Project Notes

**Project Notes**

**Main Questions**

* What **useful** insights or knowledge can I derive from this dataset?
  + What problems would be good to solve?
  + What kinds of information does this dataset provide (in general)
* What information and variables can I get rid of?
  + How can I consolidate variables to preserve meaning, while decreasing less relevant clutter?

**Initial Stages of a Data Science Project**

Goals:

1. Understand the data
   1. What does a single row represent?
   2. Are there natural groupings of observations?
      1. What do these groupings of data tell us?
      2. How can we use these groupings to better understand the variables in the dataset?
   3. How can I simplify my dataset and create a dataset that only contains useful information?
      1. We want the dataset to be manageable
      2. TIDY DATASET
   4. What are the primary keys of the dataset?
      1. What does the smallest level of granularity represent?
   5. Exploratory data analysis
      1. Check distributions of variables
      2. Examine extremes

Data Proc/understanding

**Data Exploration/Understanding (Inpatient dataset)**

* 15 years of data
  + Contains every claim of every patient
    - Each claim will have:
      * a specific set of DRG codes (Diagnosis, treatment and resources)
      * Length of stay information
        + Made up of multiple variables, so actual LOS will have to be derived

Potential variables:

Claim admission date (CLM\_ADMSN\_DT)

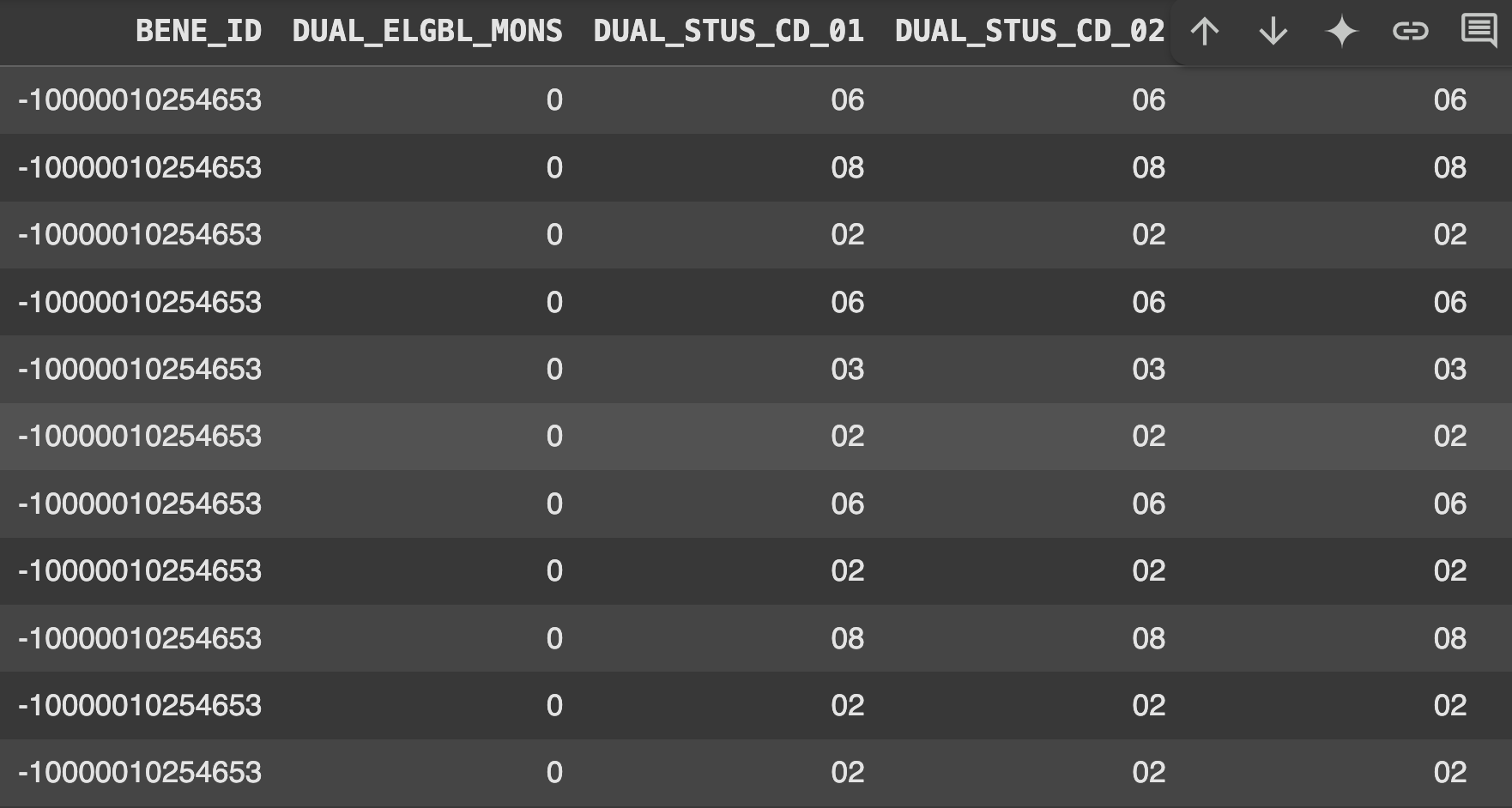
* + - * Amount of money claimed, paid out … (there are a ton of variables on this)
        + To get the total amount of money paid by medicare for a claim:

**multiply the CLM\_PASS\_THRU\_PER\_DIEM\_AMT by the CLM\_UTLZTN\_DAY\_CNT and then add to CLM\_PMT\_AMT**

Process obtained from the codebook for the dataset

* + - * Type of claim (V = Part A institutional claim record (inpatient [IP], skilled nursing facility [SNF], hospice [HOS], or home health agency [HHA])
        + This is the value for every entry in the inpatient dataset
      * Claim line number, one number for each service or procedure rendered to the beneficiary
        + HCPCS\_CD is a column of codes for the procedure or service that was performed
* What info do we want?
  + LOS
  + Date/Time
  + Financial
  + Diagnosis + treatment
    - ICD and DRG information (codes + descriptions)
  + Demographic information

**Data Exploration/Understanding (Enrollment dataset)**

* Dual Status Codes
  + Indicate whether bene was eligible for both medicare and medicaid **in a given month**
    - Consolidate: number of months for a given year
      * **Already exists**: but the number of eligible months doesn’t make sense??
        + Clearly this individual has dual status for multiple months, but the dual\_elgbl\_mons variable just says 0 for all of them?
        + Every single entry has Dual\_elgbl\_mons == 0
* Cost Sharing group
  + Not really interested in which specific cost sharing group individual is in right now
    - Dropping specific groups
* HMO indicators
  + All null values
* MDCR\_ENTLMT\_BUYIN\_IND
  + All repeated values, except for the 2025 for which there are a bunch of null values (because it is 2025
  + Totally consistent in all other years
    - Evidence suggests that it’s ok to assume that the value will stay the same across all months for 2025 as well
    - Create single indicator var for this variable
* MDCR\_STATUS\_CODE\_{i}
  + 14.4% of bene\_id/year combos have status codes that change
  + ALL changes are because people turned 65
  + So MDCR\_status\_code\_flag is an indicator of whether or not someone turned 65
    - So you really only need one indicator column for MDCR\_status code per year, and then an indicator for whether or not it changed that year
    - Can best obtain this by using age at end of ref year variable
      * Can also use birthday to get month of change if you want
* PTC\_CNTRCT\_ID{i}
  + Don’t care about this
    - Contract number of beneficiary’s medicare advantage plan (part C)
* PTC\_PBP\_ID{i{
  + Same reason as above
    - Three digit alphanumeric part C benefit package number
* PTC\_PLAN\_TYPE
  + All null values
* PTD\_CNTRCT\_ID
  + Dropping for same reasons as part C contract ID
* PTC\_PBP\_ID{i}
  + Dropping for same reasons as part C PBP id
* PTD\_SGMT\_ID
  + Only null or 0.0 values
    - Doesn’t really provide any information so dropping
* RDS\_IND
  + Drug plan by employer
  + Dropping

Potential Directions

Machine Learning

* Classification of high or low LOS
* Prediction of readmissions
* ~~Classification of severity (could look at the nyu algorithm)~~
* **Predicting LOS**

NYU Algorithm

<https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2813806>

* Concordance in Medical urgency classification of discharge diagnoses and reasons for visit
* Basically it’s really hard to accurately predict level of urgency required for people that show up to the ED for a lot of reasons
  + So maybe trying to predict that wouldn’t be a great idea
* Survey Weighted logistic regression analysis
* Could try to predict final severity of “problems” and **amount/quantity of resources required to treat a patient**?

**Classification Ideas**

* Classify individuals as high or low risk of near-future disease progression
  + Would have to watch out for patients that left the dataset
* Classify severity of disease

**Prediction Ideas**

* Predict probability of readmission
* Predict **LOS/Patient day ratio**
  + Could categorize individuals as a High Cost Patient (HCP) or not, and then run different models based on that
    - Hierarchical model (maybe talk to leontine about implementation??)
* Predict **cost/utilization**
  + Predict “High Cost Patients”
* Or just **identification of trends of costs?**
  + Forecasting resource utilization, reasons for utilization
    - Could more accurately identify times of unreasonably high utilization
* Predicting cost and use predictions to identify disparities
  + Between geographical regions, socio-demographic factors
  + Temporal variations
    - Costs of healthcare increase over time
* Predict severity of disease based on presenting symptoms?
  + Related to predicting LOS/Cost-resource utilization
* Predicton consideratiions
  + No matter what outcome is being predicted, the data is temporal, and so there should be some kind of temporal aspect factored in
    - Look into how to factor in time (day of the week, holidays, yearly trends … etc.)

**Inference Ideas**

* Are people who receive certain DRG’s more likely to be hospitalized than others?

**Variable Considerations**

* Definitely include demographic information

Grouping diagnoses / **simplifying datasets** for analysis

* CCSR groupings?
* Create new tables with subsections of columns
  + Create a patient journey table?
* Truncation of ICD codes
  + Won’t have a lookup for these values though
    - Unless you could truncate the icd descriptions as well

Look at top ICD’s, DRG’s, … etc

* Group in terms of severity?
  + How to measure severity?
  + Could use cost as an indicator/proxy for quantity of resources
  + Only have access to inpatient records
    - This means they were admitted to the hospital
    - So don’t have the “denominator” of all ED visits to work with
  + Very little evidence that discharge diagnosis (final diagnosis) is associated with the presenting reasons for the visit (at the time of arrival) and the need for ED care
    - So basically trying to predict final diagnosis is really really hard, and doesn’t work very well (not enough information upon patient arrival)
      * Probably also means that predicting final cost/resources also wouldn’t work (based solely on presenting conditions/initial diagnosis)

Data\_dict

Inpatient\_encounters.csv:

* Processed inpatient file
* # Column Non-Null Count Dtype
* --- ------ -------------- -----
* 0 BENE\_ID 20867 non-null object
* 1 CLM\_ID 20867 non-null object
* 2 CLM\_FROM\_DT 20867 non-null datetime64[ns]
* 3 CLM\_THRU\_DT 20867 non-null datetime64[ns]
* 4 YR 20867 non-null int32
* 5 LOS 20867 non-null int64
* 6 DRG 20867 non-null object
* 7 PRNCPAL\_DGNS\_CD 20867 non-null object
* 8 PTNT\_DSCHRG\_STUS\_CD 20867 non-null object
* 9 CLM\_IP\_ADMSN\_TYPE\_CD 20867 non-null object
* 10 ER\_flag 20867 non-null int64
* 11 CLM\_TOT\_CHRG\_AMT 20867 non-null float64
* 12 CLM\_PMT\_AMT 20867 non-null float64
* 13 NUM\_DIAG 20867 non-null int64
* 14 ICD\_Description 17098 non-null object
* 15 DRG\_TITLE 17952 non-null object

**Notes:**

* Icd description is the description of the principal ICD code
* All claims were fully paid for by insurance (there are no rejected claims in the synthetic dataset)

Diagoses:

* Data columns (total 4 columns):
* # Column Non-Null Count Dtype
* --- ------ -------------- -----
* 0 BENE\_ID 176690 non-null object
* 1 YR 176690 non-null int32
* 2 ICD\_DIAG\_CD 176690 non-null object
* 3 ICD\_Description 163357 non-null object

Enrollment:

* RangeIndex: 86917 entries, 0 to 86916
* Data columns (total 44 columns):
* # Column Non-Null Count Dtype
* --- ------ -------------- -----
* 0 **BENE\_ID 86917 non-null object**
* **1 STATE\_CODE 86917 non-null int64**
* **2 COUNTY\_CD 86917 non-null int64**
* **3 ZIP\_CD**  86917 non-null int64
* 4 **BENE\_BIRTH\_DT** 86917 non-null datetime64[ns]
* 5 **SEX\_IDENT\_CD 86917 non-null int64**
* **6 BENE\_RACE\_CD**  86917 non-null int64
* 7 ENTLMT\_RSN\_ORIG 86917 non-null int64
* 8 ENTLMT\_RSN\_CURR 86917 non-null int64
* 9 **ESRD\_IND** 86917 non-null object
* 10 BENE\_PTA\_TRMNTN\_CD 86917 non-null int64
* 11 BENE\_PTB\_TRMNTN\_CD 86917 non-null int64
* 12 **BENE\_DEATH\_DT** 0 non-null datetime64[ns]
* 13 **BENE\_ENROLLMT\_REF\_YR** 86917 non-null int64
* 14 BENE\_HI\_CVRAGE\_TOT\_MONS 86917 non-null int64
* 15 BENE\_SMI\_CVRAGE\_TOT\_MONS 86917 non-null int64
* 16 BENE\_STATE\_BUYIN\_TOT\_MONS 86917 non-null int64
* 17 BENE\_HMO\_CVRAGE\_TOT\_MONS 86917 non-null int64
* 18 RDS\_CVRG\_MONS 86917 non-null int64
* 19 ENRL\_SRC 86917 non-null object
* 20 SAMPLE\_GROUP 0 non-null float64
* 21 ENHANCED\_FIVE\_PERCENT\_FLAG 86917 non-null object
* 22 CRNT\_BIC\_CD 86917 non-null object
* 23 AGE\_AT\_END\_REF\_YR 86917 non-null int64
* 24 COVSTART 86917 non-null object
* 25 DUAL\_ELGBL\_MONS 86917 non-null int64
* 26 STATE\_CNTY\_FIPS\_CD\_01 74001 non-null float64
* 27 STATE\_CNTY\_FIPS\_CD\_02 74001 non-null float64
* 28 STATE\_CNTY\_FIPS\_CD\_03 74001 non-null float64
* 29 STATE\_CNTY\_FIPS\_CD\_04 65477 non-null float64
* 30 STATE\_CNTY\_FIPS\_CD\_05 65477 non-null float64
* 31 STATE\_CNTY\_FIPS\_CD\_06 65477 non-null float64
* 32 STATE\_CNTY\_FIPS\_CD\_07 65477 non-null float64
* 33 STATE\_CNTY\_FIPS\_CD\_08 65477 non-null float64
* 34 STATE\_CNTY\_FIPS\_CD\_09 65477 non-null float64
* 35 STATE\_CNTY\_FIPS\_CD\_10 65477 non-null float64
* 36 STATE\_CNTY\_FIPS\_CD\_11 65477 non-null float64
* 37 STATE\_CNTY\_FIPS\_CD\_12 65477 non-null float64
* 38 VALID\_DEATH\_DT\_SW 0 non-null float64
* 39 **RTI\_RACE\_CD** 86917 non-null int64
* 40 MDCR\_status\_code 86917 non-null int64
* 41 PTD\_PLAN\_CVRG\_MONS 86917 non-null int64
* 42 Year 86917 non-null int64
* 43 MDCR\_status\_change\_flag 86917 non-null int64

Model Ideas

Model Ideas

**Hierarchical**

* Could do hierarchical by demographic variables
  + For both LOS and high cost prediction

**LLM**

* Could use an LLM to classify ICD descriptions as severe vs not severe?
  + Unlikely

**Use data from 2015 - 2021 to predict LOS for patients in 2022**

* This would require me to use only patients for which there was data from 2015-2021
* More insurance based

Ideally we’d want to be able to use someone’s presenting symptoms and demographics to predict LOS without any prior information

* More of a hospital based approach

**We have no info on presenting symptoms besides principal diagnosis code**

* All we really have to evaluate the patient when they first show up to the hospital is principal diagnosis code, ER\_flag, demographic data, date of admission,
  + Maybe have number of diagnoses
* Unless we use their data from past years

**Could do a two part model to predict an individual's specific LOS or severity**

* Part 1 is trained on all the data except for that individual
  + This would take into account demographic and medical info
* Part 2 is trained on only the individual’s data
  + This would take into account person specific medical info (no demographic in this part)
* Then aggregate the results together

Add in temporal features somehow?

* Using claim from and through dates?

**Variables to use for baseline mod (no hierarchy, no two part):**

* Age
* State (maybe transition to county CD)
* Sex
* Race
* ESRD indicator
* Principal diagnosis code
* ER flag
* Inpatient admission type (code)
* Total RX cost of the previous year(s)?
* Total inpatient cost of previous year(s)?
* Could try to use an area deprivation index scale
  + PCA of demographic variables
  + <https://www.census.gov/topics/income-poverty/poverty/about/related-sites/rates.html>
* **Number of co-morbidities** (derived from previous years)
* **Previous surgery y/n**

**Train Test Dataset**

* **If 2023 is a full year’s worth of data use below**
* Available = 2015 - 2023
  + Train = 2015 - 2021
  + Validation = 2022
  + Test = 2023
* **If 2023 is not a full year use below**
* Available = 2015 - 2022
  + Train = 2015 - 2020
  + Validation = 2021
  + Test = 2022

**Ideas from Stats 630**

* Generative models for classification
  + Non-normal class assumptions
  + Fuzzy classes
  + Unknown number of classes

**Regression based models**

1. Random forest i
   1. Unscaled data
   2. Removed principal diagnosis code, and replaced with a “Severe indicator”
      1. Severe indicator: derived by filter diagnosis codes that had any LOS greater than 10 days
   3. One hot encoded all other categorical variables
      1. State code
      2. Race indicator
      3. Admission type
   4. Removed year from training sets
      1. Assumes that the observations are independent of each other year to year
         1. This is not true, as observations from the same individual are clearly dependent (maybe?? Need to think about this more)
         2. Ignoring this fact for this first model
   5. Removed county code from training sets
   6. Trained model according to TRAINING SCHEME 1 (year pairs)
2. RF ii
   1. RF i but model trained according to TRAINING SCHEME 2 (all years)
3. XGBoost

**Hierarchical Models**

* Pipeline idea

1. Classify observations as prolonged LOS or not prolonged LOS
2. Assign flag based on classification
3. Separate into two different datasets
4. Build separate models for each dataset

**Model Training Schemes**

1. Train models on ONLY the previous year and test models on the following year
   1. I.e. train on 2016, and test on 2017
2. Train models on ALL previous years and test on following year
   1. Would require an aggregation of some variables
      1. Number of claims
      2. Charge and payment amounts
         1. Average?
      3. ~~Number of diagnoses~~
         1. Stratification?
         2. You want model to “see” that for example, 5 years ago, someone had 45 different principal diagnoses, but summing over all the years would give people not present in the database 5 years ago an unfairly low # of diagnoses
            1. You’d want a metric that IS extremely sensitive to outliers
      4. RX payment amounts
         1. Average?
   2. Could **average** or sum over these variables depending on context
   3. Not all beneficiaries would be present in every year
      1. This would lead to some beneficiaries having artificially higher total payment amounts than others
      2. Should probably average
         1. An average wouldn’t reflect any trends in total cost, so taking an average over the number of years for which every beneficiary is present
            1. I.e. looking at years 2016-2022, and one observation is only present in 2021, then you’d only take an average during that year?

Pipeline Plan

**Pipeline Plan**

* Considerations

1. No missing data, that was taken care of in the data pre-processing
2. Variable Encoding (for regression models)
   1. Target encoding
      1. Take values of the categorical variable, find the mean of the target variable for each category, and replace the categorical values with that number (the mean of the target value per category)
   2. One hot encoding
   3. Other encoding techniques … ??
3. Data scaling
   1. Standard scaling
   2. Min-max scaling
   3. Robust scaling
4. Variable importance
   1. SHAP
   2. PCA
      1. PCR uses PCA and combines it with linear regression (so could be a technique used)
5. Data partitioning
   1. Train set : 70%, Test set : 30%
   2. Train set : 80%, Test set : 20%
   3. Train set: 90%, Test set : 10%
   4. Create a column for each of these partition schemes and evaluate results
6. Model Training
   1. Write function to train model and put results into dictionary
7. Model Evaluation
   1. For regression:
      1. Root Mean Squared Logarithmic Error (RMSLE)
   2. Variable Importance:
      1. Permutation Importance
      2. Partial Dependence

Meeting Notes

**Meeting notes 2/4/2025**

* Edits and comments in colab script
* Think about ways of creating comorbidities from diagnoses
  + Number of
  + Creating a system of taking diagnosis codes and turning them into general ideas of disease classification
    - For an entire person over a period of time
    - Could use one DRG as a proxy for a single comorbidity
* Enrollment data
  + Member beneficiary tables
    - Process these for **thursday 2/6/2025**
* How big of a sample size should we require for testing hypotheses
  + Depends on whether or not it can be backed up with a literature review
* Synthetic data qualities
  + Would be interesting to see whether or not data holds up to real world results
* Ai lit review
  + <https://consensus.app/>
  + <https://www.scholarcy.com/>
  + <https://eds.p.ebscohost.com/eds/search/advanced?vid=0&sid=50a6bea3-2d59-498c-ade0-c3692b24e2de%40redis>
    - UMASS library

**Meeting notes 2/11/2025**

How do we analyze the data?

* Putting the data into long format

Grouping diseases / diagnoses

* DRG: designed to group like inpatient stays
  + ICD codes are included in this
  + As well as procedures, medications, length of stay … etc.
  + More comprehensive than a single ICD
* ICD code truncation

Member months

* One row for every month the member is enrolled

Dual status enrollment: could be used as an indicator of socioeconomic status

How do you know what your research question is?

* What are the top diagnoses in the file
* Top DRGs
  + Comparing groups of patients
* Could compare outcomes before and after a policy change
* Geographic variations
  + In cost
  + Lengths of stay for similar diseases
  + Occurrences of specific diseases

Machine Learning

* Given a certain condition, group into high, low LOS
  + Classification problem
    - What factors predict high or low LOS the best

**Meeting notes 02/18**

* Watch Drug database video on yummy data science
  + Drugs themselves are meaningless, but number of drugs could be useful
* Could build models based on different “stages of stay”
  + Present on admission, new diagnosis during stay, drg given retrospectively
    - Drg is “almost always” assigned retrospectively
    - Drg is supposed to be for billing purposes
* Prediction of resource needs for next year
  + LOS is a big part of this
  + Patient day ratio is equivalent to the LOS in this dataset
    - To reduce patient day ratio, reduce admissions or reduce average LOS

**Meeting notes 02/25**

* Conditions from past years can be used as discriminatory qualities from cohort analyses
  + Which makes data from past years important
  + In order to compare current health statuses, you need to know whether or not both individuals came from the same health background
    - I.E., comparing people with advanced stage cancer, you shouldn’t compare someone who already had cancer, with someone who hasn’t had cancer
      * The person who has had cancer will obviously be more likely to transition to a higher stage of cancer
  + The assumption here is that once you reach a high risk of cancer or of some disease, you will remain at that risk level until you die
    - This is a big assumption, and might not be true
      * Could go from pre-diabetic, to diabetic, and then back to not even pre-diabetic
* Ben’s project:
  + Looking at diabetes (and different stages of them)
    - Doing survival analysis?
  + Also including liver disease (s)
    - Which ones?
  + Initial goal was to construct diabetes cohort and follow their transition to kidney failure (from diabetes stage to stage then eventually to kidney failure)
    - Three classifications of diabetes
    - Outcome of nx3 table with 0’s and 1’s for specific diabetes stages
      * With the specific times for each of the above 0’s and 1’s
    - Can’t create cohort to only include people who have had prediabetes at some point because some people won’t ever be diagnosed with pre-diabetes
      * People who collapse (for example) and are diagnosed with diabetes
* Margaret’s Project
  + Firearm injury
  + Causal inference
    - Occurrences and costs change after implementation of policies?
  + How to define cost
    - What medicare was charged
    - Out of pocket charges (to patient)
    - Cost to facility vs provider vs for drugs
* My project
  + Hurdle model
  + Depends on the factors used in the models
  + We know from the past that the DRG is a good indicator of resource utilization
    - Summary dataset of costs aggregated by DRG
      * This can be derived from the population
      * The DRG does the job of the model
      * Second model would be only demographic
        + Could do inference
    - With this, can use a mapping of DRG → diagnosis to map to the individual diagnosis
      * Then can use the rest of predictors (not DRG)
  + Limit scope of project to data contained in one dataset
  + Could treat each obs as independent from all others in the dataset
    - Regress cost on principal diagnosis
  + Start with total cost model and throw everything into it
  + Could try geographic hierarchical model
  + Need to be really careful about double dipping the data
  + **Predicting readmissions?**
* Claim line
  + What information would you be losing by filtering to only include claim lines == 1?
* Could try to figure out payments are actually deconstructed

**Meeting Notes 03/04/2025**

Questions:

* Data is broken up into years, what would my model be missing by ignoring the year?
  + How could I incorporate the temporal aspect into a model not initially trained on year
  + How could I incorporate the temporal aspect initially?
    - Train models based on all past years (2015-2022) and validate with data from 2023?
    - Train models based on one year (t) to predict the following year (t+1)?
    - <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0279540>
      * Trains on 2016 features with 2017 outcomes, validates on 2017 features and 2018 outcomes

Ben’s project

* Multistage modelling
  + Kind of survival analysis?
  + Markov model meets survival analysis I think
* Resampling time dependent observations
  + Divide data into blocks (3 yr blocks for ex.)
    - Overlapping blocks
  + 2018-2020, 2019-2021
* Simplicity
  + Just start by ignoring year
* Kaiser didn’t go back more than 6 years
  + Last 3 years and last 6 years (with a greater weight on the most recent 3)
  + Score model for every year in the past
    - Because then model can capture prediction accuracy regardless of year
  + But main model presented to hospital administrators is model trained on all data

Tracker

Last updated Feb 25, 2025

Project Tracker

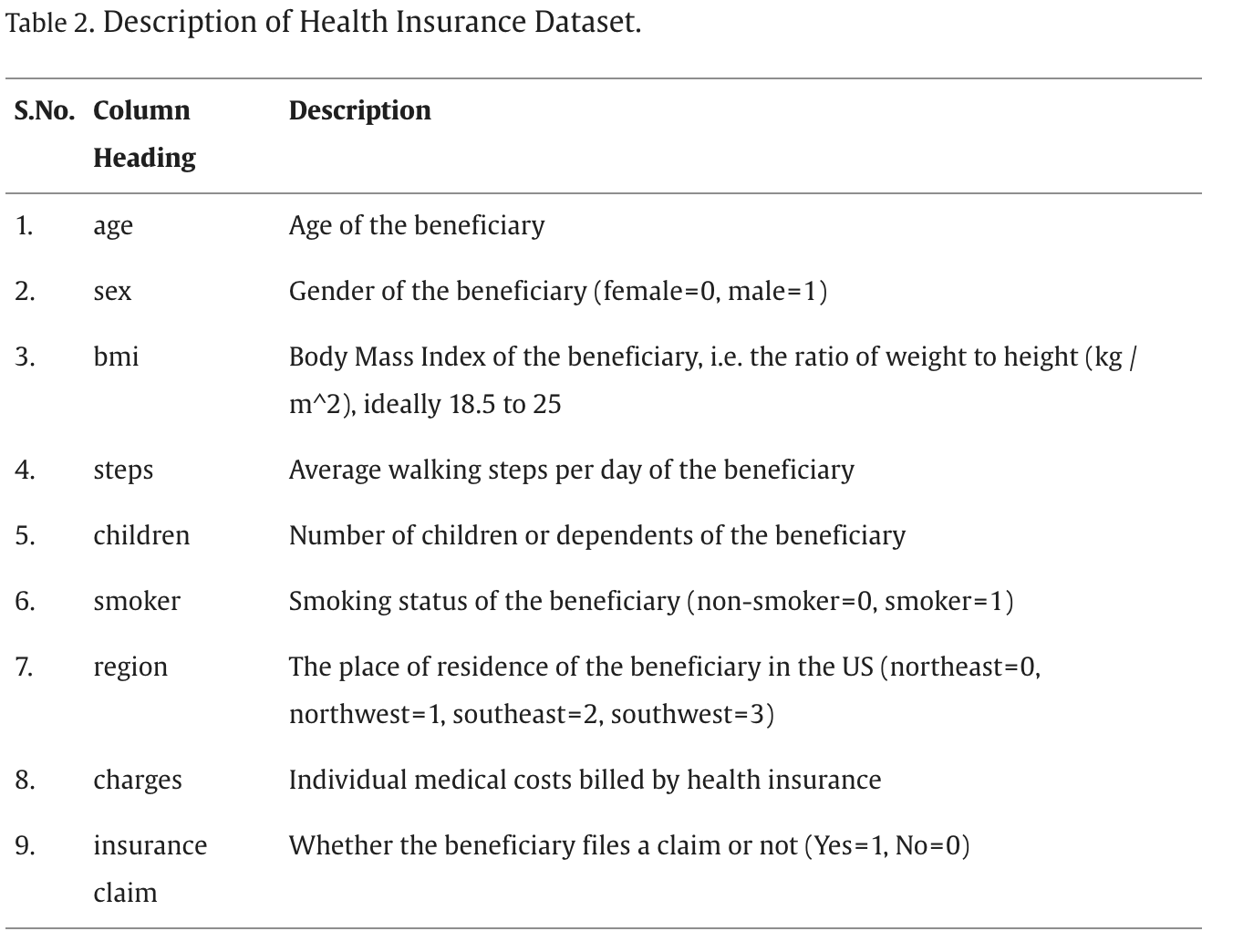
# Project Name

| ID | Task | Impact | “Due” Date | Status |
| --- | --- | --- | --- | --- |
| [01](?tab=t.nw2t34qdgpa) | Literature Review | Medium | Feb 28, 2025 | In Progress |
| [02](?tab=t.s1zrxb7xd556) | Data Pre-Processing | Medium | Feb 27, 2025 | Finished |
| [03](?tab=t.86rny357c0i) | Preliminary Analytics | Low | Mar 4, 2025 | In Progress |
| [0](?tab=t.86rny357c0i)4 | Model Testing | High | Mar 7, 2025 | In Progress |
| 05 | High powered computing access | High | Feb 27, 2025 | In Progress |
| 06 | Watch cohort analysis ymy DS | Low | Feb 26, 2025 | Finished |
| 07 | Add code to git Repo | Low | Date | Finished |
| 08 | Develop question enough for an abstract | Low | Date | Finished |

Literature Review

Literature Review

<https://www.sciencedirect.com/science/article/pii/S2667096821000057#sec0020>



| **Resource** | **Status/notes** |
| --- | --- |
| Application of machine learning and data visualization techniques for decision support in the insurance sector   * <https://www.sciencedirect.com/science/article/pii/S2667096821000057#sec0026> | Need to read fully  So far: Could give some insights as to potential directions, Used above variables to predict whether or not beneficiary had filed a claim (one example) |
| Predicting surgical department occupancy and patient length of stay in a paediatric hospital setting using machine learning: a pilot study   * <https://informatics.bmj.com/content/29/1/e100498> | Goal of predicting next day census. Hierarchical methods used to incorporate temporal features, predict admissions and discharges to the surgical unit. Used RF’s, nlp to analyze doctor orders, and glm to predict discharges based on prob classifications of RF for individual level data, accuracy was roughly 2-3 for admission and discharge change, while the median admission and discharge was 12 for both categories each day |
| Application of Artificial Intelligence And Big Data for Fighting COVID-19 Pandemic.   * <https://link.springer.com/chapter/10.1007/978-3-030-87019-5_1> |  |
| Predicting the length-of-stay of pediatric patients using machine learning algorithms   * <https://www.tandfonline.com/doi/epdf/10.1080/00207543.2023.2235029?needAccess=true> | Good baseline, pretty simple models, pretty bad accuracy (accurate +- 4 days or so), could definitely improve on this |
| The application of machine learning to predict high-cost patients: A performance-comparison of different models using healthcare claims data  * <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0279540> | Dataset used: - demographics, care dependency, disease management program participation, out and inpatient diagnosis, data on prescription drugs, costs of each in and outpatient treatment + prescriptions were available as well   * **They didn’t use # of comorbidities as a predictor** * Diagnoses were based on ICD 10 codes (german version)   Outcome was “high cost patient” ~ top 5% of the cost distribution.   * Used ANN, RF, GBM * ANN: feed-forward, [1,2,3] **hidden layers**, [0.003, 0.005, 0.007] **learning** **rate**, [sigmoid (with/without dropout), maxout (with/without dropout), rectifier (with/without dropout)] **activation function**, [10,20,100] units within each hidden layer, cross entropy loss function as a measure of weight fits * Predicted outcomes for year t1 trained on data from year t0(so the previous year)   + Used k-fold cross validation * Testing data came from predictors of t1 and outcome of year t2   + So following year compared to training data (might be biased?) * Variable selection based on importance   + SHAP analysis?   + Important variables:     - Total cost     - Age     - Number of hosp. Diagnoses     - Number of ATC prescriptions * Identified data balancing as a potential improvement   + Could try different methods * Identified the inclusion of text-based data sources as a potential improvement   + Could try to incorporate an NLP/ LLM model into this |
| Machine Learning for Health Services Researchers[https://www.sciencedirect.com/science/article/pii/S1098301519301469#abs0010](https://www.sciencedirect.com/science/article/pii/S1098301519301469#section-cited-by) | General “guide” to approaching healthcare data using a machine learning framework   * Discusses pros and cons of different models and frameworks in different scenarios * Explains some specifics of different machine learning methods   + I.e. general principles of neural networks and decision trees |
| Using machine learning for healthcare challenges and opportunities  * <https://www.sciencedirect.com/science/article/pii/S2352914822000739> | * General overview of ML applications in healthcare   + Not super specific about how it could be used for prediction of LOS or severity |
| Identifying Racial Disparities in Utilization and Clinical Outcomes of Ambulatory Hip Arthroscopy: Analysis of Temporal Trends and Causal Inference via Machine Learning  * <https://pmc.ncbi.nlm.nih.gov/articles/PMC11418677/> | * No prediction of LOS or severity |
| Medicare fraud detection using neural networks  * <https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0225-0#Sec12> |  |
| Demographics, comorbidities, and comedications in newly diagnosed patients with Alzheimer's disease and related dementias: Findings from United States Medicare claims data  * <https://journals.sagepub.com/doi/10.3233/JAD-231488?icid=int.sj-full-text.similar-articles.2> |  |
| dxpr: an R package for generating analysis-ready data from electronic health records-diagnoses and procedures<https://pubmed.ncbi.nlm.nih.gov/34141876/> |  |
| Association of COVID-19 Pandemic with Colorectal Cancer Screening: Impact of Race/Ethnicity and Social Vulnerability  * <https://pmc.ncbi.nlm.nih.gov/articles/PMC10997707/#Sec1> |  |
| Machine-Learning-Based Electronic Triage More Accurately Differentiates Patients With Respect to Clinical Outcomes Compared With the Emergency Severity Index  * <https://www.sciencedirect.com/science/article/pii/S0196064417314427?via%3Dihub> | * Can’t access |
| Concordance in Medical Urgency Classification of Discharge Diagnoses and Reasons for Visit  * <https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2813806> | * You can’t classify medical urgency based on patients’ reasons for visit using their algorithm that was based on the NYU algorithm |
| Moving Beyond the NYU Algorithm for Emergency Department Visit Appropriateness  * <https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2813809> | * NYU algorithm is not good for prospective classification, and new methods for that need to be designed |
| A Comparative Analysis of Machine Learning Models in Predicting Emergency Department Patient Volumes in a University Hospital<https://www.annemergmed.com/article/S0196-0644(24)00799-6/fulltext> | * Used RF, LR, SVM to predict patient volumes in the ER with varying time lags   + Most effective as a day lag |
| Lit Review (2022)<https://pmc.ncbi.nlm.nih.gov/articles/PMC9931263/#sec008> | * Not a whole lot of ML applied to LOS predictions * Compartmental Modeling   + Used to model the flow of patients between units   + Queueing systems     - Can help planners in testing for bottlenecks   + Markov models are a type of these “compartmentalized models”   + Ex. of non markov model compartmentalized model is one that assumes arrivals into each compartment are exponentially distributed   “ML”   * RF regression * LOS is a continuous outcome, so you need to use some kind of regression * Classification of LOS   + Low, medium, high     - Decision trees, SVMs, ANNs |
| <https://pmc.ncbi.nlm.nih.gov/articles/PMC7283236/>  * Deep Learning for Improved Risk Prediction in Surgical Outcomes |  |
| <https://machinelearningmastery.com/xgboost-for-regression/>  * XGBoost for regression | * Can use on the existing dataset to get a regression model that incorporates optimization and decision trees |
| <https://pmc.ncbi.nlm.nih.gov/articles/PMC7137462/>  * Predicting Inpatient Length of Stay After Brain Tumor Surgery: Developing Machine Learning Ensembles to Improve Predictive Performance | * Used 29 different models (adhered to the **no free lunch theorem**) * Used **Root Mean Squared Logarithmic Error**   + Penalizes larger errors less, when predicted and actual values are both large   + Interpreted as the standard deviation of the log of the unexplained variance (eg. error) * Uses **permutation importance and partial dependence** to identify which features are important   + Would be good to try to implement |
| <https://bmchealthservres.biomedcentral.com/articles/10.1186/s12913-024-11238-y#Sec4>  * Predicting hospital length of stay using machine learning on a large open health dataset | * Target encoding: Take values of the categorical variable, find the mean of the target variable for each category, and replace the categorical values with that number (the mean of the target value per category)   + This could be very interesting for LOS prediction     - Take ICD-10 code (technically a categorical variable) and replace that with the mean LOS for each code * Shapely Additive Explanation (SHAP) method for feature importance * Classification into 6 bins   + 1 day, 2 days … 6+ days   + CatBoostClassifier: Gradient boosted decision trees (python library also contains regressors and rankers)     - Catboost handles categorical data natively, whereas XGBoost requires data preprocessing * For regression, they used a nn.Linear() layer from pytorch? |

Abstract

Abstract

The length of a patient's hospital stay is a key factor in efficiently planning and managing hospital resources. Accurately predicting length of stay (LOS) is of great interest, as it can enhance patient care, reduce hospital costs, and improve service efficiency. However, LOS prediction has mainly been studied on an ad-hoc basis, making it hard to generalize the findings to other hospitals [(Stone et al., 2022)](https://pmc.ncbi.nlm.nih.gov/articles/PMC9931263/#sec008). Therefore, I will try to generalize the methods used in some of these ad hoc studies to synthetic CMS data. In addition I will attempt to replicate classification methods of patient LOS, based on different definitions of classes of LOS. Using methods such as Neural Networks, … (ADD OTHER METHODS HERE)

* Identification of need
  + LOS and disease severity is hard to predict
  + Very few algorithms or methods available to accurately predict these metrics
  + Some studies have been able to achieve decent results ([Predicting Inpatient Length of Stay After Brain Tumor Surgery: Developing Machine Learning Ensembles to Improve Predictive Performance](https://pmc.ncbi.nlm.nih.gov/articles/PMC7137462/#sec2), 2018)
    - Replicating these methods on different datasets is a useful experiment to test the generalizability of said methods
  + These metrics are important because … (below)
* Justifications
  + LOS predictions and categorizations can help hospital administration allocate resources, identify anomalously needy patients and areas with potential for improvement
  + Risk of high severity predictions can do the same as above
  + Both predictions can also enhance patient care, reduce hospital costs, reduce unnecessarily long lengths of stay through quicker interventions, and improve service efficiency
  + Therefore my goals are to … (below)
* Goals:
  + Predict length of stay (regression methods)
  + Predict risk of categorical outcomes
    - Risk of Prolonged length of stay
    - Risk of outcome with high severity
  + I’m going to develop models to predict these by … (below)
* Methods
  + Data:

1. Source
2. Data cleaning
3. Data transformations
4. Variable selection
   1. Try PCA?
   * Models:
     + Take general approach from [Predicting Inpatient Length of Stay After Brain Tumor Surgery: Developing Machine Learning Ensembles to Improve Predictive Performance](https://pmc.ncbi.nlm.nih.gov/articles/PMC7137462/#sec2)
5. Partition Data into training set and test set
   1. Training set will be used to conduct 5-fold CV
   2. 80-20 split
6. No free lunch approach to model selection (i.e. try everything)
   1. Try all models
      1. Train models using 5-fold CV
   2. Tune models
      1. bayesian optimization (see [Deep Learning for Improved Risk Prediction in Surgical Outcomes](https://pmc.ncbi.nlm.nih.gov/articles/PMC7283236/#Sec2))
      2. Grid search (scikit learn)
      3. Random search (scikit learn)
7. Ensemble best performers
   1. Combined in paper with elastic net
   2. Train on entire training set, and then evaluate using test set
8. Variable importance
   1. Use permutation importance and partial dependence to evaluate variable importance
      1. This is very important in any medical setting
      2. The black box alone is not enough, and there needs to be more sound reasoning to actually convince medical practitioners
9. Final evaluation/validation (ideal scenario)
   1. Somehow obtain real CMS data
   2. Evaluate model on the real data for improved reliability and credibility

### **Abstract**

Predicting hospital length of stay (LOS) and disease severity remains a critical challenge in healthcare management. Accurate predictions can assist hospital administrators in optimizing resource allocation, identifying high-risk patients, and improving overall care efficiency. However, LOS prediction has mainly been studied on an ad-hoc basis, making it hard to generalize the findings to other hospitals [(Stone et al., 2022)](https://pmc.ncbi.nlm.nih.gov/articles/PMC9931263/#sec008). Replicating and extending these methodologies on different datasets is essential to assess their generalizability and effectiveness.

This study aims to develop predictive models for LOS (using regression methods) and categorical risk outcomes, including the risk of prolonged hospitalization and severe patient outcomes. Using a synthetic Center for Medicaid and Medicare Services dataset, we will apply rigorous data preprocessing techniques such as cleaning, transformation, and variable selection, potentially incorporating principal component analysis (PCA).

Our modeling approach follows a structured methodology inspired by previous research ([Muhlestein et al., 2018](https://pmc.ncbi.nlm.nih.gov/articles/PMC7137462/#sec2), [Jain et al, 2024](https://bmchealthservres.biomedcentral.com/articles/10.1186/s12913-024-11238-y#Abs1) ) on LOS prediction, incorporating machine learning ensembles, classification and regression, and gradient boosting methods . We will partition data into training and test sets, employ five-fold cross-validation, and explore a broad range of models using Bayesian optimization, grid search, and random search for hyperparameter tuning. The best-performing models will be ensembled using elastic net regression to improve predictive accuracy.

To ensure interpretability in a medical setting, we will analyze variable importance using permutation importance and partial dependence plots. Given the complexity of machine learning models, interpretability is crucial for gaining trust from medical practitioners. Finally, in an ideal validation scenario, we aim to evaluate the model on real Centers for Medicare & Medicaid Services (CMS) data to enhance credibility and real-world applicability. This study contributes to improving hospital efficiency, reducing unnecessary costs, and enhancing patient care through data-driven decision-making.

Results

**Results**

**Data Processing**

**EDA**

* No major correlation between LOS and number of diagnoses
  + Conclusion based on visual evidence from eda script
    - Justifies averaging the number of diagnoses from the past n years, as discussed in the model ideas doc

**Modelling**

Simple Linear Regression

* Both models (training scheme 1 vs 2) had pretty similar results
  + Scheme 1:
    - 'RMSLE': 0.46742167078115165,
    - 'R2': 0.30129147112574084,
    - 'MAE': 0.9067930411502675,
    - 'MSE': 5.296115062132563
  + Scheme 2:
    - 'RMSLE': 0.45243282611375374,
    - 'R2': 0.3600572124500635,
    - 'MAE': 1.071987453311116,
    - 'MSE': 8.456505768259749