Audiobooks Business Case

1. What phenomenon do I want to model?

- Duration of the phenomenon: 2 years obtained from the audiobook app plus 6 months to verify the purchase.
- It is supervised learning, so a goal or objective is needed.
- The objective is to observe if a customer has purchased a new book in the last 6 months (1- if they bought again, 0 - if they did not buy).
- ***The phenomenon is the new purchase of the product.***
- ***Therefore, the task is to create a machine learning algorithm that can predict whether a customer will buy again.***

1.1 Analysis unit

Customers

1.2 Result of interest (Y)

Whether the customer will buy again or not.

1.3 Time horizon

The last 6 months.

1.4 Intervention

1.5 Context

Based on data from an audiobook app. The targets column shows whether a customer has purchased a new book in the last 6 months (1 - if they bought again, 0 - if they did not).

2. What does each variable represent?

Variable	*Role*
ID	None
Book_length_mins_overall	Predictor (does not cause the purchase but helps to anticipate it)
Book_length_mins_avg	Predictor (does not cause the purchase but helps to anticipate it)
Price_overall	Predictor (does not cause the purchase but helps to anticipate it)
Price_avg	Predictor (does not cause the purchase but helps to anticipate it)
Review	Contextual Predictor
Review 10/10	Preacher
Minutes_listened	Preacher
Completion	Preacher
Support_requests	Preacher
Last_visited_Minus_Purchase_date	Preacher
Targets	Outcome

3. What do I expect and why?

Hypothesis

A positive change in the X variable increases Y.

Mechanism

Why should this happen? Various metrics used in the business can help predict whether a customer will buy again on the audiobook platform.

Conditions

When should it not occur? (heterogeneity).

For example:

- If the total or average price is very high, then the customer buys again.
- If the review is very low and you buy again.
- If the number of minutes listened to is close to zero and you buy again.
- If you completed very little of the book and buy it again.
- If you didn't use the app and you buy again.

Alternatives

—

4. What would it mean for the model to learn?

Learning here means predicting the effect using log-loss since it is measuring how good the predicted probabilities are and the result is binary.

5. What does success mean?

Here it means reducing out-of-sample error to accurately predict new data.

6. What would be a trap?

It would be overfitting, meaning the model memorizes past data too well and therefore fails to predict future data accurately.

7. Write the model in 3 formats

Conceptual relationship

The likelihood that a customer will buy an audiobook again depends on their past behavior, their experience with the app, and some economic factors.

Y←f(X)

- Y = new customer purchase (yes / no)
- X = set of customer variables

Your X variables (in words):

- What type of books do you read (total and average reading time)
- How much have you spent (total and average price)
- How good was your experience? (reviews)
- How much do you use the app (minutes listened to, completion)
- If it has had friction (support)
- How recent was your activity?

Interpretable model

Since your result is binary (buy / don't buy), a typical interpretable model is a logistic regression.

Model shape

$$logit\left(P(Y_i = 1)\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_{10} x_{10i}$$logit(P(Yi=1))=β0+β1x1i+β2x2i+⋯+β10x10i

where:

- $Y_i = 1$Yi=1 → el cliente **Yes, buy**
- $Y_i = 0$Yi=0 → el cliente **does not buy**

Replacing with your variables

$$logit\left(P(Compra_i)\right) = \beta_0 + \beta_1 Book\_length\_mins\_overall_i + \beta_2 Book\_length\_mins\_avg_i + \beta_3 Price\_overall_i + \beta_4 Price\_avg_i$$

logit(P(Buyi))=β0+β1Book_length_mins_overalli+β2Book_length_mins_avgi+β3Price_overalli +β4Price_avgi $+ \beta_5 Review_i + \beta_6 Review10_i + \beta_7 Minutes\_listened_i + \beta_8 Completion_i + \beta_9 Support\_requests_i + \beta_{10} Last\_visited\_minus\_purchase_i$

+β5Reviewi+β6Review10i+β7Minutes_listenedi+β8Completioni+β9Support_requestsi+β10Last_visited_minus_purchasei

Each β answers this question:

“If this variable increases, what happens to the probability that the customer will buy?”

Causal structure (is there mediation?)

Here we ask a different question:

Do some variables not only help predict, but also explain the "path" to purchase?

In this case, there is a very plausible mediation: engagement.

Las características del cliente
→ influyen en su uso de la app
→ y ese uso influye en la recompra

Examples:

- libros adecuados + buena experiencia
  → más minutos escuchados
  → mayor probabilidad de compra

We define a mediator (M)

M = Customer engagement, captured by:

- Minutes_listened
- Completion

Equation 1: What determines engagement?

$M_i = \alpha_0 + \alpha_1 Book\_length + \alpha_2 Price + \alpha_3 Review + \alpha_4 Support\_requests + u_i$

My=α0+α1Book_length+α2Price+α3Review+α4Support_requests+ui

Translation

The app's usability depends on the type of books, the price, the experience, and the frictions.

Equation 2: What determines the purchase?

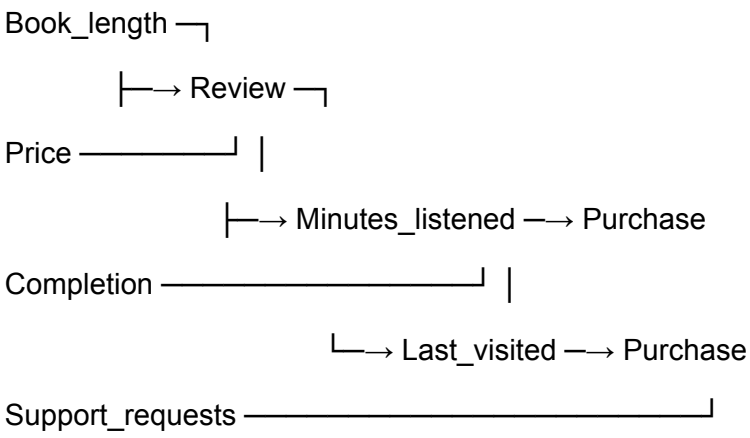
$logit\left(P(Compra_i)\right) = \gamma_0 + \gamma_1 M_i + \gamma_2 Book\_length + \gamma_3 Price + \gamma_4 Review + \gamma_5 Last\_visited + v_i$

logit(P(Buyi))=γ0+γ1My+γ2Book_length+γ3Price+γ4Review+γ5Last_visited+vi

Translation

The purchase depends on:

- directly from engagement
- and also some customer characteristics



5. Pseudocode structure

Step 1. What information do I have? Do I have input-output pairs (x, y)?

The model is supervised.

Step 2. What do I want the model to learn?

Purchase probabilities

Step 3. General model type (without libraries)

It is a more flexible relationship; it is linear, but the result is not continuous.

Step 4. How will I know if the model is “good”?

The model's performance is evaluated using log-loss on data not used during training, since the goal is to estimate well-calibrated repurchase probabilities and not just correct classifications.

4.1 How do I measure the error or the quality of the result?

Assigning good probabilities, that's why Log-loss (or cross-entropy) will be used .

Because?

- It penalizes heavily when the model is very confident and makes a mistake.
- It rewards well-calibrated probabilities.
- It is standard in issues of repurchase, churn, risk.

4.2 What does “making a lot of mistakes” mean in this problem?

Making a lot of mistakes means:

- Decir que un cliente seguro compra (probabilidad alta) → y no compra.

- O decir que un cliente no comprará → y sí compra.
- Assigning extreme probabilities that do not reflect reality.

4.3 Which metric best reflects my objective?

The metric that best reflects the business objective is the log-loss, because it measures the quality of predicted probabilities and penalizes errors with high confidence.

Step 5. What does “learning” mean in this problem?

Minimize the probability prediction error.