Final Project

March 25, 2024

1   Project Description

Our insurance company provides several ways for its policyholders to make payments. While our service counselors can take payments over the phone, it is more cost-efficient for customers to make payments through our self-service channels, such as online or through the automated phone system. We would like to use a predictive model to select people to receive a pre-emptive e-mail message designed to encourage them to pay online.

You have been tasked with identifying which customers are likely to make a service payment call in the next 5 days. The attached file contains data on customers who have had a bill due in the next 5 days and whether they made a service payment call. Construct a model that predicts the likelihood that each policyholder will make a service payment call (CALL_FLAG=1). You may use whatever models *you learned in this class* fit and Python packages you choose to complete your analysis.

2   Requirements

Include any relevant workflow, code, or scripts you used. Summarize your data preparation, variable selection, modeling methods, and fit statistics into a presentation. You should, at a minimum, describe how you selected model inputs, any data pre-processings performed, how you dealt with missing or extreme values, reasons for your selected model, and how your analysis addresses the business problem. **Assume the audience of your presentation is fellow data scientists with limited knowledge about this specific problem.**

3   Criteria

Your submission will be assessed by the criteria outlined below (The example questions are just guidelines, not conditions, for passing the assessment):

- Model quality:

- What insights do you identify by analyzing the data? How did those insights contribute to your formulation of the model, the features you selected/engineered, and the ML models you chose?
- What metrics do you use to evaluate the model? How do you set up your training experiments so that the metrics are as unbiased as possible?
- How do you decide whether your model is good enough for real-world application? Why do you choose your model over others?

- Code quality:

  - Is your code easy to follow for potential collaborators and/or reviewers? Do you make comments to your code?
  - Is your codebase reproducible, extensible, and portable?
  - What are hyperparameters in your code? How do you tune them?
  - Python is the only programming language you can use in this project.
  - Do you split your data into training and validation (or test) sets?

- Presentation quality:

  - Is your presentation organized coherently and concisely? Is your presentation friendly to the audience with limited background?
  - Does your presentation include a background introduction, data cleansing, model training, evaluation, and conclusion?
  - Does your presentation include visualizations of your data, model performance, and results?
  - Does the overall presentation make a compelling case for your recommendations?

4   Grading Policy

- [**20pts**] Project report
- [**80pts**] In-class presentation
  - [**10pts**] Peer evaluation
  - [**70pts**] Professor evaluation

5   What you need to submit online

1. All of your Python scripts.

2. A project report formatted in PDF.

3. Your presentation to the audience. (There is no need to record your speech in voice or video. Slides formatted in PDF are enough.)

4. Both report and presentation must include, but not be limited to

   - Overview/outline of your presentation
   - Background introduction of the project
   - Data cleansing procedures
   - Your learning model
   - Model performance and evaluation
   - Your observations/conclusions
   - Take home messages for customers