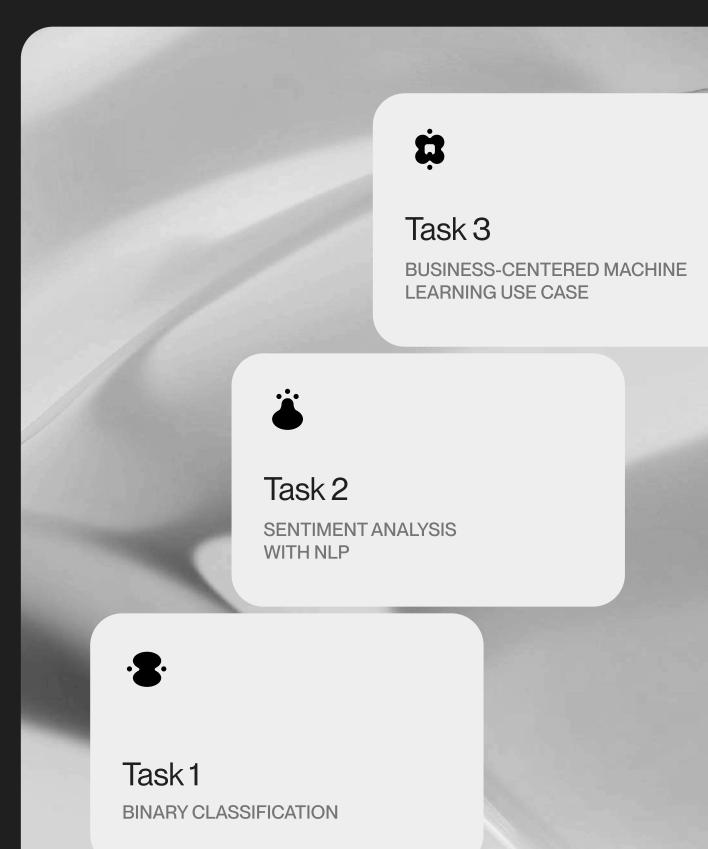


Al Interview Kit

# 3 Technical Tasks with Solutions, Evaluation Tips & Red Flags

### About these tasks

These technical tasks are designed to assess how AI developers approach real-world challenges—from data cleaning and modeling to evaluating results and aligning with business goals. Use the solutions, red flags, and evaluation tips to make your interviews structured, efficient, and insightful.



# Binary classification



You're building a basic spam filter. You receive a CSV with features like num\_links, has\_attachment, email\_length, sender\_reputation\_score, and a binary target column: is\_spam (1 for spam, 0 for not spam). Assume that only 5% of emails are spam.

#### Task for the candidate

- (01) BUILD AND EVALUATE A BINARY CLASSIFICATION MODEL
- (02) HANDLE ANY NOISY OR MISSING DATA
- (03) PRESENT EVALUATION METRICS (ACCURACY, PRECISION, RECALL)
- 04) COMPARE AGAINST A DUMMY CLASSIFIER THAT ALWAYS PREDICTS "NOT SPAM"
- (05) EXPLAIN ANY PREPROCESSING OR FEATURE ENGINEERING DECISIONS



# Task 1 Binary classification



- PREPROCESSING
   Fill missing values (median), scale features (optional), remove outliers.
- MODELING
   Use Logistic Regression, RandomForestClassifier, or XGBoost. Stratified train/test split.
- EVALUATION
   Precision, recall, F1-score, ROC-AUC. Compare to baseline.
- EXTRAS
   Discuss overfitting and model drift concerns.

#### What to look for

- HANDLES IMBALANCE
   (E.G., CLASS WEIGHTS, RESAMPLING)
- INTERPRETS CONFUSION MATRIX CLEARLY
- EXPLAINS MODEL AND METRIC CHOICES
- MENTIONS PRODUCTION CONSIDERATIONS

#### ► Red flags

- ONLY REPORTS ACCURACY
- IGNORES CLASS IMBALANCE
- APPLIES DEEP LEARNING UNNECESSARILY
- DOESN'T EXPLAIN PREPROCESSING DECISIONS

Bonus flags

DEPLOYMENT

How would you monitor precision/recall drift over time?

ALTERNATIVE

Could use anomaly detection for evolving spam patterns.



# Sentiment analysis with NLP



You're given a dataset of customer reviews labeled as **Positive**, **Negative**, or **Neutral**.



- (01) BUILD A SENTIMENT CLASSIFIER USING A SIMPLE NLP PIPELINE
- (02) TOKENIZE AND VECTORIZE THE TEXT
- (03) TRAIN A LIGHTWEIGHT MODEL (E.G., LOGISTIC REGRESSION, NAIVE BAYES)
- (04) OUTPUT CLASSIFICATION METRICS AND A CONFUSION MATRIX
- (05) ADDRESS NEUTRAL CLASS DIFFICULTY
- 06) SHOW EXAMPLES OF MISCLASSIFIED REVIEWS FOR ERROR ANALYSIS



# Sentiment analysis with NLP



- PREPROCESSING
   Lowercasing, remove punctuation, stop words, optional lemmatization.
- VECTORIZATION
   TF-IDF or CountVectorizer. Limit vocab if needed.
- MODELING
   Logistic Regression or Naive Bayes. Stratified split.
- EVALUATION
   Macro F1, per-class recall, confusion matrix.
- IMPROVEMENTS
   Mention oversampling, class weights, potential use of transformers.

#### What to look for

- BALANCED EVALUATION METRICS (NOT JUST ACCURACY)
- INSIGHT INTO WHY NEUTRAL CLASS IS HARD
- JUSTIFIES MODEL AND PREPROCESSING CHOICES
- ACKNOWLEDGES MODEL LIMITATIONS

### ► Red flags

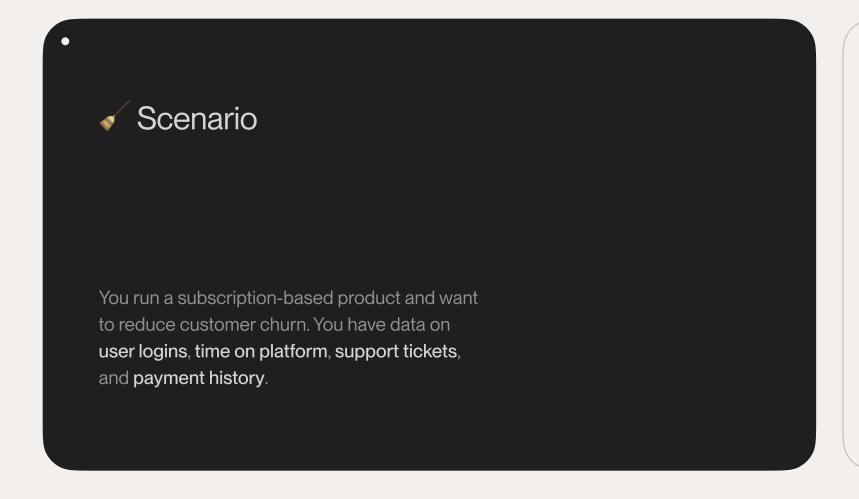
- SKIPS TEXT PREPROCESSING
- JUMPS TO DEEP LEARNING UNNECESSARILY
- DOESN'T ADDRESS NEUTRAL CLASS OR CONFUSION MATRIX
- IGNORES CLASS IMBALANCE

Bonus flags

DEPLOYMENT What if too many reviews are labeled Neutral? ALTERNATIVE
 Combine rule-based logic with ML for edge cases.



## Business-centered machine learning use case





- (01) OUTLINE A MACHINE LEARNING SOLUTION TO PREDICT CHURN
- (02) IDENTIFY NEEDED DATA AND LABELING STRATEGY
- (03) SUGGEST A MODELING APPROACH AND JUSTIFY IT
- 04) DEFINE EVALUATION STRATEGY, INCLUDING BUSINESS IMPACT
- (05) CONSIDER COST-SENSITIVE EVALUATION AND INTERVENTION PRIORITIZATION



# Business-centered machine learning use case



- FRAMING
   Binary classification; churn = inactive 30+ days or cancelled.
- DATA
   Logins, session length, support tickets, subscription type, payment history.
- MODELING
   Start with Logistic Regression, then Gradient Boosted Trees. Use SHAP for explainability.
- EVALUATION
   ROC-AUC, recall. Track impact on retention rate or revenue lift.
- IMPROVEMENTS
   Use model scores to trigger retention offers.

#### What to look for

- BUSINESS-ORIENTED FRAMING OF THE PROBLEM
- SOLID FEATURE SUGGESTIONS
   WITH CLEAR RATIONALE
- DISCUSSION OF TRADE-OFFS AND THRESHOLDS
- TIES PREDICTIONS TO ACTIONS (E.G., MARKETING, RETENTION)

### ► Red flags

- DOESN'T DEFINE CHURN PRECISELY
- NO MENTION OF COST OF FALSE NEGATIVES
- CHOOSES HIGH-COMPLEXITY MODELS WITHOUT JUSTIFICATION
- IGNORES HOW MODEL FITS INTO BUSINESS WORKFLOW

Bonus flags

DEPLOYMENT
 How would you monitor drift in user behavior?

ALTERNATIVE
 Segment users with clustering before modeling churn.



# A summary table comparing tasks side-by-side

TASK	SKILLTESTED	COMMON PITFALLS	GOOD SIGNS
TASK 1: SPAM FILTER	Data prep, model evaluation	Ignores imbalance, uses accuracy only	Discusses metrics, class weighting
TASK 2: SENTIMENT	NLP basics, multiclass	Skips preprocessing, uses BERT too early	Chooses simple models, analyzes errors
TASK 3: CHURN	Business thinking, feature selection	Ignores imbalance, uses accuracy only	Discusses metrics, class weighting

