Project Description: Quality Assurance Machine Learning Task

Overview

This project is a structured machine learning task taken from a quality assurance and data science collaborative project. The goal is to analyze LLM outputs related to call transcripts to classify the part of the sales process each call belongs to using machine learning models.

Both incoming and outgoing call transcripts are automatically forwarded to a service, which sends them to an API endpoint. The API then evaluates the transcript based on approximately 200 predefined questions to assess compliance with business standards. However, not all standards are always required, and calls can vary significantly depending on where the application is in the sales process.

Our focus in this task is on two main categories:

Category 6 (Introduction Calls)

These calls focus on gathering customer preferences and briefing them on the loan application process. Examples of LLM outputs in this category include:
- "Did the agent take the time to understand the customer's car preferences?"
- "Did the agent brief the customer on what tasks they will complete to improve the chances of a successful application?"

Category 7 (Final Loan Details & Next Steps)

These calls focus on confirming the final loan details and ensuring the customer understands additional benefits. Examples of LLM outputs in this category include:
- "Did the agent clearly explain each benefit of additional insurance?"
- "Did the agent confirm the next steps of the application process?"

Data Structure

1. **High-Level Key Segments**: Scans the entire call for high-level information without going into detail.

2. **Business Standards:** Compliance checks and general sales practices that apply to all call categories.

3. **Low-Level and Mid-Level Key Segments:**
   - **Low-Level:** Provides key details of what was expected from a Category 6 call.
   - **Mid-Level:** Provides key details of what was expected from a Category 7 call.

4. **Legacy Segments:** This segment contains information from a previously used service.

While this service is no longer active, its data is retained for business continuity and may still contribute to model performance.

Task Requirements

- You are provided with approximately **400 labeled examples** and **2000 unlabeled examples**.
- The objective is to **Train a machine learning model** on the labeled examples and predict the classification of the unlabeled examples.
- The priority is to **maximize recall, even if it comes at the expense of some accuracy**, as the current heuristic-based approach has **low recall (2%) but high accuracy (80%)**.
- A recall and accuracy of **80%** for Category 6 and 7 calls is considered **acceptable**, while values in the **high 90s** are achievable with a structured machine learning method.
- An **F1 score of 80-90%** is achievable and should be considered a strong benchmark.
- Students may justify alternative evaluation metrics beyond F1, accuracy, and recall, but should provide clear reasoning.

Deliverables

1. **Two Model Approaches:**

- A **standard** machine learning approach:
  - Create a reusable script that pre-processes the data for a standard ML algorithm (e.g., a supervised Scikit-learn model).
  - Fit the ML algorithm.
  - Evaluate the model.
  - Report how many Category 6 and 7 calls are classified as a result.

- A **more advanced** semi-supervised approach:
  - Experiment with a novel model selection and feature engineering process.
  - Evaluate the model.
  - Explain the difference in outputs compared to the supervised model.
  - Provide a layman-friendly explanation of why this approach did or did not improve the process.

Suggested Advanced Methods

For the advanced modeling section, students may consider:

- **Autoencoders and Variants**
- [Unsupervised Learning in Scikit-Learn](https://scikit-learn.org/stable/unsupervised_learning.html)
- [Semi-Supervised Learning in Scikit-Learn](https://scikit-learn.org/stable/modules/semi_supervised.html)

- **Ensemble models or sequential prediction methods**

These techniques can help improve recall and accuracy by leveraging unlabeled data or combining multiple models for better decision-making.

Additional Notes

- The dataset may contain noisy or seemingly irrelevant data. While some information might not appear useful at first glance, it has been retained as it may hold value from a data science perspective.
- There are no intentional "traps" or misleading data points designed to catch you out.

By completing this task, you will contribute to the development of a **reliable, high-recall machine learning model** for classifying business calls, improving upon the limitations of heuristic-based methods.