Customer Churn and Win-Back Targeting

A. Business problem in one page

You work for a subscription business. Leadership wants two things that can be used soon.

1. A reliable score for each active customer that estimates the chance they will cancel in the next period.
2. A simple, budget aware rule that tells the retention team whom to contact and how many to contact per 1,000 customers.

Your job is not only to produce a model. Your job is to deliver an end-to-end process that turns raw data into a calibrated probability and then into an action a manager can follow.

B. Data and release plan

You will receive two files on Canvas so every team works from the same schema.

- train.csv is available now. It contains features and labels. Treat it as the past.
- holdout_features.csv will be released on Monday, Nov 10,2025. It contains features only.

The holdout may differ from training in two ways:

1. the overall churn rate may shift by as much as five percentage points, and
2. the mix of add-on services may shift modestly.

C. Tools and where to work

Choose one path.

- **Python in Google Colab** as a single notebook.
- **R in Posit Cloud** as a single R Markdown.

Your notebook or document must run end-to-end from a clean runtime without manual fixes.

Suggested headings inside your notebook or document:

1. Title, team, date
2. Business problem
3. Leakage policy
4. Setup
5. Load train.csv

6. Features
7. Baseline model (logistic)
8. Other models and calibration
9. Decision rule and cost table
10. Save figures and files
11. Holdout scoring cell or chunk
12. Clear recommendation for managers
13. Appendix: AI use, prompts, and reflection

D. Leakage policy and how you will enforce it

Definition. Leakage is any use of information that would not exist at the time you decide whom to contact.

Assume you score customers at the end of a billing period to plan outreach for the next 30 days. Only information available up to that scoring time is allowed.

Your report must include one paragraph that lists the leakage risks you checked and how you prevented different sources of leakage.

E. Training, validation, and calibration

Work only on train.csv until the holdout is released. Use either 5-fold cross-validation or a simple time-aware split after sorting by tenure. Report:

- **Discrimination** on validation data (AUC).
- **Calibration** on validation data (Brier score).

Freeze your final pipeline before you touch the holdout file.

F. Models you may use

Use exactly three models.

- Baseline: **logistic regression** with regularization.
- Two additional Models: random forest and gradient boosted trees. [**Caution:** The churn label is class imbalanced (the 'Yes' class is a minority).]

Calibrate the final model (Platt scaling) and include the reliability curve. If you compress many related fields with **PCA**, explain in plain words what each component roughly captures.

About Platt Scaling
Platt scaling is a probability calibration technique used to convert a model's uncalibrated prediction scores into well-calibrated probabilities. It works by fitting a logistic regression model on the validation predictions of your chosen classifier (e.g., Random Forest or XGBoost) and their true labels. The logistic model learns a mapping that aligns the model's raw scores with the actual observed frequencies of the positive class (here, churn). This step ensures that predicted probabilities (such as 0.7) accurately reflect real-world likelihoods (about 70% of those customers actually churn). The goal is to make your model's predicted probabilities reliable and interpretable for business decision-making, as reflected in the required reliability (calibration) curve.

References if you are interested in more information about Platt Scaling:

- Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers, 10(3), 61-74.
- **Scikit-learn User Guide:** Probability Calibration. https://scikit-learn.org/stable/modules/calibration.html
- Niculescu-Mizil, A., & Caruana, R. (2005, August). Predicting good probabilities with supervised learning. In Proceedings of the 22nd international conference on Machine learning (pp. 625-632). https://www.cs.cornell.edu/~alexn/papers/calibration.icml05.crc.rev3.pdf

---

G. Required execution evidence

Your notebook or document must generate the following **from code**:

1. oof_predictions.csv built on train.csv.
2. figures/calibration.png with the reliability curve.
3. An ablation table[1] that shows how AUC and Brier[2] change when you drop each feature group you defined.

[1] An ablation table is a small results table that shows what happens when you remove one part of your model or features at a time.
- Each row: the part you removed.
- Columns: your metrics after removal, and the change from the full model.

[2] Brier score checks how close your predicted chances are to what actually happened. For each case, take your predicted probability p and the real outcome y. $y = 1$ if it happened, 0 if not.
Compute $(p - y)^2$. Average that number over all cases. That average is the Brier score.

We will re-run your notebook and we should be able to reproduce your results.

H. Holdout protocol

On Monday, Nov 10, you will receive your team's holdout_features.csv.

- Run the single Holdout Cell to read the file, score it with your frozen pipeline, and write predictions.csv with two columns in the exact input order: customer_id,p_churn
- After labels are released, add a short "Holdout results" paragraph and a holdout-only calibration plot.

I. Turning scores into actions

A manager can spend a fixed amount per 1,000 customers. You must choose one threshold on p_churn that fits the budget.

- Present a small **cost table** at the chosen threshold. Show contacts per 1,000, expected saves, expected cost, and expected net value.
- State the rule in one sentence that a manager can follow. Example: "Contact anyone with p_churn ≥ 0.28, except customers with tenure under two months."

J. AI use, citation, and reflection

You may use an LLM for writing help, boilerplate code, or idea generation. You may not use it to fabricate results.

Include a final section titled **AI use, prompts, and reflection** with:

- what you used the model for,
- prompts or a brief prompt log,
- a citation such as:
  "OpenAI, ChatGPT (GPT-5 Thinking). Conversation with the author, November 2025. Prompts and excerpts documented in the appendix."
- a short reflection on what helped, what did not, and one rule you will follow next time.

Variable Description

- **ID** — Unique customer identifier.
- **gender** — Male or Female.
- **SeniorCitizen** — 1 if the customer is a senior citizen, 0 otherwise.
- **Partner** — Customer has a partner: Yes or No.
- **Dependents** — Customer has dependents: Yes or No.
- **tenure** — Number of months the customer has stayed with the company.
- **PhoneService** — Customer has phone service: Yes or No.
- **MultipleLines** — Customer has multiple phone lines: Yes, No, or "No phone service."
- **InternetService** — Type of internet connection: DSL, Fiber optic, or No.
- **OnlineSecurity** — Online security add-on: Yes, No, or "No internet service."
- **OnlineBackup** — Online backup add-on: Yes, No, or "No internet service."
- **DeviceProtection** — Device protection add-on: Yes, No, or "No internet service."
- **TechSupport** — Tech support add-on: Yes, No, or "No internet service."
- **StreamingTV** — Streaming TV add-on: Yes, No, or "No internet service."
- **StreamingMovies** — Streaming movies add-on: Yes, No, or "No internet service."
- **Contract** — Contract term: Month-to-month, One year, or Two year.
- **PaperlessBilling** — Billing is paperless: Yes or No.
- **PaymentMethod** — Payment method: Electronic check, Mailed check, Bank transfer (automatic), or Credit card (automatic).
- **MonthlyCharges** — Current monthly charge amount.
- **TotalCharges** — Total amount charged to date (may contain blanks that must be coerced to numeric during cleaning).
- **Churn** — Target label: Yes if the customer left during the target window; No otherwise.