Quantitative Methods and Decision Analysis (224HCA203)

Final Project - Thomas Locke

June 16, 2023

Cohort: Patient Readmissions

Patient readmissions are very costly in healthcare systems and have negative impacts on patient's health outcomes. According to CMS, unplanned readmissions within 30 days of hospital discharges cost Medicare billions of dollars annually. I chose this cohort to study patient readmission within 60 days of discharge based on age, chronic conditions, and total costs. Based on my study using Machine Learning, I want to find out if we can predict when a patient will be readmitted after discharge so that we can better prepare to "take care" of them to minimize any chance from coming back to the hospital (readmission).

Data Preparation

I chose the provided dataset from CMS for analytics:
DE1_0_2008_to_2010_Inpatient_Claims_Sample_13.csv
DE1_0_2008_Beneficiary_Summary_File_Sample_13.csv

Initially I was going to include all inpatients from 2008 thru 2010 for my study but as I started analyzing patient conditions (Column names with SP_*), I found that some chronic conditions appeared to one patient in 2008 but not in 2009 or 2010 of the same patient (matching DESYNPUF_ID). In order to have definitive chronic conditions for my study I decided to use beneficiary 2008 data only. See attached SQL code for how I extracted only inpatients that matched 2008 beneficiary.

I did some of the preliminary data analysis, extract and additions using SQL. These include:

- A new inpatient table 'Inpatient_Claims_2008_FUTURE_CHKIN' with additional calculated columns:
- 'FUTURE_CHECKIN': Future check-in date (taken from the 'next' readmission date)
- 'FUTURE_GAP': Number of days between future readmissions
- '60_DAYS_READMIT': Binary value either that patient readmitted within 60 days (1 or 0)

(See **Appendix A** for SQL code)

Modified the beneficiary table with the following additions:

- 'BENE_AGE_ASOF_2008': Patient's age
- 'BENE_AGE_CATS': Age categories ('AGE UNDER 55', 'AGE 55-64', 'AGE 65-74', 'AGE 75-84', 'AGE 85-94')
- All SP_* fields: Convert '1' to '1' and '2' to '0' (1=Yes, 0=No)
- 'TOTAL_CONDITIONS': Total number of chronic conditions of a patient
- 'TOTAL_READMISSION': Total number of readmissions of a patient

(See **Appendix B** for SQL code)

Data Exploratory in Jupyter Notebook (Python)

See **Appendix C** for Python code

Files used:

```
InPatient = pd.read_csv("Final Project/data/CMS_Inpatient_Claims_2008_FUTURE_CHKIN.csv")
Bene_2008 = pd.read_csv("Final Project/data/CMS_Beneficiary_2008_v4.csv")
```

Total Readmissions	Count	Percentage
7	5	0.03
6	6	0.04
5	17	0.11
4	80	0.5
3	271	1.71
2	841	5.3
1	3,012	18.98
0	11,635	73.33
Total	**15,867**	

Readmission percentage: (15,867-11,635) / 15,867 = 26%

% By Race (1=White, 2=Black, 3=Other, 5=Hispanic)

```
1    84.100355
2    10.862851
3     3.050292
5     1.986502
Name: BENE_RACE_CD, dtype: float64
```

% By Sex (1=Male, 2=Female)

```
2    57.051893
1    42.948107
Name: BENE_SEX_IDENT_CD, dtype: float64
```

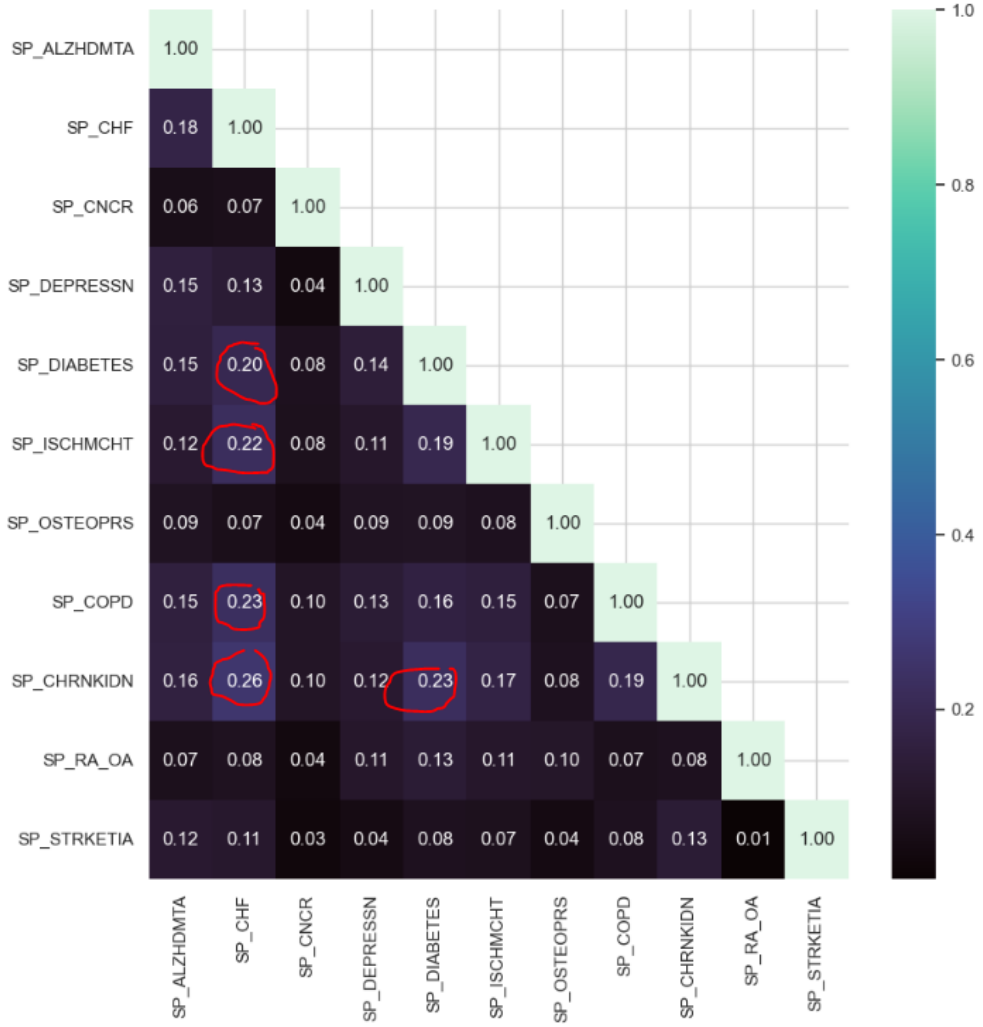
Correlations among chronic conditions heat map:

Patient's Chronic conditions (From Beneficiary table):

SP_ALZHDMTA	Alzheimer or related disorders or senile
SP_CHF	Chronic Condition: Heart Failure
SP_CHRNKIDN	Chronic Condition: Chronic Kidney Disease
SP_CNCR	Chronic Condition: Cancer
SP_COPD	Chronic Condition: Chronic Obstructive Pulmonary Disease
SP_DEPRESSN	Chronic Condition: Depression
SP_DIABETES	Chronic Condition: Diabetes
SP_ISCHMCHT	Chronic Condition: Ischemic Heart Disease
SP_OSTEOPRS	Chronic Condition: Osteoporosis

SP_RA_OA Chronic Condition: rheumatoid arthritis and osteoarthritis (RA/OA)
SP_STRKETIA Chronic Condition: Stroke/transient Ischemic Attack

Here, the circles indicate patients with those conditions are highly correlated.
(Chronic Kidney Disease and Heart Failure are highly correlated, 0.26, for example)



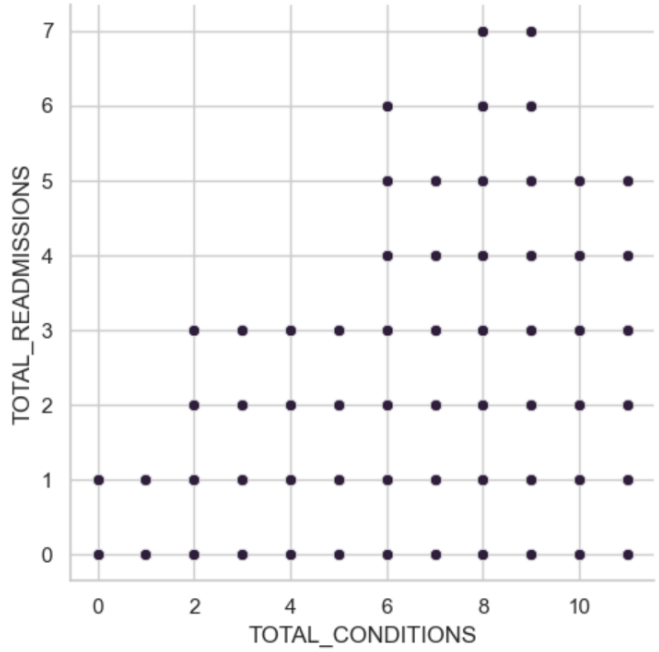
Total Chronic Conditions

Patients that had between 5 and 7 chronic conditions were readmitted the most (see percentage ranking below):

```
7     18.290312
6     18.191177
8     15.499295
5     13.669120
4      9.879132
9      8.697144
3      6.462806
2      3.587906
10     2.848210
1      1.830175
0      0.556678
11     0.488047
Name: TOTAL_CONDITIONS, dtype: float64
```

Total Readmissions vs Total Conditions

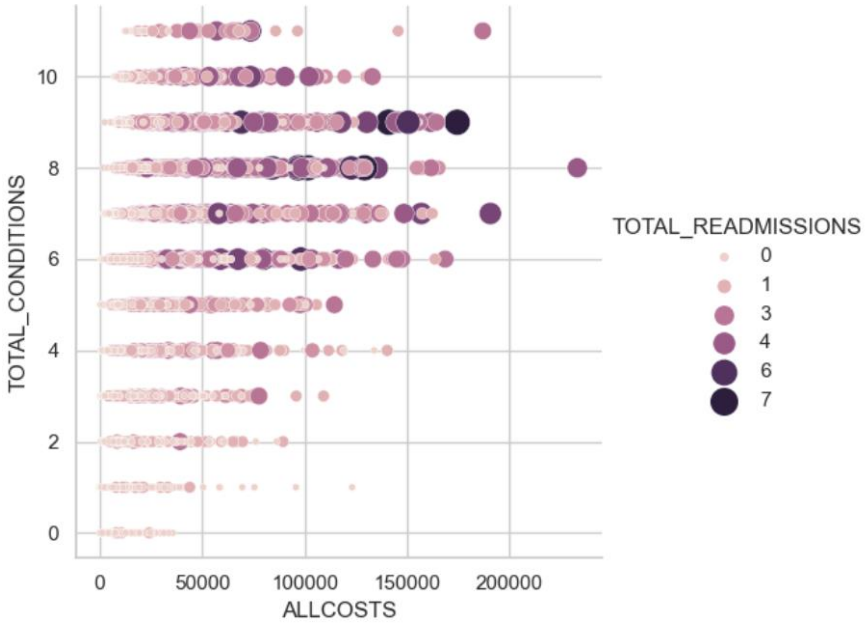
The more chronic conditions a patient had, the chance of being readmited are higher.



Total Conditions vs Total Costs

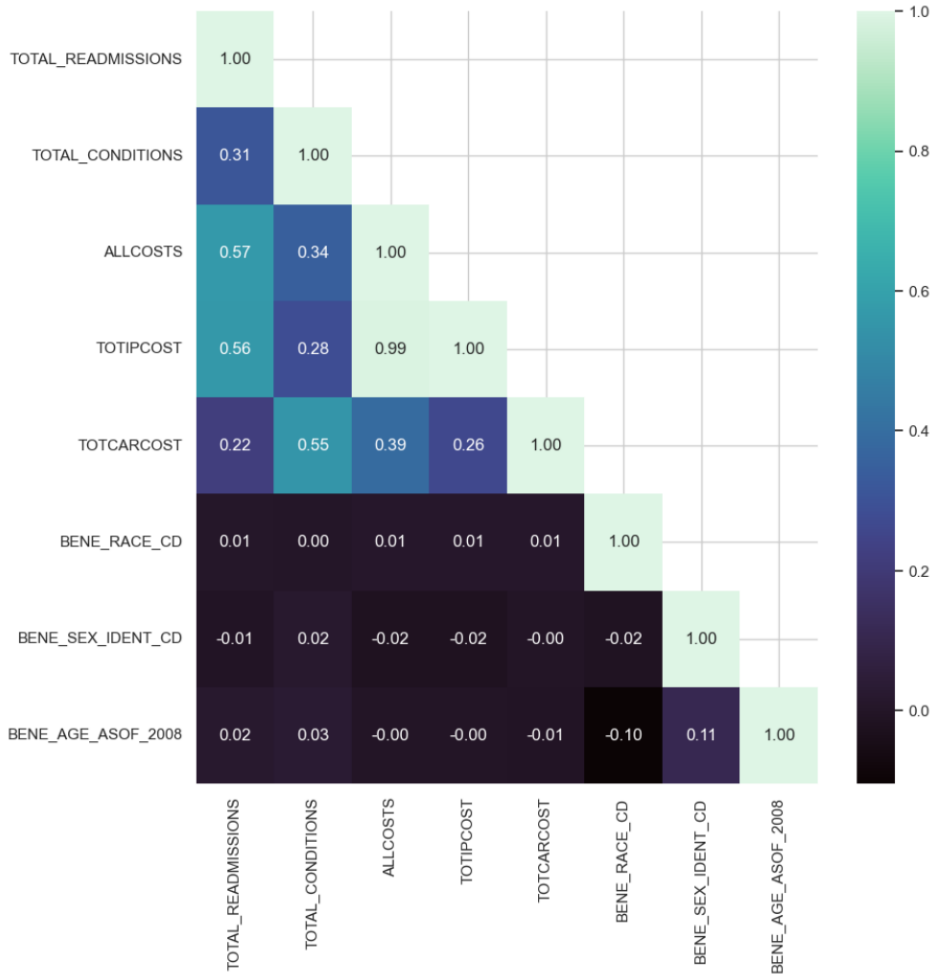
Total Costs = MEDREIMB_IP + BENRES_IP + PPPYMT_IP + MEDREIMB_CAR + BENRES_CAR + PPPYMT_CAR

The more chronic conditions a patient had the higher the total cost (darker color, below). The largest number of readmissions appear to occur when patients have between 6 and 10 chronic conditions, and this very much aligns with total costs.



Comparing Total Readmissions, Total Conditions and Total Costs, Sex, and Race

Not surprisingly, total costs correlate with total conditions, but surprisingly, sex, and race were less correlates with total readmissions.



Readmit Category Percentage

Low=0-1, Moderate: 2-5, High: 6+

```
LOW         73.328291
MODERATE    18.926073
HIGH         7.745636
Name: READMIT_CATEGORY, dtype: float64
```

Output file: Two files were combined and saved as **readmissions_visits.csv**

Predicting Patient Readmissions (Machine Learning)

See **Appendix D** for Jupyter (Python) Code.

Using **L1 Regularized Logistic Regression** and **Cross-Validated Recursive Feature** Elimination (**RFE**)

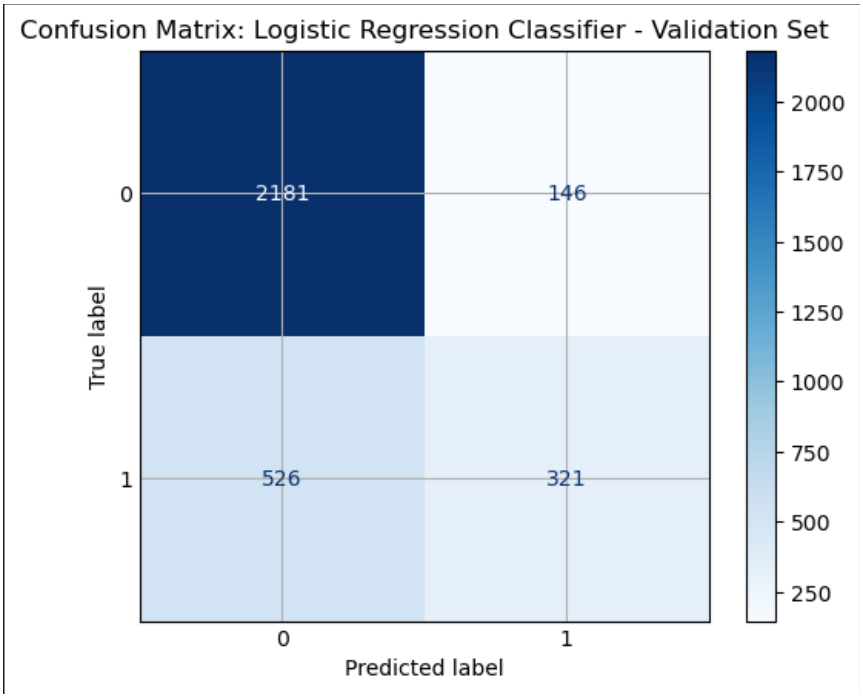
RFE works by searching for a subset of features by starting with all features in the training dataset and successfully removing features until the desired number remains.

The model did best up to about 13th feature. The accuracy topped out at 13th. Here are the results (highlighted yellow):

'AGE', 'SP_ALZHDMTA', 'SP_CHF', 'SP_CHRNKIDN', 'SP_CNCR','SP_COPD', 'SP_DEPRESSN', 'SP_DIABETES', 'SP_ISCHMCHT', 'TOTAL_CONDITIONS', 'TOTIPCOST', 'TOTCARCOST', 'ALLCOSTS'

Interestingly, age, Alzheimer or related disorders or senile (SP_ALZHDMTA), and Congestive Heart Failure (SP_CHF) are the top 3 features that are highly correlated.

Confusion Matrix using Validation Set



N: Total # of samples in the validation set: 3,174

True Positives (TP): 321

True Negatives (TN): 2,181

False Positives (FP): 146

False Negatives (FN): 526

Accuracy: 80.1%
The accuracy is the % of samples CORRECTLY CLASSIFIED by the model out of all the samples in the set.

Precision: 68.7%,
Precision is the number of samples that actually belong to the positive class out of all the samples predicted to belong to it.

Recall: 37.9%
The number of samples PREDICTED CORRECTLY as belonging to the POSITIVE class out of all the samples that belong to the positive class.

F1-Score: 48.9%
The harmonic mean of the precision and recall scores obtained for the positive class.

Specificity: 93.7%
The # of samples PREDICTED CORRECTLY as belonging to the NEGATIVE class out of all the samples that actually belong to the negative class.

(See frame [24] for detailed calculations in Jupyter code in **Appendix D**)

Class Likelihood Ratios: LR+6.040, LR-0.663
A higher LR+ value of 6.040 implies that the model has some ability to differentiate between patients who will be readmitted and those who will not be readmitted. However, it's not very great (See frame #25 in **Appendix D**)

Thoughts/Action Plans

The model predicted very well on patients that don't come back (93.7%) but what we really care about is patient readmissions. Unfortunately, our model didn't do so well for predicting patient readmissions. The Precision and Recall (68% and 38%, respectively) support that. However, based on the logistic regression function **predict_proba**() function which predicts chances of readmission or no readmissions (see frame #22 in **Appendix D**), we can use the results to follow up those patients with 80% or higher with exceptional care.

```
lr.predict_proba(X_val)
```

```
array([[0.971, 0.029],
       [0.87 , 0.13 ],
       [0.753, 0.247],
       ...,
       [0.858, 0.142],
       [0.828, 0.172],
       [0.954, 0.046]])
```

Patient A: Probably of readmission: 97%, probability of no readmission: 3%
Patient B: Probably of readmission: 87%, probability of no readmission: 13%
Etc.

Lessons Learned

The following highlights what this project has discovered:

- Age: The older patients are the higher chance that they may readmit
- Chronic conditions: The more conditions a patient has, the higher the chance that they may be admitted.
- Total costs are highly correlated with readmissions with patients that have many chronic conditions.

We could try using **decision tree** classification algorithms to predict which patient's age group may readmit. In this project, I did categorize patients into age group ([BENE_AGE_CATS] column in Beneficiary_2008 dataset) but have not had a chance to use it for predictive modeling.

Healthcare Intervention

As mentioned earlier we could target patients with a higher probability of readmission to provide them with exceptional care and attention, and education in the hope that we can reduce the chance of them from coming back to the hospital.

Appendix A

Create InPatient 2008_with Future readmission future gap.sql

Appendix B

Create Beneficiary_2008_v4.sql

Appendix C

Final_Project_v3.html

Appendix D

Final_Project_v2_Part_2.html