Applied Machine Learning 2025: mid-term examination

Please refer to the following use case in order to answer the questions below. Your answer must contain the choice A or B or both, together with a brief explanation - why. The explanation is as important as the correct answer. Please summarize the explanation succinctly in two to three sentences. If you absolutely feel the need to use more space to explain, please use an extra sheet and note down the appropriate reference to the question number. In case of apparent ambiguity, please specify your assumptions in the why part to clearly express your thought process.

Use Case:

Consider a scenario when a telecom company wants to minimize the customer attrition rate by giving one month of unlimited data usage to 20% of its customers. One of the manager suggests to send the gift in an unbiased way to every 5th customer ordered by their last name. Some others believe that there are better data driven strategies for this marketing promotion. Let us say that the company has about 1 million customers and the monthly churn rate is 1-2%. The budget for promotion can only support 200k customers.

Problem Framing

1.1 (big picture)

What is the burning issue? Why?

A. churn
B. targetted upsell

1.2 (treatment)

What is a proposed treatment? Why?

A. identify likely churners and offer one month promotion to them
B. identify high value offers and promote them to all customers

1.3 (goal)

What is the optimization goal? Why?

A. minimize budget for promotion s/t very limited churn
B. minimize churn s/t limited budget for promotion

1.4 (why ML)

What is the ML value proposition? Why?

A. effective targeting based on demographics, latest usage pattern and complaints
B. personalized upselling of higher value products based on demographics

Solution Design

2.1 (data)

What could be an issue with data? Why?

A. class imbalance
B. data sparsity

2.2 (model)

Which metric would you use for model selection? Why?

A. AUC ROC
B. f1

2.3 (action)

What action would you take based on the model? Why?

A. rank by pr(churn) and offer promotion to top 20%
B. if pr(churn) > 0.8 then offer promotion

2.4 (impact)

How would you measure impact? Why?

A. cancellations/month
B. data usage/month

System Design

3.1 (requirements)

Which one is a functional requirement? Why?

A. obtain sentiment of user complaint
B. obtain propensity score of churn

3.2 (estimations)

Which configuration would you choose to serve the system? Why?

A. single node
B. cluster

3.3 (process flow)

Which is better? Why?

A. prepare features offline every month
B. prepare features online in real-time

3.3 (architecture)

Which is better? Why?

A. batch
B. streaming

MLOps

4.1 (unit testing)

What would you use to unit test the model? Why?

A. threshold value
B. regularization param

4.2 (model debugging)

If train metrics are high and test metrics are low what does it indicate? Why?

A. overfit
B. underfit

4.3 (monitoring)

What would you use to measure data drift? Why?

A. complaints/month
B. monthly bill

4.4 (feedback loop)

What might be an issue in next round of modeling? Why?

A. selection bias
B. lack of data

Statistical Learning

5.1 (methods)

Which method would you use to identify different segments of data pattern usage and complaints? Why?

A. regression
B. clustering

5.2 (limitations)

What would most impact the Bayes error rate? Why?

A. label ambiguity
B. data sample size

5.3 (data imbalance)

Would you evaluate your model metric on balanced or imbalanced data? Why?

A. imbalanced
B. balanced

5.4 (feature normalization)

Which feature might have to be normalized? Why?

A. number of complaints last month
B. data usage last month