Take Home Exercise

Note

1. For all questions that need coding, prepare both code and answer.
2. Pick a programming language of your choice to answer the questions (We have SAS, R, python, SQL on site)
3. You are expected to answer all the questions independently
4. Follow-up questions given during on-site interview will be based on the level of theoretical and practical competence revealed from this take home exercise.
5. CONFIDENTIAL. PLEASE DO NOT DISTRIBUTE.

Dataset Description

Target variable:

- Readmitted (True/False)

Independent Variables:

- Demographic Info (age, race, gender)
- Admission and Discharge Type
- Time in hospital
- Specialty
- Procedure and drug Information (number of procedures/drugs billed)
- Diagnosis Information (Three ICD9 codes provided, diag_1, diag_2, diag_3)
- Lab results

Task

Predict the likelihood of readmission to hospital in future given diabetes patient's last hospital visit history

Questions

1. Data Processing
    a. Remove the records where discharge_disposition_id = "Expired", which means patient passed away during last visit.

2. EDA (Exploratory Data Analysis)
   Run EDA on the dataset. Your analysis should include, but not limited to:
    a. Frequency analysis for class variables
    b. Summary statistics and missing value analysis on numeric variables

    c. Document key findings, data manipulation, transformation and variable creation recommendations
3. Create a training set and a test set while training set has 75% of the records
4. Pick a few variables you are comfortable with, perform transformations that you feel necessary, and fit a logistic regression model (your model should include both categorical and numeric variables).
5. (Optional) Build another model using a different approach.
6. (Optional) Evaluate how each model performs.
7. Discuss your final model selection and how you think this model can be further improved.