2024 Machine Learning Take Home Assignment: Sentiment Analysis

Problem

You are driving the team in understanding how sentiment expressed in a consumer finance call correlates with business outcomes. For this purpose, you have collected a sample dataset with:

1. Call transcriptions, i.e. transcribed text of Agent and Borrower utterances, with timestamps
2. Call outcomes, called "dispositions", as specified by the agent.

The de-identified data has been shared along with this assignment. By downloading the data, you are agreeing to use this data solely for the purpose of the interview process, and that you will not retain it beyond the interview process or share it with any third parties.

Your goal is to generate two scores for each call, representing the overall agent and borrower sentiment. These scores should correlate with, and be predictive of, the provided call dispositions.

Deliverables

- A `csv` containing call IDs with the agent and borrower sentiment scores for each call
- All the python code you wrote to solve the problem, including data exploration, modelling and evaluation. Jupyter notebooks are preferred.
  - Be sure to include a `requirements.txt` file to help us install relevant packages and include a `README` with instructions on how to run your code.
- A short document answering the following questions:
  - What was your overall strategy for this problem?
  - What were your hypotheses?
  - How did you evaluate your proposed method? Why did you choose this method?
  - What were your findings?
  - What features would you add and what analysis would you do if you had more time?
  - What other things would you want to try before deploying your model to production?

Guidelines

- Regardless of the time given to you for this assignment, plan to spend 6 to 8 hours overall, including the writing of the report. Come up with a plan of action before starting to get the most out of this limited time. Scope down your exploration to fit within the time frame and mention other directions you would like to explore in your document.
- Feel free to use publicly available sentiment models, such as those on HuggingFace as a starting point. Be sure to justify, both qualitatively and quantitatively if possible, why the chosen model is a good fit for this dataset.
- You do **not** need to train a model from scratch for this problem or even finetune a pretrained model on the provided dataset. However, you have the freedom to do either of these if you wish.
- What we are looking for here is a high quality **evaluation** of your approach and the generated sentiment scores. Our goal is to find insights explaining how sentiment correlates with various dispositions. Asking the right questions and presenting a well-written report is more important than training the best possible model.

Please DO NOT share this assignment or your solution with anyone, and delete the data once you are done with the interview process. We sincerely appreciate the fact that you’re willing to spend your time on this. Thank you and have fun!