INF 428/528 – Final Project

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 methods you see fit and any modeling package you choose to complete your analysis.

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 transformations preformed, how you dealt with missing or extreme values, reasons for your selected algorithm, and how your analysis addresses the business problem. Assume the audience of your presentation are fellow data scientists with limited knowledge about this specific problem.

Grading criteria:

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

- **[30 pts]** 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 select/engineer, and the ML algorithms you choose?
  - What metrics do you use to evaluate the model? How do you set up your training experiments such 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?
- **[30 pts]** 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?
- **[40 pts]** Presentation quality:
  - Is your presentation organized coherently and concisely? Is your presentation friendly to the audience with limited background?
  - Does your presentation include 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?

What you need to submit:

1. All of your Python files.
2. Your presentation to the audience. (There is no need to record your speech in voice or video. Slides formatted in PDF are enough.) Your presentation must include, but not be limited to
   a. Overview/outline of your presentation
   b. Background introduction of the project
   c. Data cleansing procedures
   d. Your learning model
   e. Model performance (metrics or visualizations)
   f. Your observations/conclusions