Master's degree in Banking and Financial Regulation

University of Navarra

Introduction to Modeling evaluation assessment

1. a) We have a dataset of about 5,900 rows and 65 columns. Whether this is truly enough depending on a few things:
    - First, we have to check if the sample really represents all the types of customers in the bank's credit card portfolio. If it's truly random, that's great; but if it was cherry-picked or limited somehow, we might miss whole segments of customers (like high-risk groups or certain income brackets), making our analysis incomplete.
    - Second, even though 5,900 rows might seem like a lot, predicting credit risk or building detailed customer segments often need larger datasets, especially to capture rare events or niche patterns.
    - Third, we should also consider data quality and completeness. In this case there aren't any missing values or important missing fields, which allows us to have a more thorough analysis. Overall, it might be enough for an initial exploration but we might need more data—particularly if we want solid, bank-wide insights or advanced modeling.

    b) From this smaller dataset, I can begin some exploratory analysis to get a sense of how our credit card portfolio looks. However, before I do anything too detailed or start recommending big decisions, I need to ensure that the data covers the different types of customers we have. If this subset is truly representative, meaning it includes a good mix different income brackets then I'll start by checking how well this sample reflects our overall customer base.

2. a) No, I wouldn't accept the model right away. Even though it showed remarkable performance in the initial phase, its real-time performance during the pre-trial doesn't match those impressive metrics. In a real business setting, consistent performance—especially when deployed—is crucial. If it's no longer reliable in production or under real operating conditions, we shouldn't accept it as is.

    b) Overfitting- The model might have been too closely tailored to the training data, learning patterns specific to that dataset rather than general rules. Once it faced real-world data during the pre-trial, it failed to perform as well.

    c) Encourage them to use k-fold cross-validation or bootstrapping to ensure the model generalizes well and reduce the risk of overfitting.

3. a) Many machine learning models (like linear regression, logistic regression, neural networks) can only work with numerical data. They don't inherently understand labels like "Male/Female" or "High/Medium/Low" unless those labels are converted to numbers.

b) - Label Encoding: Assigns each unique category an integer (e.g., "Red" = 0, "Green" =1, "Blue" =2)

- One-Hot Encoding: Creates new binary columns for each category (e.g., "Red" -> [1,0,0], "Green" -> [0,1,0], "Blue" -> [0,0,1])

c) Yes, there are some conceptual similarities: Numerization can be as straightforward as assigning integers or using dummy variables. Filling missing values might involve simple rules like using the mean, median, or a special placeholder. In both cases, you apply a systematic, consistent approach.

4. Provider 1 Report:
    - Provider 1's Logistic Model stands out with an AUC of 0.92, which is quite high. This suggests strong predictive performance compared to its Random Forest and Boosted Decision Trees.

    Provider 2 Report:

    - All three models from Provider 2 have the same AUC (0.88). This shows consistent performance across different model types, but none exceed the 0.90 mark.

    Provider 3 Report:

    - Provider 3's best model is the Boosted Decision Trees at 0.89 AUC, which beats their Random Forest (0.87) and Logistic Model (0.81).

Conclusion: I would endorse Provider 1 because it currently leads the top AUC score. However, it's also wise to look at other factors (like model explainability, operational ease, implementation costs, and track record of each provider).

5. a) Yes, there is a potential issue. the PSI results suggest that 6 features show signs of distribution shift (i.e., they may be behaving differently compared to when the model was originally trained). This discrepancy could mean the model is starting to see data that doesn't match its original training distribution for those specific features.

b) Since the overall model performance (AUC) is still above 0.80, the immediate impact on the bank's risk metrics might be limited for now. The model is still classifying relatively accurately in aggregate.\

Right now, the cost of risk probably isn't spiking because performance remains acceptable, but it's worth keeping a close watch on those 6 features to ensure they don't further derail the model's accuracy.

c) - Economic changes, new regulations, or changes in customer behavior

- Changes in how data is collected, coded, or stored could lead to differences in feature values.

d) Not necessarily urgent, given that the AUC is still healthy. But we shouldn't ignore the PSI warnings.

e) - Identify why these 6 features have changed. Is it a macro-level shift (e.g., economy) or something internal

- Instead of a full re-training, see if re-calibrating the model or adjusting thresholds for these 6 features is sufficient.