Emergency Department Patient Segmentation Using Unsupervised Machine Learning Techniques

Amanda D. Hanway

Western Governors University

Table of Contents

A. Project Highlights

A.1 Research Question or Organizational Need

This project addressed the research question, “What shared characteristics are present among patients with emergency encounters?” This question was brought about by the organizational need for healthcare payers to identify individuals with a higher propensity to experience an emergency encounter. Armed with this insight, the healthcare payer can take pre-emptive action to prevent the encounter as a cost-savings measure.

A.2 Scope of Project

The scope of this project included the identification, collection, preparation, and optimization of a patient dataset for a machine learning model, the application of the model on a separate dataset, and the analysis of results. The creation of a database to store the data and the development of a machine learning pipeline were out of scope for this project.

A.3 Overview of Data Analytics Solution

The data analytics solution included the use of Exploratory Data Analysis (EDA), Principal Component Analysis (PCA), Unsupervised Machine Learning, and aspects of project management. The python programming language and python libraries such as pandas, numpy, matplotlib, sklearn, and seaborn were used within a JupyterLab environment to conduct analysis and develop the model.

EDA was first conducted on a dataset of synthesized Electronic Health Records to understand its characteristics and identify underlying issues. Following the data cleaning and feature engineering steps identified through EDA, PCA was applied to the data to investigate and select important features. The data was then trained using the K-Means Clustering model. The initial model was developed, then tuned in iterations using the Agile project management

methodology. The final model was applied to a second, unseen dataset, and the results were evaluated.

B. Project Execution

B.1 Project Plan

The goal of this project was to develop an unsupervised machine learning model to cluster patients who had an emergency encounter based on their medical history and demographics. There were no variances from the original plan to achieve this goal, outlined in Task 2. All of the following objectives and deliverables established in Task 2 were met.

- **Objective 1:** Identify and collect a dataset with synthetic patient-level data that can simulate what would appear in Electronic Health Records.
    - **Deliverable 1:** A dataset with patient-level data, including encounters, conditions, allergies, care plans, and demographic information.
- **Objective 2:** Conduct a thorough Exploratory Data Analysis to gain an understanding of the dataset and the cleaning and feature engineering steps required.
    - **Deliverable 1:** A list of issues identified for cleaning and feature engineering.
    - **Deliverable 2:** Charts and tables with descriptive analytics about the dataset.
- **Objective 3:** Conduct Principal Component Analysis on the dataset to understand the variance accounted for by each component.
    - **Deliverable 1:** A Principal Component Analysis with a bar chart displaying Explained Variance Ratio and a scree plot displaying Cumulative Explained Variance Ratio.

- **Deliverable 2:** An optimal selection for the number of components to retain for the model, with reasoning.

B.2 Project Planning Methodology

The Agile methodology was used throughout this project. Agile allowed for quick development of a working prototype, followed by iterative improvements. There were no variances from the stages defined in Task 2, as follows:

1. **Requirements:** This step involved identifying a dataset appropriate to answer the business question, and researching and choosing the machine learning algorithm that was leveraged to develop a model.
2. **Development:** This step involved writing python code to load the data files into dataframes for exploratory analysis, cleaning, feature engineering, and model development.
3. **Testing:** This step involved reviewing the data after cleaning and initially running the model to ensure no data had been dropped unintentionally and that the model results were appropriate.
4. **Delivery:** This step involved conducting an internal review of the model to determine improvements that should be made.
5. **Feedback:** For this project, the feedback was determined in the form of improvements that should be made to the data or the model for the next round of updates.

B.3 Project Timeline and Milestones

The project timeline and milestones established in Task 2 concluded one week earlier than scheduled. The Model Development, Model Application, and Analysis steps did not take the full duration planned, which allowed the rest of the project to move along earlier than scheduled.

Milestone or Deliverable	Projected Start Date	Anticipated End Date	Actual Start Date	Actual End Date
Data Identification & Collection	*1/27/2024*	*1/29/2024*	*1/27/2024*	*1/29/2024*
Data Cleaning & Feature Engineering	*1/30/2024*	*2/12/2024*	*1/30/2024*	*2/12/2024*
Principal Component Analysis	*2/13/2024*	*2/15/2024*	*2/13/2024*	*2/15/2024*
Model Development	*2/16/2024*	*2/20/2024*	*2/16/2024*	**2/18/2024*
Model Application	*2/21/2024*	*2/23/2024*	**2/19/2024*	**2/20/2024*
Analysis of Results	*2/24/2024*	*3/1/2024*	**2/21/2024*	**2/24/2024*

C. Data Collection Process

C.1 Data Selection and Collection

The data selection and collection process followed the plan outlined in Task 2. The data was downloaded from the Synthea website (SyntheticHealth, 2024) in two zipped folders which contained 16-18 files each. After reviewing the files individually in Excel, the five following files were chosen for further use in the model:

1. allergies.csv
2. careplans.csv
3. conditions.csv
4. encounters.csv
5. patients.csv

C.2 Data Collection Obstacles

There were no obstacles encountered in the data collection process. The files were downloaded as planned, reviewed in Excel, and then loaded into dataframe using the python programming language for further cleaning and review.

C.3 Data Governance, Privacy, and Security

There were no unplanned data governance issues encountered. As designed, no data governance, privacy, or security concerns were identified for this project due to the use of synthetic data.

C.4 Advantages and Limitations of Data Set

C.4.A Advantages

Working with medical data requires utmost attention to compliance to protect the privacy of the individual. As such, data for real patients is not readily available to the public. An advantage of the Synthea dataset was that it provided synthetic but realistic patient data to mimic what would be found in a true healthcare setting, eliminating any concerns around compliance.

This dataset was advantageous in that it provided a complete 360 degree view of individual patients that would not have otherwise been available. The machine learning model had access to learn from a variety of details about each patient, from county of residence, age, and ethnicity, to a full medical history of allergies, conditions, and types of encounters.

C.4.B Limitations

While an advantage of the Synthea dataset was that of being synthesized, this was also a limitation. Given the dataset was not comprised of real patient medical histories, it may not truly

represent patterns in patients who have emergency encounters. The model generated from the dataset is not a reflection of how a group of real patients would be segmented. Rather, it represents segments only within the scope of the Synthea data.

An example of a disadvantage in this dataset was that it was comprised of data only for one state, for patients of largely the white race/ethnicity. To see a full picture of characteristics that may indicate an individual will have an emergency encounter (according to the Synthea algorithm), a larger dataset including a more diverse population would be an important next step to improve the model.

D. Data Extraction and Preparation

D.1 Data Extraction

The Synthea data was openly available for download from https://syntheticahealth.github.io/synthea-sample-data/. The datasets were appropriately collected by downloading the following zipped folders:

- Synthea_sample_data_csv_apr2020.zip – contained 16 csv files
- Synthea_sample_data_csv_nov2021.zip – contained 18 csv files

Once the zipped folders were downloaded and the files extracted, each file was reviewed in Excel to gain an understanding of contents and format. It was determined that the five following files would be adequate for developing a model to segment patients, and these files were subsequently loaded into pandas dataframes, using the python programming language within JupyterLab.

1. allergies.csv
2. careplans.csv
3. conditions.csv
4. encounters.csv
5. patients.csv

D.2 Data Preparation

D.2.A Exploratory Data Analysis

After loading the five chosen files from the 2020 dataset into individual dataframes, EDA was conducted to gain an understanding of its contents. The pandas info() method was run on each dataframe to generate descriptive statistics, including total rows, total columns, and counts of non-null values in each column. Then, additional investigation into each dataframe was conducted as follows:

- Bar charts were created using matplotlib to understand the percent of total patients by gender, race, ethnicity, and county of residence.
- The pandas describe() method was used to generate summary statistics for the healthcare expenses column.
- The pandas nunique() method was used to understand the total number of unique conditions, allergies, and unique patients with encounters, conditions or allergies in the dataset.
- The pandas groupby() function was applied to summarize counts of rows and patients by condition, encounters, and allergies.

D.2.B Data Cleaning & Feature Engineering

During the exploratory data analysis phase, the data was found to be primarily of good quality and completeness, values appeared to be standardized, and an adequate count of patients to utilize for this analysis was available. However, some work would be needed to prepare the data for further analysis. The cleaning and engineering steps required to transform data elements into data types appropriate for a machine learning model were as follows:

- Removed patients who did not have an emergency encounter
- Summarized counts of each encounter class per patient
- Summarized conditions per patient
- Created a column for patient age

- Identified and removed outliers for patient age
- Created patient age numerical categories
- Created healthcare expense numerical categories
- Created gender numerical categories
- Created race numerical categories
- Created county numerical categories
- Created a careplan indicator
- Created allergy one-hot encoded columns
- Joined datasets together into one main dataset
- Assessed and handled missing data
- Identified and removed outliers for emergency encounters

These steps were performed using appropriate logic and were primarily implemented with pandas methods and functions. As a final step in the data preparation phase, a reusable cleaning function comprised from all steps was created for use on the second dataset.

E. Data Analysis Process

E.1 Data Analysis Methods

E.1.A Exploratory Data Analysis

EDA was the first analytic technique performed on the dataset. This was a crucial first step that revealed insight into the file contents: number of data points available, data formats, patient demographics, types of conditions, allergies, and encounters represented. Information gained from EDA determined the data cleaning and feature engineering steps that would be required before further developing the model.

E.1.B Principal Component Analysis

PCA was performed after cleaning the data and engineering features. The use of PCA was appropriate in this project because the dataset, containing over 150 features, was moderately dimensional. Through PCA, an optimal number of components which explained most of the variance in the dataset was identified. Reducing the number of dimensions prior to training the machine learning model dealt with issues that could arise from the "curse of dimensionality".

E.1.C K-Means Model

The K-Means clustering algorithm was selected to perform unsupervised machine learning. K-Means was an appropriate choice for this project because it segments data based on shared characteristics, which addresses the project's goal to understand shared characteristics among individuals who have emergency encounters.

E.2 Advantages and Limitations of Tools and Techniques

E.2.A Exploratory Data Analysis

- **Advantages:** EDA was advantageous in that it revealed information about the dataset, such as dataset size, format, and presence of outliers. This was a crucial first step to understand the contents of the dataset and determined the cleaning and feature engineering steps needed before further developing the model.
- **Limitations:** A limitation of EDA was that it took up a large amount of time for this project. Even so, this stage could have been extended to further investigate and identify nuances in the dataset that were not already revealed.

E.2.B Principal Component Analysis

- **Advantages:** An advantage of PCA was that it identified the features in the dataset that had the greatest impact on variability. Removing the lesser features helped to reduce the noise in the dataset and allowed the model to better identify the underlying structure in the data. Additionally, removing the less important features reduced the computational resources required of the model, which would make this model scalable to larger datasets.
- **Limitations:** A limitation of PCA is that it may have resulted in loss of information that would have affected the model output. Reducing the components from 152 original variables to 80, which explained around 80% of the variance in the dataset, removed variables that explained the remaining 20%. These variables may have informed the model on additional patterns, and some potential clusters may have been missed as a result.

E.2.C K-Means Model

- **Advantages:** An advantage of the K-Means model was that it was quickly implemented and the results were relatively easily visualized and analyzed. Implementation time and explainability are important aspects when working with a time-dependent organizational need, as well as with stakeholders and partners outside of the realm of data analytics.
- **Limitations:** A limitation for the K-Mean algorithm is that a parameter for the resulting number of clusters must be defined before training the model. Selecting for an optimal number of clusters was a multi-step process. One method used to select for clusters was the elbow method, which was not very effective for this dataset as it was difficult to determine an optimal elbow point. Identifying the cluster parameter required running the

model on a range of clusters, picking the one with the best score, then training the final model using the chosen number.

E.3 Application of Analytical Methods

E.3.A Exploratory Data Analysis

The EDA phase involved generating both programmatic and visual assessments of each dataset to gain an understanding of its contents and to identify underlying issues. The detailed applications of EDA were as follows:

- **Programmatic Assessments:**
    - The pandas info() method was run on each dataframe to generate descriptive statistics, including total rows, total columns, and counts of non-null values in each column.
    - The pandas describe() method was used to generate summary statistics for the healthcare expenses column.
    - The pandas nunique() method was leveraged to understand the total number of unique conditions, allergies, and unique patients with encounters, conditions or allergies.
    - The pandas groupby() function was applied to understand counts of rows and patients by condition, encounters, and allergies.
- **Visual Assessments:**
    - Bar charts were created to understand the percent of total patients by gender, race, ethnicity, and county of residence, the count of patients with each type of encounter or allergy, and count of patients with a careplan.
    - A histogram was created to show the distribution of patients by healthcare expense.
    - A few rows of each dataset were printed to visually assess the formatting for each column.

E.3.B Principal Component Analysis

PCA was performed in combination with feature scaling. After the data cleaning and feature engineering steps concluded, the following steps were taken to scale features and conduct PCA:

- **Feature Scaling:** Feature scaling was a critical step before PCA. By rescaling all features to a standard deviation of 1 and a mean of 0, this step ensured that variances subsequently calculated in PCA were not influenced by differences in scale. The scikit-learn StandardScaler class was utilized.
- **PCA:** The scikit-learn PCA class was applied to the scaled data. The explained variance ratios were plotted in a bar chart and the cumulative explained variance ratios were plotted in a scree plot to investigate the results. After reviewing the charts, 80 components, which explained 80% of the variance, were chosen as an optimal number to retain for the model. The data was again run through PCA using the parameter of 80 components to be retained. The features making up the first three principal components were then extracted and their weights were examined to generate an analysis.

E.3.C K-Means Model

The K-Means clustering algorithm was leveraged to perform unsupervised machine learning. The scikit-learn KMeans class was fit to the PCA data for a range of one to 30 clusters, and the resulting Sum of Squared Errors (SSE) scores and Silhouette Scores were extracted for each number of clusters. The SSE for each model was plotted in a line chart. The models were evaluated using Elbow Criterion on the SSE chart, in combination with the Silhouette Scores. Based on the results, three clusters were chosen as the optimal number of clusters for the final model. The final model was fit to the PCA data again, using the parameter of three clusters.

F. Data Analysis Results

F.1 Statistical Significance

The model developed in this project was an unsupervised machine learning model using the K-Means clustering algorithm. The metrics to assess performance were Sum of Squared Errors (SSE), Elbow Criterion, and Silhouette Score. A Silhouette Score over 0.5 was the benchmark established to determine a strong model.

SSE and Elbow Criterion were used in combination with Silhouette Score to select for an optimal number of clusters for the model. SSE was calculated after training the model for a range of one to 30 clusters, then plotted in a line chart for each range of clusters. Elbow Criterion was applied to the chart to investigate an optimal number of clusters, and the highest Silhouette Score (0.29) was identified for three clusters. After selecting three clusters as the optimal number based on this analysis, the final model was fit again using three clusters as the parameter.

This Silhouette Score of 0.29 did not achieve the benchmark for a strong model. Rather, the score indicates the model clustered data points that were of a fair match to other data points in the same cluster. The conclusion drawn from the score supports the hypothesis in that the model begins to segment patients into shared characteristics, however there is still room for further work to be done to refine and improve the model.

F.2 Practical Significance

Practical significance was determined by the model's ability to identify patterns among patients who had emergency encounters. The model identified three clusters of patients:

- Cluster 0 was driven by lack of or minimal presence of allergies, Hypertriglyceridemia disorder, Metabolic syndrome X, Suspected lung cancer or Carcinoma in situ of prostate.
- Cluster 1 was driven by presence of allergies.

- Cluster 2 was driven by presence of Hypertriglyceridemia disorder or Metabolic syndrome X.

Using this information in a practical setting, healthcare payers equipped with a better understanding of individuals who have emergency encounters can create preventative action plans targeted toward each cluster. For example, individuals sharing characteristics of Cluster 1 may be sent information on managing allergies and preventing allergy attacks that may result in an emergency encounter. Preventing visits to the Emergency Department generates cost-savings for the payer, while improving the overall health of its members.

F.3 Overall Success

This project's overall success was measured on the completion of the defined goals, objectives, and deliverables. The project was successful in that a machine learning model was developed to cluster patients based on shared characteristics. In fulfilment of this goal, all objectives - to identify and collect an appropriate dataset, to conduct a thorough exploratory analysis of the dataset, and to conduct Principal Component Analysis on the dataset - along with their associated deliverables, were completed.

The results generated by the machine learning model indicated there is room for improvement. Although the model did not achieve the benchmark for a strong model, it performed at a fair level and resulted in somewhat-defined clusters of patients. Further work can be done to investigate and implement methods to tune the model.

G. Conclusion

G.1 Summary of Conclusions

The goal of this project was to identify shared characteristics among patients who had an emergency encounter. A patient dataset was identified, collected, cleaned, and then underwent Principal Component Analysis to select for an optimal number of components. Following PCA, the K-Means model was trained on a range of clusters, and then assessed to select for an optimal number of clusters. The final model was then applied to a new dataset for evaluation.

The machine learning model performed at a fair level with a Silhouette Score of 0.29, indicating the model clustered data points that were of a reasonable (but not strong) match to other data points in the same cluster. The conclusion drawn from the score suggests there is still room for further work to be done to refine and improve the model.

Even so, applying the trained model to the second patient dataset identified interesting patterns among the population. The largest cluster was Cluster 0 with 677 patients, followed by Cluster 1 with 71 patients and Cluster 2 with 52 patients. The following prominent characteristics were identified for each cluster:

- Cluster 0 was driven by lack of or minimal presence of allergies, Hypertriglyceridemia disorder, Metabolic syndrome X, Suspected lung cancer or Carcinoma in situ of prostate.
- Cluster 1 was driven by presence of allergies. All patients in this cluster had at least five allergies.
- Cluster 2 was driven by presence of Hypertriglyceridemia disorder or Metabolic syndrome X. Most of the patients in this cluster had at least one of these conditions.

G.2 Effective Storytelling

Throughout the project, the matplotlib library was leveraged to create numerous data visualizations. Data visualizations were a key method to quickly communicate underlying patterns in the data and to make decisions on how its intricacies should be handled.

- Bar charts and histograms were used throughout EDA. These charts supported effective storytelling by quickly displaying the dataset's make up in terms of demographics and conditions, as well as identifying outliers in the data.
- A bar chart and a scree plot were created to investigate variance during PCA. These charts supported effective storytelling by visually depicting the variance for over 150 features, facilitating the selection of PCA components.
- A line chart was used to analyze the SSE by cluster when identifying an optimal number of clusters for the model. This chart supported effective storytelling by plotting each point in comparison to all other points, enabling the use of Elbow Criterion to evaluate the results.
- Scatter plots were used to analyze the resulting cluster predictions. These charts supported effective storytelling by making the results easily understandable for audiences outside of the realm of analytics, through use of color to visually group datapoints into their respective clusters.

G.3 Recommended Courses of Action

The following courses of action are recommended in response to the outcomes of this project:

1. **Distribute Educational Materials:** The first recommended course of action is to utilize the model's initial results to fulfil the organizational need to identify shared characteristics among patients with emergency encounters. The model identified segments of the population based on the prominence of specific conditions, where Cluster 1 was driven by presence of allergies and Cluster 2 was driven by presence of Hypertriglyceridemia disorder or Metabolic syndrome X. The healthcare payer can generate informational material related to treating and managing these conditions, and distribute it to individuals who have had a diagnosis. Using email as a delivery method, this would be a low-cost approach to potentially prevent an expensive emergency encounter resulting from the condition going untreated.
2. **Continuous Improvement:** The second recommended course of action is to continue refining and tuning the machine learning model. Although the current model successfully segmented patients into clusters, it did not meet the benchmark for a strong model. Additional work is needed to investigate and remove outliers in the dataset and engineer additional features. Improving the model will help the healthcare payer to better address its organizational need to identify shared characteristics present among patients with emergency encounters.

H. Panopto Presentation

Panopto Link:

- [removed]

References

Hanway, A. (2024). Patient_segmentation. GitHub. https://github.com/mandi1120/patient_segmentation/blob/main/Capstone.ipynb

Synthetic patient generation. Synthea. (2024). https://synthetichealth.github.io/synthea/

Synthea-sample-data. SyntheticHealth. (2024). https://synthetichealth.github.io/synthea-sample-data/