Developing tree-based prediction model using the CMS 2008-2010 Data Entrepreneurs' Synthetic Public Use File (DE-SynPUF)

Michael Seo

March 23, 2022

Introduction

- ✓ Although the Medical DE-SynPUF has very **limited inferential research value** to draw conclusions about Medicare beneficiaries due to the synthetic processes used to create the file, DE-SynPUF does increase access to a **realistic Medicare claims** data file in a timely and less expensive manner to spur the innovation necessary to achieve the goals of better care for beneficiaries and improve the health of the population.

Aim

- ✓ Given medical data from 2008 to 2010, our aim is to use first 30 months of data (Jan 01, 2008 to June 30, 2010) to predict the last 6 months claim payment amount (July 1, 2010 to Dec 31 2010).
- ✓ We will focus on outpatient claims for illustration.

Summary of variables

Beneficiary summary file contains:

- ✓ Patient characteristics such as date of birth/death, sex, race, and state/county
- ✓ Total number of months of (insurance) coverage for the beneficiary
- ✓ Chronic condition indicator (i.e. heart failure, chronic kidney disease)
- ✓ Outpatient annual medicare reimbursement amount, beneficiary responsibility amount, primary payer reimbursement amount

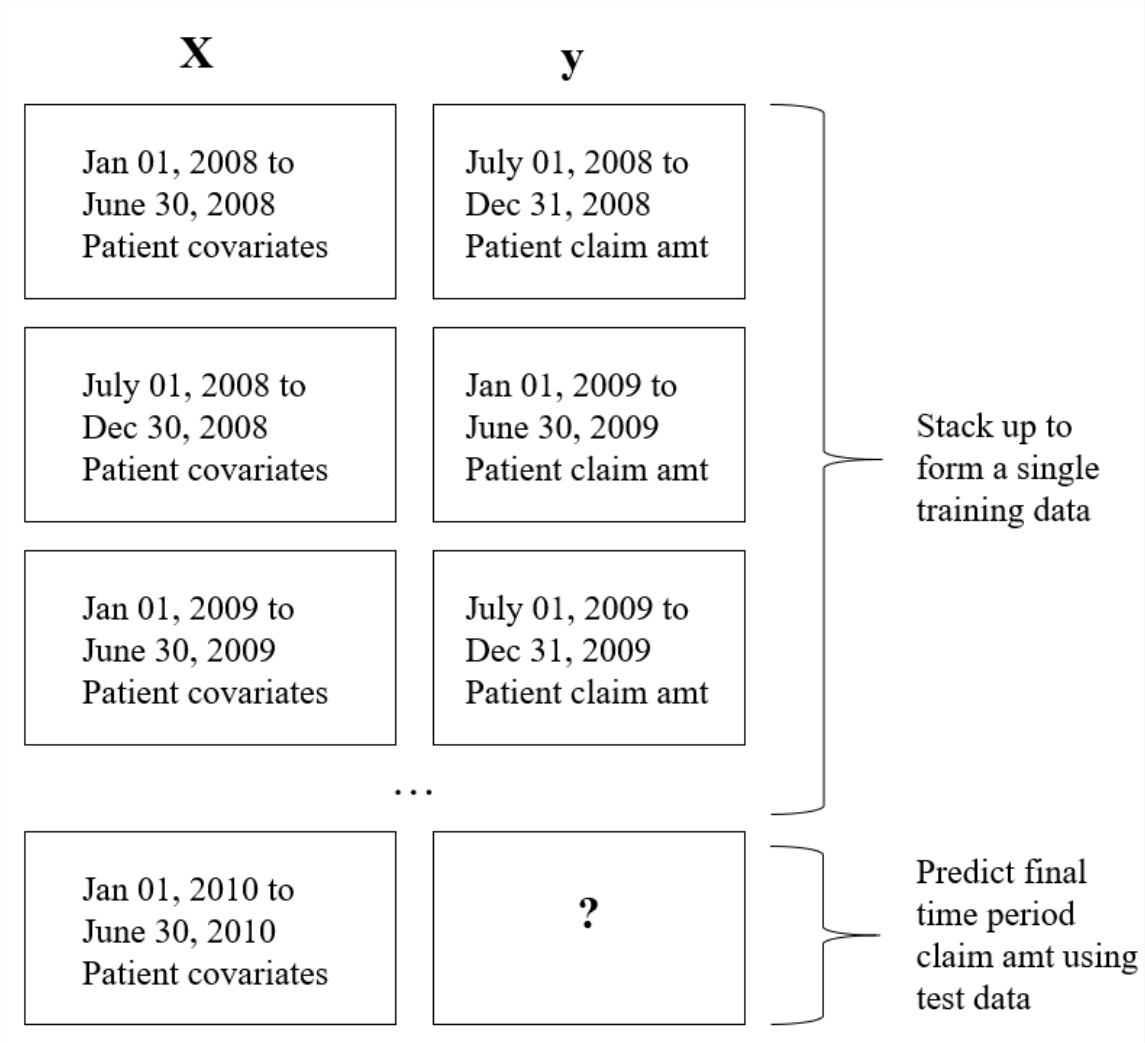
Summary of variables II

Outpatient claims file contains:

- ✓ Claim start/end date, **Claim payment amount**
- ✓ Other variables: Provider institute, Attending/Operating physician, Claim diagnosis code, Claim procedure code

Model training

Idea is to divide the data into different time frames.



Time-varying and time-invariant variables

- ✓ We can derive time-varying variables that might be useful for prediction purposes, such as number of visits within time period, total number of visits thus far, last visit from period end, and total payment in the given time period.
- ✓ This is probably where feature engineering begins.

Predicting claim payment amount

- ✓ We see that many patients have 0 claim payment amount.
- ✓ We divide modelling into two parts and fit tree-based gradient boosting algorithm (lightgbm in python) separately for each part.
- ✓ First model, we develop a model where we predict whether a patient would file a claim in the next period (binary outcome).
- ✓ Second model, we develop a model where we predict each patient claim amount (continuous outcome).
- ✓ Final predicted claim amount would be the product of the two: Pr(patient files a claim) x Expected claim amount for each patient.

Model result

Although there is limited inferential research value for this project, with real claims data we would be able to find important predictors.