Data Science & LLM Technical Assessment for RLDatix

Reynald Havard

Predictive Modelling (Binary classification)

The task began with data cleaning and exploratory analysis. The dataset included 200 rows, 7 features (age, gender, diagnosis code, num_previous_admissions, medication_type, length_of_stay, and discharge_note), and a binary target, readmitted_30_days. There were no missing values or anomalies, but the target was imbalanced (67.5% non-readmitted, 32.5% readmitted).

Although no strong correlations were found, certain patterns emerged: younger patients, males, those with no previous admissions, diagnosis code D003, type C medication, and discharge notes mentioning blood pressure issues were more likely to be readmitted.

Feature engineering followed. Numerical features were centered and scaled for logistic regression, and categorical features, including discharge notes, were one-hot encoded. The dataset was split 80/20 into training and testing sets, with stratification to preserve the readmission ratio. Three models were trained: logistic regression, random forest, and XGBoost. Evaluation metrics included ROC AUC, precision, recall, F1-score, accuracy, and confusion matrices.

Model performance was generally poor. While predictions for non-readmissions were accurate, all models struggled with readmissions, and ROC AUC scores were low. A different data split improved these scores slightly, but overall performance remained weak.

Finally, a random forest was trained on the full dataset and SHAP values were computed. Although the model was not highly accurate, it did reflect some patterns observed during analysis, such as the influence of age, blood pressure issues, and previous admissions.

Overall, the models failed to predict the minority class effectively. Future improvements could involve generating synthetic data (for example using SMOTE), applying repeated stratified k-fold cross-validation, using more advanced processing for discharge notes (for example using a TfIdf vectorizer or Word2Vec), improving feature engineering, tuning hyperparameters, and, most importantly, collecting more data, both in terms of patient numbers and feature richness. While the SHAP values seemed to show some interesting relationships between features and the binary target, evaluation shows that we cannot use these models to predict non-readmissions and readmissions.

Named Entity Recognition from Discharge Notes (LLM/NLP)

The goal of this task was to extract entities such as diagnosis, treatment, symptoms, medication, and follow-up, from discharge notes using NLP techniques or large language models (LLMs). The approach used the "Qwen/Qwen2.5-1.5B-Instruct" model from Hugging Face, using prompt engineering to guide the model in identifying entities. Other models, like "google/flan-t5-large," were tested but performed worse, either failing to extract entities or producing inconsistent outputs. Instruct models generally followed directions more reliably.

Some entities were correctly extracted, but the discharge notes were unstructured and often lacked certain categories: only one note mentioned a specific medication (antibiotics). There were also overlaps between entity types, with diagnosis and symptoms sometimes mixed, as well as treatment, medication, and follow-up.

While this method shows promise, it would benefit from a domain-specific model. Annotating a subset of the data and fine-tuning either a named entity recognition model or a text-generation model like the one used here could improve clarity and accuracy in identifying entity types.

Finally, LLMs can hallucinate. In some cases, especially for follow-ups, the model generated information not present in the notes, despite prompts explicitly instructing otherwise. Future improvements could include verifying that extracted text appears as in the original notes.