ZERVE AI DATATHON — PROBLEM STATEMENT

---

1. Introduction

Health insurance companies continuously evaluate the risk associated with individual customers. Accurately identifying high-risk customers—those who are more likely to file significant health insurance claims—helps insurers improve customer service, reduce fraud, and optimize pricing strategies.

In this datathon, your goal is to develop a machine-learning model that predicts the likelihood of a customer filing a health insurance claim. You will work with an anonymized dataset containing 50 engineered features.

This challenge tests your ability to build robust ML pipelines, handle mixed-type data, and optimize models for imbalanced classification.

---

2. Objective

Develop a machine-learning model that predicts:

The probability that a customer will file a health insurance claim (target = 1).

Your final submission must contain **predicted probabilities for each row in the test dataset**, not the training dataset.

---

3. Dataset Description

Two datasets will be provided for this datathon:

A. Training Dataset

- Contains 50 anonymized features
- Includes the binary **target column (`target`)**

- Used for model training, experimentation, validation, and tuning

B. Test Dataset

- Contains the same 50 anonymized features
- **Does NOT contain the `target` column**
- Your predictions must be generated on this dataset
- Evaluation will be performed using hidden true labels

All personally identifiable information has been removed, and original domain-specific feature names have been replaced with generic identifiers.

3.1 Feature Types

Binary Features (0/1)

feature_4
feature_5
feature_6
feature_11

feature_14
feature_16
feature_18
feature_19
feature_20
feature_21
feature_22
feature_27
feature_30
feature_32
feature_41
feature_44
feature_46

Categorical Features

feature_3
feature_7

feature_8
feature_12
feature_15
feature_23
feature_25
feature_28
feature_31
feature_34
feature_35
feature_39
feature_42
feature_49

Numeric / Continuous Features

feature_1
feature_2
feature_9
feature_10
feature_13
feature_17
feature_24
feature_26
feature_29
feature_33
feature_36
feature_37
feature_38
feature_40
feature_43
feature_45
feature_47
feature_48
feature_50

3.2 Target Variable

Column: `target`

- `1` → Customer filed a significant health insurance claim
- `0` → Customer did not file a claim

The target variable is **present only in the training dataset**.
The test dataset does not contain the target and is used solely for prediction.

The dataset is imbalanced, meaning that claim events are relatively rare. Handling this imbalance effectively is essential for achieving strong performance.

---

4. Evaluation Metric — Normalized Gini Coefficient

The official leaderboard metric is:

Normalized Gini Coefficient

The Gini coefficient measures how well the ranking of predicted probabilities differentiates between claim and non-claim customers.

Why Gini?

- It directly measures model discrimination ability
- It is highly sensitive to ranking quality
- It is widely used in insurance risk modeling
- It performs well with imbalanced datasets

Formula Overview

Normalized Gini =
(Gini(model predictions) / Gini(perfect model))

Your model must output a **probability between 0 and 1** for each test dataset row.

---

5. Scoring & Judging Criteria

Final evaluation will be based on two equally weighted components:

1. Model Performance (50%)

- Evaluated using the **Normalized Gini Coefficient** on the hidden test dataset

- Higher Normalized Gini → better ranking

2. Final Presentation (50%)

Shortlisted teams will present:

- Methodology and approach
- Feature engineering
- Handling of missing values and imbalance
- Validation strategy
- Model selection and tuning
- Insights, reasoning, and interpretation

Judges will assess:

- Technical clarity
- Methodological rigor
- Creativity and innovation
- Communication quality

Final Score Formula

Final Score = 0.5 × Normalized Gini + 0.5 × Presentation Score

---

6. Submission Format

Participants must submit:

A. Prediction File (CSV)

For every row in the test dataset, the file must contain:

- `id` → Row identifier (same order as test dataset)

- `target` → Predicted probability (between 0 and 1)

B. Zerve Canvas

- Your final working code and notebook must be submitted in a **Zerve Canvas**
- This will be reviewed during the final presentation round

---

7. Timeline

- **Submission Deadline (Prediction CSV):** 20th December 2025
- **Final Presentations:** 23rd December 2025