

The marketing team aims to identify a **selection of customers (no more than 10% of the customer base)** who are the most likely to make a purchase again in the near future.
As an initial step, the **Director of Data Science** wants to test your skills by assigning you the task of building a performant propensity score.

Your mission is to:

1. **Predict** whether a customer will make a purchase in the next month (Kaggle submission).
2. **Recommend** how the marketing department should leverage these predictions to optimize the campaign's targeting strategy (oral presentation).

Data:

id_client	Identifier for a unique customer
transaction_date	Date and time of purchase. Be careful, there can be multiple rows for one id_client and one transaction_date
stores_nb	Identifier for the store where the purchase was made; the first 2 digits look like a regional code… but do the last 3 digits look like this is forming a postal code?
item_count	number of items purchased
gross_amount	total value of purchases
discount_amount	total value of discounts for the purchase
basket_value	value paid by the customer
payment_gift	has purchase been paid (partly) by gifts to the customer
payment_cheque	has purchase been paid (partly) by cheque by the customer
payment_cash	has purchase been paid (partly) by cash by the customer
payment_card	has purchase been paid (partly) by card by the customer
email_domain	email domain if given by the customer
civility	Mister, Miss, or not given
zip_code	INSEE code of the customer
card_subscription	has the customer subscribed to a loyalty card
multicard	has the customer subscribed to loyalty programs from other brands (unrelated to payment cards)
price_segmentation	segmentation of the type of products purchased: 8: products with different prices but within a same range of products, Access: affordable products, Mixte: sometimes affordable and at other times high-end, Quali: high-end, xx: products too different within a same purchase

Expected work:

Your project will be evaluated on 2 criteria:

- Your final ranking in this competition (based on AUC). You can make 20 submissions per day, and your final ranking (private leaderboard) will be based on your best submission. Try to do better than my submission (basic model)! The weight of this ranking in the project 2 grade will account for 30% of the grade. So don't spend too much time trying to be the 1st on the leaderboard, just because you want to beat the 2nd group even if there's a 0.0001 difference! It is about learning, not about competition. And because the oral presentation weighs more than the ranking.

- Your oral presentation: you will present your work in a group, with some time for questions. Do not give too much details but focus on your best submission and present the data preparation/cleaning you have made, the changes you made on the features, the features included in your model, the type of algorithms used, the different version you have compared, what could further improve your model… Explain how your model can be used for targeting the next marketing campaign, and do not hesitate to criticize your work, the features at your disposal, what you did but failed to give a boost in performance! The overall presentation will account for 70% of the grade.

As an example (not mandatory at all), your coding work can follow the following framework:

1- Context & objectives: Why is this problem important? What is the business goal of this propensity model? Provide preliminary details on why this scoring approach is valuable and how it could be applied in practice.
2- Data Overview : Present key statistics about the dataset (similar to Project 1). Include time period, number of observations, variables, and simple descriptive statistics. Do not overlook this step, basic exploration may reveal important insights about the scope and quality of the data.
3- Feature: Explain the new features you created, transformations applied, and your rationale behind them. Show relevant descriptive statistics and visualizations to highlight how certain features relate to the likelihood of repurchase.
4- Modeling Approach & Insights : Briefly describe the models tested, their main strengths, and your performance metrics. Present the best-performing model. Identify the most influential features, describe the profile of a likely repurchaser, and discuss model interpretability.
5- Recommendations : Suggest how the marketing and CRM teams could leverage your model. Provide actionable ideas on campaign targeting, personalization, and overall brand strategy.

Note: this is a mix of a technical and business presentation.

- This should not be “okay, I’ve got 103040 rows, I’ve got these variables, I implemented feature engineering, I removed missing values, here are 4 models we tried, here is a complete explanation of how XGBoost works, and we got this score, and voilà”. You’re not blindly presenting technical details; you are answering a business question. Everything you do can enhance the marketing efforts of the brand, so what you present should be justified / criticized / etc.
- This is not only a business presentation: I want to know how and why you implemented this feature engineering technique, how you treated base variables, why you are using the model you used, etc.
  → present as if you were in front of a senior data analyst / scientist that has a marketing academic background AND the head of marketing that has a strong data acumen.

Grading criteria

Sorted on importance:

- Details on data preparation: missing values, outliers, data quality, suppression of variables, statistical handling (normalization, dummification, quantization, …)
- Feature engineering: **creation** of new variables, modification of existing
- Algorithms: the ones you tried, strengths and weakness, concise presentation if you want, important parameters and their modification / tuning
- Recommendations: same as project 1
- Quality, fluidity and clarity of presentation
- Introduction and context
- Enhancements: more complex features handling, detailed insights of existing data, detailed ideas of new variables (with data available, with public data, …), other technical details not already graded.
- Descriptive statistics
- Basic details about the scope of the project and data
- Bonus points for other strong points

This kind of project is a never-ending task. It could take you months full-time if you want to have perfectly clean variables, compare all possible algorithms, all possible settings… That's not the point! Given the time you have to do it, it is up to you to be pragmatic, to define the list of indicators you think are relevant and on which you want to spend time, to define some model scenarios that you want to compare and to have a critical look at your work to list the sources of improvements. Think first, act second: you will always find a way to achieve what you want to do in Python / R using the internet, but you need a plan of action so you don't get lost.

The AUC will not be the only criterion, it is your overall approach that will be rated.

Metric used in the competition

Each submission is evaluated with the AUC metric.

Submission format

You will need to score each id_client in the test set. The submission (in CSV) must contain 2 columns: ID (which is equal to the client ID) and Expected (the repurchase probability you computed). Your file must have as separator the comma: ",".

Practical work guidelines

Organization

During the 2 practical sessions (and then at home), you will work in groups of 3-4 students on this project which you will present at the oral.

Each group organizes itself as it sees fit for the project and can parallelize the different components of the project (**Python (or R code)**, data analysis, PPT presentation, overall organization) but each member must be comfortable with all the steps.

Marketing recommendations

Focus on one segment and give marketing recommendations to the marketing direction using analysis and statistics tools. This part is mandatory for the presentation.