Introduction

* Describe what Synthea is
  + Open source generator of synthetic patient data
  + Realistic patient data for modeling
* Describe goal of project
  + Develop model to predict 2024 patient encounters costs based on their most recent health data/health data from 2023
  + Useful for patients and doctors to determine how they can reduce healthcare costs through lifestyle interventions
* Outline of paper
  + Data Generation
  + Data Wrangling

Data Generation

* Follow instructions from Synthea github page to download data
  + 100 living patients from each of the 50 states
* List all the csv files it spits out (18?)
  + Used –
    - Allergies – patient allergy data
    - Encounters – patient encounter data
    - Immunizations – patient immunization data
    - Medications – patient medication data
    - Observations – patient observations including vital signs and lab reports
    - Patients – patient demographic data
    - Procedures – patient procedure data including surgeries
  + Unused – careplans, claims, claims\_transactions, conditions, devices, imaging\_studies, organizations, payer\_transitions, payers, providers, supplies

Data Wrangling

* Patients
  + Load csv into workspace
  + Contains 5871 entries – 100 patients from each of the 50 states, plus dead ones
  + Contains birth/deathdate, name, race, ethnicity, gender, state, income
  + Id, BIRTHDATE, DEATHDATE, SSN, DRIVERS, PASSPORT, PREFIX, FIRST, MIDDLE, LAST, SUFFIX, MAIDEN, MARITAL, RACE, ETHNICITY, GENDER, BIRTHPLACE, ADDRESS, CITY, STATE, COUNTY, FIPS, ZIP, LAT, LON, HEALTHCARE\_EXPENSES, HEALTHCARE\_COVERAGE, INCOME
  + Check number and percentage of missing values
    - lots for SUFFIX, DEATHDATE, MAIDEN, MARITAL, PASSPORT, PREFIX, MIDDLE, DRIVERS, FIPS
    - These wont be useful for analysis so drop them
  + Remove rows with value in deathdate to keep living patients – now contains 5000 entries
  + Calculate patient age at the end of 2023 by subtracting their birthdate from December 31st 2023
* Encounters
  + Load csv into workspace
  + Id, START, STOP, PATIENT, ORGANIZATION, PROVIDER, PAYER, ENCOUNTERCLASS, CODE, DESCRIPTION, BASE\_ENCOUNTER\_COST, TOTAL\_CLAIM\_COST, PAYER\_COVERAGE, REASONCODE, REASONDESCRIPTION
  + Use START column as a timestamp for encounter. Convert to datetime and filter for just 2023
  + Calculate total 2023 encounter cost per patient (encounters\_cost) and number of encounters per patient (num\_encounters)
  + Merge with patient dataframe based on Id keeping just encounters\_cost and num\_encounters
* Graphs
  + Very right skewed, most cost values in 0-50,000 range

A graph of a graph

Description automatically generated

* + Very right skewed, most visit values in 0-20 range

A graph of a number of people

Description automatically generated

* + From this plot, we can see that Arizona has the highest average medical encounter cost, and Wisconsin has the lowest

A graph of a number of blue lines

Description automatically generated

* + The plot is truncated to the $0-50,000 range so we can see the boxplots more clearly. There are many high outliers, which agrees with what we saw in the right skewed histogram earlier. The median total medical encounter cost does not seem to vary too much between states.

A graph of blue and white columns

Description automatically generated

* + The average number of medical encounters per person in 2023 for all states appears to be within 3-9. Interestingly, Arkansas and Louisiana have the highest average number of medical encounters per person, but Arkansas has the 31st and Louisiana has the 9th average medical cost per person in 2023.

A graph of a number of patients

Description automatically generated

* + The plot is truncated to the 0-20 range so we can see the boxplots more clearly. Again, we see that this data has a lot of higher outliers and is right skewed, confirming what we saw in the histogram. The median number of medical encounters does not seem to vary too much between states.

A graph with blue and black lines

Description automatically generated with medium confidence

* + When plotting medical cost by gender, we see that women tend to have higher total average medical costs in 2023 compared to men. Arizona has the highest total average medical cost for women, which may contribute to Arizona having the highest total average medical cost overall. South Dakota has the highest total average medical cost for men, which appears to be much higher than New York in 2nd place.

A graph of a number of different colored lines

Description automatically generated with medium confidence

* + To visualize the spread of the data, the data can be displayed with boxplots. The plot is truncated to the $0-60,000 range so we can see the boxplots more clearly. As expected, the data is very right skewed with many high outliers. Women tend to have greater interquartile ranges and median values than men. Again, the median cost of medical encounters does not seem to vary as much between states and genders compared to the mean cost.

A graph of different colored columns

Description automatically generated

* + There does not seem to be a large difference between White, Native, Black, and Asian races. Hawaiian is slightly higher, and Other has the highest total average medical encounter cost.

A graph of a bar chart

Description automatically generated with medium confidence

* Medications
  + Load csv into workspace
  + START, STOP, PATIENT, PAYER, ENCOUNTER, CODE, DESCRIPTION, BASE\_COST, PAYER\_COVERAGE, DISPENSES, TOTALCOST, REASONCODE, REASONDESCRIPTION
  + Use START column as a timestamp for medication. Convert to datetime and filter for just 2023
  + Calculate total 2023 medication cost per patient (medications\_cost) and number of medications per patient (num\_medications)
  + Merge with patient dataframe based on Id, keeping just medications\_cost and num\_medications
* Procedures
  + Load csv into workspace
  + START, STOP, PATIENT, ENCOUNTER, SYSTEM, CODE, DESCRIPTION, BASE\_COST, REASONCODE, REASONDESCRIPTION
  + Use START column as a timestamp for procedure. Convert to datetime and filter for just 2023
  + Calculate total 2023 procedure cost per patient (procedures\_cost) and number of procedures per patient (num\_procedures)
  + Merge with patient dataframe based on Id, keeping just procedures\_cost and num\_procedures
* Immunizations
  + Load csv into workspace
  + DATE, PATIENT, ENCOUNTER, CODE, DESCRIPTION, BASE\_COST
  + Use DATE column as a timestamp for immunization. Convert to datetime and filter for just 2023
  + Calculate total 2023 immunization cost per patient (immunizations\_cost) and number of immunizations per patient (num\_immunizations)
  + Merge with patient dataframe based on Id, keeping just immunizations\_cost and num\_immunizations
* Allergies
  + Load csv into workspace
  + START, STOP, PATIENT, ENCOUNTER, CODE, SYSTEM, DESCRIPTION, TYPE, CATEGORY, REACTION1, DESCRIPTION1, SEVERITY1, REACTION2, DESCRIPTION2, SEVERITY2
  + Keep all allergies (not just ones found in 2023)
  + Calculate number of allergies per patient (num\_allergies)
  + Merge with patient dataframe based on Id, keeping just num\_allergies
* Observations
  + Load csv into workspace
  + DATE, PATIENT, ENCOUNTER, CATEGORY, CODE, DESCRIPTION, VALUE, UNITS, TYPE
  + Use DATE as timestamp for observation. Convert to datetime and get rid of ones after 2023 (keeping older ones)
  + For this study, we will look at the most recent observations for each patient. So, the dataframe is first sorted such that the most oldest observations are at the top and the most recent observations are at the bottom. Then, it is checked for any duplicate observations for each patient, and the last (or most recent) one will be kept.
  + Pivot table such that each patient is a row and each column is an observation. There are 305 different observations
  + Convert into its proper type
  + Count missing values and percentages, remove columns where there are no more than 70% missing values
  + Merge with patient dataframe based on Id
* Cleaning up data
  + It is important to manually check over the columns and determine which ones will be useful to include for analysis. Some columns may be repeats of one another, and others may not be useful for statistical analysis and modeling (such as first and last name, zip code, etc). The columns are examined and a list of the most useful columns is created to filter the dataframe.
  + Certain columns of type 'string' or 'object' that can be converted into numerical values are converted appropriately.
    - Stress level – 1:Not at all, 2:A little bit, 3:Somewhat, 4:Quite a bit, 5:Very much, 0:everything else
    - Tobacco smoking status – 1:Ex-smoker, 2:Smokes tobacco daily, 0:everything else
* Summary of section
  + In this section of the capstone project, several csv files containing synthetic medical records from Synthea were loaded into the workspace. These data tables were modified so that they can be merged together to create one large dataset for modeling purposes. During the data wrangling steps, it was determined that the total medical encounter cost for 2023 was skewed to the right, with many high outliers. This variable will be the target feature for modeling later in the project, so the shape of this data will have to be taken into account in future steps. The observations csv file contains the results of many useful medical tests and measurements, which may prove useful in predicting a patient's medical encounter costs for the year. However, there are many missing values throughout this data set. It will be important to infer why these values might be missing, such as the test not being required or appropriate for the patient. Additionally, the best method to impute missing values must be determined, including techniques such as mean, median, multiple imputation using chained equations (MICE), probabilistic principal component analysis (PPCA), or an imputation technique from Python's sci-kit learn package. Different techniques can be tested, and the shape of the data and model performance can be analyzed. In the next section of this project, exploratory data analysis will be performed to understand relationships between features.

Exploratory Data Analysis

* In the exploratory data analysis (EDA) section of this project, features of the data are visualized and explored to determine which attributes may be most useful in modeling steps. All string data will be transformed into numerical data where possible to assist in analysis and in preparation for modeling. Any trends or correlations in the data will be noted.
* Turn strings into numeric
  + Data contains 5000 patients and 63 features
  + GENDER – male = 0, female = 1
  + ETHNICITY – nonhispanic = 0, Hispanic = 1
  + RACE – one hot encoding
  + Now 67 features
* States
  + Next, the states column will be turned into a numeric format. There are multiple ways in which to do this, including one hot encoding, label encoding, and target encoding. Since there are 50 states, one hot encoding would add 49 additional columns, which could cause problems with dimensionality. Label encoding would give each state an integer value from 0 to 49, but since there is no inherent ranking to the states, the behavior of this feature could be disruptive to modeling. To introduce target encoding, the population of each state will be used. This could give some information on the availability of healthcare within a state, as a state with a larger population would likely have more cities and thus more healthcare facilities. To do this, state data from Wikipedia is imported and the population column is merged with the patients dataframe. Then, the state and city columns can be dropped.
  + Import state information from Wikipedia
  + Extract state name and population
  + Merge to DF on state name
  + Drop state and city names from dataframe
  + Now 66 features
* Healthcare expenses and columns
  + Finally, the 'HEALTHCARE\_EXPENSES' and 'HEALTHCARE\_COVERAGE' columns will be removed as these are lifetime totals, and this analysis is just for a yearly estimate.
  + Patient ID will be removed as well since this is a unique patient identifier and is not useful for data visualization.
  + Now 63 features
* Visualization
  + Histograms for each variable
  + There is a lot of variety in the distributions of the features shown above. Some follow very normal distributions, including body weight, BMI, and diastolic and systolic blood pressure. Others follow somewhat uniform distributions, including age, erythrocyte distribution width, erythrocytes in blood, platelet distribution width, platelet mean volume, and platelets in blood. Many features have right skewed distributions, such as income, state population, number of encounters, number of procedures, number of immunizations, generalized anxiety score, heart rate, pain severity, respiratory rate, and stress level.

A collage of blue and white graphs

Description automatically generated

* + Next, scatterplots are created with each feature on the x axis and the target feature, encounters cost, on the y axis. This will reveal any correlations between encounters cost and the features that would be useful to know heading into modeling.
  + There are not too many clear correlations with encounters cost. However, there are a couple of key relationships that should be noted. There is a strong positive correlation between procedures cost and encounters cost. This makes sense, as procedures cost refers to the cost of whatever procedure was performed on a patient during an encounter. Since this variable is so strongly related to the target feature, it will be removed from the analysis. There are also positive correlations between the number of encounters and the number of procedures and encounters cost, which also makes sense. These variables will remain in the analysis for now, since individual procedures and encounters would have variable costs. Another visible positive correlation is height, and a few negative correlations include income, immunization cost, and number of immunizations. Other than that, there are not too many very clear relationships, so it will be informative to see which features are most important in the modeling stage of this project.

A screenshot of a graph

Description automatically generated

* + Additionally, a heatmap of correlation coefficients is created to visualize correlations between features.
  + As identified in the scatterplots, some features that appear to be correlated with encounters cost include number of procedures, number of encounters, and number of medications.

A red square with a white line

Description automatically generated

* + Lastly, boxplots of integer features will be observed to determine differences in distributions of these features versus encounter cost. Since there are many high outliers in encounter cost, the y axis will be limited to $100,000 so that the box can be seen clearly.
  + There do not appear to be large differences between ethnicity or race. Males have a higher interquartile range of encounters costs compared to females.

A group of blue and black graphs

Description automatically generated with medium confidence

* Summary of section
  + In the EDA section of this project, relationships between variables were analyzed to see if there were any potential correlations between encounters cost and all other features of the dataset. The histograms of the features showed a wide variety of distributions for the variables, which makes sense since there are so many different types of features and medical measurements present. The scatterplots of encounter cost versus each feature also displayed a variety of relationships, most of which did not appear too strongly positive or negative. The heatmap of correlation coefficients further confirmed that there were only a few features with a strong positive or negative correlation with the target feature. Lastly, the boxplots showed some differences in the distribution of encounters costs when grouped by different integer variables. In the next step of the project, it will be necessary to decide how to impute missing values and/or which features to remove from the analysis. Additionally, different models and parameters will be tested to determine if medical encounters costs can be predicted from patient data.

Data imputation

* Check missing values in each column
* It is assumed that missing values for cost/number of encounters, cost/number of medications, number of procedures, cost/number of immunizations, and number of allergies can be set to 0 as the patient likely did not have values for any of the above.
* Data imputation strategies
  + Mean – fill in missing values of the column with the mean value
  + Median – fill in missing values of the column with the median value
  + KNN – fill in missing values by using values of similar neighboring data points
  + MICE – multivariate imputation by chained equations. Uses regression models to fill in missing values
* Assessing data imputation strategies
  + The different imputation techniques are evaluated through comparing the R squared values for a simple ordinary least squares (OLS) model.

|  |  |
| --- | --- |
| Imputation Technique | R squared |
| Mean | 0.4934 |
| Median | 0.4941 |
| KNN | 0.5313 |
| MICE | 0.5139 |

* + Created histograms of distributions
  + In general, median, mean, and MICE imputation seems to create unimodal distributions that do not reflect the distribution of the original dataframe without imputed values. The KNN imputation method seems to make the distribution more normalized and is about the closest to the original distributions.
* PCA
  + Performed PCA on each data imputation technique to see if dimensionality reduction should be considered

A graph of a line

Description automatically generated

A graph of a function

Description automatically generated

A graph with a blue line

Description automatically generated

A graph of a function

Description automatically generated with medium confidence

* + All imputation techniques showed that reducing the data to approximately 40 components would account for ~80% of the variance. This may be useful as the dataset has a lot of features, so reducing the dimensionality of the dataset may help with modeling and computation time.
  + Will not use in this project
* The KNN imputation technique is selected as it produced the highest R squared score, resulted in somewhat normalized distributions of imputed variables, and had similar cumulative variance ratios to the other imputation techniques when performing PCA.

Modeling

* Preparation for modeling
  + To begin model development, the data is split into train and test sets, with 25% of the data reserved for testing.
* Dummy regression
  + For a baseline model, the mean value is used as a predictor. This is performed using a dummy regressor.

|  |  |
| --- | --- |
| Mean value of training data | 15,040.24 |
| Training R squared | 0.0000 |
| Testing R squared | -0.0006 |
| Training MAE | 18,029.71 |
| Testing MAE | 17,276.07 |
| Training MSE | 1,214,707,483.91 |
| Testing MSE | 1,141,617,626.70 |

* + The R squared values are very low and the mean absolute error and mean squared error values are very high, showing that this is not a great fit for the model, as expected. These values will be saved to be compared to the final model.
* Scale the data with standard scaler
* Linear regression
  + Fit default model
  + Select best 10 features using f regression
  + 5 fold cross validation
  + Grid search to determine best number of features to include (37)
  + Plot mean R squared score versus number of features

A graph with blue lines

Description automatically generated

* + The graph shows that the R squared value flattens out around 20 features, so it is not necessary to include 37.
  + Final model results

|  |  |
| --- | --- |
| Training R squared | 0.4464 |
| Testing R squared | -0.1396 |
| Training MAE | 13,050.87 |
| Testing MAE | 12,799.56 |
| Training MSE | 672,463,349.73 |
| Testing MSE | 1,300,162,762.60 |

* + Since the testing R squared value is lower than the training R squared value, it can be concluded that the model is overfitting. The negative R squared value for the testing data set shows that this model is not a good fit. However, it is an improvement over the dummy regressor prediction when comparing R squared, MAE, and MSE.
* Ridge regression
  + Different values of alpha are tested to assess various levels of controlling regularization strength
    - Test alpha = [0.1, 1.0, 10.0, 100.0, 1000.0, 10000.0]
    - Fit ridge model and predict
    - Calculate r squared
    - alpha value of 1000 is the highest so that is selected.
  + Select best 10 features using f regression
  + 5 fold cross validation
  + Grid search to determine best number of features to include (46)
  + Plot mean R squared score versus number of features

A graph of a line

Description automatically generated

* + Just like the linear regression model, the graph shows that the R squared value flattens out around 20 features, so it is not necessary to include 46.

|  |  |
| --- | --- |
| Training R squared | 0.4073 |
| Testing R squared | 0.2567 |
| Training MAE | 12,431.39 |
| Testing MAE | 12,021.69 |
| Training MSE | 719,984,233.37 |
| Testing MSE | 848,007,970.02 |

* + Since the testing R squared value is lower than the training R squared value, it can be concluded that the model is overfitting. This model is a slight improvement over the linear regression model when comparing R squared, MAE, and MSE, but it is still not great.
* Lasso regression
  + Different values of alpha are tested to assess various levels of controlling regularization strength
    - Test alpha = [0.1, 1.0, 10.0, 100.0, 1000.0, 10000.0]
    - Fit lasso model and predict
    - Calculate r squared
    - alpha value of 1000 is the highest so that is selected.
  + Select best 10 features using f regression
  + 5 fold cross validation
  + Grid search to determine best number of features to include (43)
  + Plot mean R squared score versus number of features

A graph of a graph

Description automatically generated

* + Once again, the graph shows that the R squared value flattens out around 20 features, so it is not necessary to include 43.

|  |  |
| --- | --- |
| Training R squared | 0.4223 |
| Testing R squared | 0.4184 |
| Training MAE | 12,130.95 |
| Testing MAE | 11,354.02 |
| Training MSE | 701,754,831.80 |
| Testing MSE | 663,539,670.31 |

* + The testing R squared value is closer to the training R squared model, suggesting this model performed well with the testing data and is not overfitting to the training data. However, these R squared scores are quite low. This model is a slight improvement over the ridge regression model when comparing R squared, MAE, and MSE.
* Random forest
  + Fit default model and select best 10 features using f regression
  + 5 fold cross validation
  + Grid search to find best number of features and optimal number of trees in random forest
    - K = [10, 15, 20, 25, 30, 35, 40, 45, 50 55, 60]
    - N\_est = [10, 20, 30, 40, 50, 60, 70, 80, 90, 100]
    - Optimal: k = 45, n\_est = 80
  + Plot mean R squared score versus number of trees

A graph of a forest

Description automatically generated

* + Plot mean R squared score versus number of features

A graph of blue lines

Description automatically generated

* + Results of final model

|  |  |
| --- | --- |
| Training R squared | 0.9426 |
| Testing R squared | 0.6350 |
| Training MAE | 2,519.64 |
| Testing MAE | 6,167.80 |
| Training MSE | 69,728,646.81 |
| Testing MSE | 416,450,743.75 |

* + Cross validation on final model

|  |  |
| --- | --- |
| Mean R squared | 0.7065 |
| Standard deviation R squared | 0.1263 |

* Gradient Boosting
  + Fit default model and select best 10 features using f regression
  + 5 fold cross validation
  + Grid search to find best number of features and optimal number of boosting stages
    - K = [10, 15, 20, 25, 30, 35, 40, 45, 50 55, 60]
    - N\_est = [10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200]
    - Optimal: k = 55, n\_est = 190
  + Plot mean R squared score versus number of boosting stages

A graph of a number of boosting stages

Description automatically generated

* + Plot mean R squared score versus number of features

A graph of blue lines

Description automatically generated

* + Results of final model

|  |  |
| --- | --- |
| Training R squared | 0.9475 |
| Testing R squared | 0.6404 |
| Training MAE | 4,236.62 |
| Testing MAE | 6,439.04 |
| Training MSE | 63,764,250.97 |
| Testing MSE | 410,288,724.39 |

* + Cross validation on final model

|  |  |
| --- | --- |
| Mean R squared | 0.7129 |
| Standard deviation R squared | 0.1234 |

* Model selection
  + Both the random forest model and the gradient boosting model performed much better than any of the other models tested. The difference between these two is so minimal and is likely due to how the data was split into training and testing sets. Since both models are similar, the random forest model will be selected for this dataset. This model is saved and exported.
  + The random forest model scores are compared to the dummy regressor model scores to show its improvement.

|  |  |
| --- | --- |
| Percent change training R squared | 100.00% |
| Percent change testing R squared | 100.10% |
| Percent change training MAE | 615.57% |
| Percent change testing MAE | 180.10% |
| Percent change training MSE | 1642.05% |
| Percent change testing MSE | 174.13% |

* + Final model is fit on dataset

|  |  |
| --- | --- |
| R squared | 0.9495 |
| MAE | 2,379.54 |
| MSE | 60,360,773.59 |

* + The R squared value is very close to 1, suggesting that this model is a good fit for the data. The MAE and MSE values are also quite low, with the MAE of 2,379.54 suggesting that on average, you could expect to estimate a patient's yearly medical encounters cost within about 2,500 dollars.
  + Finally, the top 20 features are listed in order of importance.
    - num\_encounters
    - num\_procedures
    - DALY
    - Glomerular filtration rate/1.73 sq M.predicted [Volume Rate/Area]
    - Leukocytes [#/volume] in Blood by Automated count
    - Hematocrit [Volume Fraction] of Blood by Automated count
    - Body mass index (BMI) [Ratio]
    - Pain severity - 0-10 verbal numeric rating [Score] - Reported
    - AGE
    - meds\_cost
    - Urea nitrogen [Mass/volume] in Blood
    - Chloride [Moles/volume] in Blood
    - Cholesterol in HDL [Mass/volume] in Serum or Plasma
    - Potassium [Moles/volume] in Blood
    - num\_meds
    - Triglycerides
    - Carbon dioxide total [Moles/volume] in Blood
    - QALY
    - Creatinine [Mass/volume] in Blood
    - STATE\_POPULATION
  + Unsurprisingly, the number of medical encounters and number of procedures are the top two features, as these are expected to be correlated with the cost of medical encounters. However, some other interesting factors include DALY (disability-adjusted life year, a metric used to measure overall disease burden of a population), BMI (body mass index), pain severity, age, and number/cost of medications. Some of these factors are out of a patient's control, such as age, but it would be beneficial to investigate which factors could be altered through lifestyle interventions to reduce a patient's medical encounters costs.

Conclusion

* Final summary
* Things to improve/keep for next time
  + Patients – take into account, address, healthcare expenses and coverage
  + Encounters – take into account type of encounter/reason, payer coverage
  + Medications – take into account type of medication/reason, payer coverage, dispenses
  + Procedures – take into account type of procedure/reason
  + Immunizations – take into account type of immunization
  + Allergies – take into account type of allergy, reaction, severity
* Future work
  + Make dataframe of 2022 data, predict costs for 2023, compare to 2023