Data science take-home assignment

Overview

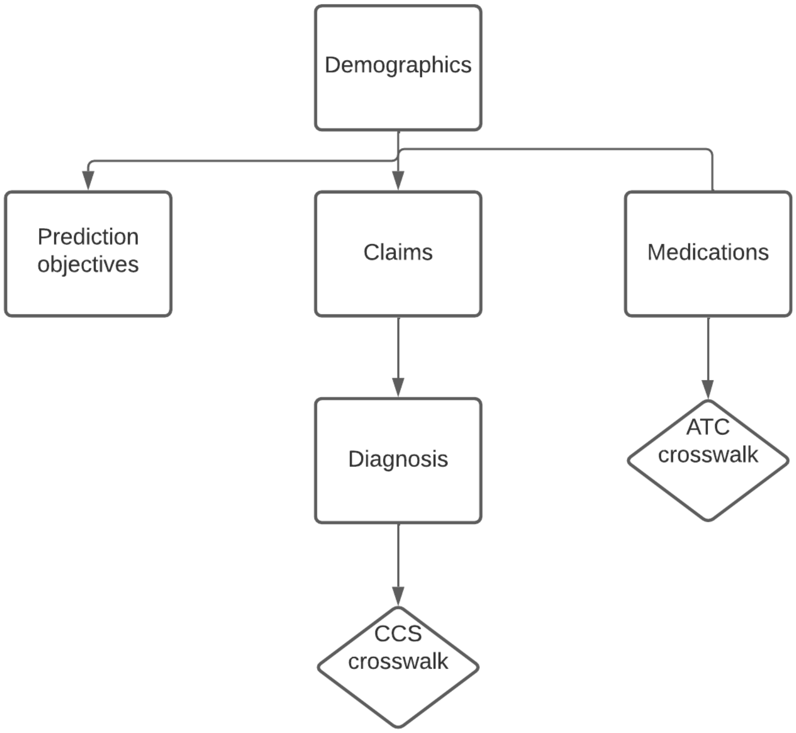
The objective of this assignment is to create a model that predicts the likelihood that a patient will have a hospital admission in the next 180 days. This assignment gives the candidate the opportunity to demonstrate their data wrangling, feature engineering and model building capabilities.

It is strongly recommended that the output of this assignment be a well structured jupyter notebook. Pandas and Sklearn are recommended, but not required. If you want to prototype in Matlab, R or Excel, feel free to do so. If that is the case, a transformed jupyter notebook or source code plus write-up can help us post-process the results.

The minimum model performance metrics that must be reported are the model's ROC AUC and precision-recall curve AUC.

Data Dictionary

All patient data has been de-identified and event dates have been replaced with *days until prediction*.



File	Description
patient_data/prediction_objective.csv	Data on whether the patient had an admission in the next 180 days.
patient_data/demographics.csv	Basic demographic information on the patient.
patient_data/medications.csv	Medication prescriptions.
patient_data/claims.csv	Medical billing records.
patient_data/diagnosis.csv	Diagnosis codes associated with billing records.
reference_data/atc_crosswalk.csv	Classification system for grouping medications.
reference_data/ccs_crosswalk.csv	Classification system for grouping diagnosis and procedure codes.

prediction_objective.csv

Field	Description

PATIENT_ID	The patient’s ID.
HAS_ADMISSION	Whether the patient had an admission in the next 180 days. 0 = No admission, 1 = had an admission.

demographics.csv

Field	Description
PATIENT_ID	The patient’s ID.
SEX	The patient’s sex. have to 1-hot encode
AGE	The patient’s age at time of prediction.

medications.csv

Field	Description
PATIENT_ID	The patient’s ID.
DATESTART	The prescription’s start date **relative to the date of prediction**. For example 5 would be 5 days until the prediction date.
MEDICATION_NAME	Name of the medication. not needed bc ATC_code encompasses it
NDC_CODE	National drug code. → ATC_code
DOSAGE	Dosage of the medication. too many NaNs
DISPENSING_QUANTITY	Quantity of the medication. heavily right skewed
DAYS_SUPPLY	Number of days this medication is supposed to be taken over. heavily right skewed
ROUTE	The route the medication is taken. too many NaNs
STRENGTH	Strength of medication. too many NaNs

claims.csv

Field	Description
PATIENT_ID	The patient’s ID.
CLAIM_ID	The claim’s ID.
ADMISSION_DATE	Start date of the claim **relative to the prediction date**. For example 5 would be 5 days until the prediction date.
DISCHARGE_DATE	End date of the claim relative to the prediction date.
DRG_CODE	Diagnosis related group code. too many NaNs
REVENUE_CODE	The claim’s revenue code.
CPT_CODE	The claims procedure code.
PLACE_OF_SERVICE	The medical setting of the claim. For example inpatient hospital or urgent care facility.

diagnosis.csv

Field	Description
CLAIM_ID	ID of the claim the diagnosis is associated with.
PRIORITY	Priority of the diagnosis within the claim. 1 is the highest priority.
CODE_TYPE	The type of diagnosis code.
CODE	The diagnosis code.
CODE_DESCRIPTION	Description of the diagnosis code.

atc_crosswalk.csv

Reference table that maps a medication’s NDC code to an Anatomical Therapeutic Chemical (ATC) code. The mapping is used to organize ~750,000 unique NDC codes into 2,736 categories. The mapping is many to many, i.e. a single NDC code can map to multiple ATC codes.

Field	Description

NDC	National drug code.
ATC	Anatomical Therapeutic Chemical code.
ATC_LABEL	Label for the ATC code.

ccs_crosswalk.csv

Reference table that maps diagnosis codes to a Clinical Classification Software (CCS) code. The mapping is used to organize ~ 35,645 unique diagnosis codes into 280 categories. The mapping is one to many, i.e. a single diagnosis code only maps to one CCS code.

Field	Description
diag_code	The diagnosis code.
diag_code_type	The type of diagnosis code.
ccs_code	The CCS code.
ccs_code_type	The type of CCS code.
label	The CCS code’s label.