Data Science & LLM Assessment

Predictive Modelling

For the first task, the individual features are studied by plotting the distributions for the numerical features and displaying the distributions for the categorical features for readmitted and not readmitted patients. The distributions and the takeaways from those distributions are outlined within the Jupyter notebook shared.

Since the number of instances were very few (200), a simple linear model (logistic regression) is tried to predict if a patient will be readmitted or not within 30 days. This model seemed to perform better than a dummy model, which makes a decision by random sampling using the prior distributions that is learnt from the dataset. The metrics used for comparing the models were ROC AUC, average precision and log loss.

Since the number of instances is very low, cross-validation is used to try to evaluate the model. A separate test set is not used to measure the generalization error because of the limited amount of data. After the model was trained trying to maximize the average precision, a threshold is chosen for the model by using the cross-validation dataset and trying to optimize balanced accuracy. Then the F1-score and confusion matrix are reported using the cross-validation approach. The reason why F1-score was not used for choosing the threshold was that the model tries to predict everything as positive if F1-score will be chosen to be maximized.

The problem with logistic regression is underfitting but in order to use more complex models with a high number of parameters, a larger dataset will definitely be more useful. The model does not seem to perform very well and more work should be done to improve this model like thinking about new features and trying to obtain more data while going for complex models. The most useful feature based on permutation importance seems to be the age feature. Different feature importance techniques can be looked at as well to see the most influential features like Shapley values or looking at the coefficients of the logistic regression model.

NER

The stanza library from Stanford University is used to try to get entities like problem, treatment and test performed on the patient. I did not have the time to try different models or perform any fine tuning by labelling the discharge notes myself. So an out of the box model is used because of time limitations. Based on some discharge notes, the entities returned by the model seem reasonable like antibiotics - treatment, pneumonia - problem, follow up scan - test.

But using an out of box model without any fine tuning and proper evaluation is definitely dangerous to use in a clinical setting. These general purpose models do not perform well on specific tasks and the returned entities do not always make sense like complications - problem.