Predicting Multiple Emergency Room Visits

Bellevue University - DSC 630-T302 - Project Group 1
Madhukar Ayachit, Kouevi Dosseh-Adjanon, Ryan Long

Table of Contents

Project Status

Milestones	Status	Notes
Team Information / Communication Plan	Complete	Weekly touchpoints on Teams and WhatsApp
Data Selection and Project Proposal	Complete	Problem statement defined, dataset obtained, initial exploratory data analysis performed
Preliminary Analysis	Complete	Initial model developed, problem statement further refined, and report drafted
Project Presentation and Status	Complete	Further refinement of model, selected features, presentation drafted
Final Project Paper and Presentation	Work in Progress	Delegate the following: Final Paper (Executive Summary and Technical Report), recording of presentation.

Background

Overview

- In the healthcare industry, emergency room (ER) visits represent one of the highest cost medical services.
- Every year, many patients have to file bankruptcy mainly due to increasing hospital and medical bills mostly made up of ER visits which lead to hospitalizations.
- Although most ER visits are warranted and have saved countless lives, a considerable amount of ER visits are avoidable and could be addressed with a visit to a primary physician or even an urgent care facility, most of which have lesser cost involved.
- It is likely individuals who utilize the emergency room more than 4 times per year could be doing so unnecessarily.
- Several healthcare focused entities from hospitals to insurance providers have had to implement measures to help prevent unnecessary emergency room visits.
- Not only do these unnecessary visits cost patients thousands of dollars out of pocket, but a significant portion is paid by healthcare insurance providers.
- These companies are now focusing on developing clinical outreach programs to get to patients in time in order to avoid these unnecessary costs.

Problem Statement

Overview

The problem to be evaluated will focus specifically on Emergency Room (ER) utilization data to help build a predictive model to understand the likelihood of a patient returning to the ER more than 4 times per year.

Scope

- The data, model development, and deployment of the results of this project will focus explicitly on repeated ERs visits.
- The dataset leveraged was obtained from a leading government sponsored health care provider in the United States of America and contains demographics, various metrics, and associated categorical information of healthcare patients who visited the ER over the course of a 12 month period.
- Consideration of the time period, demographic information, specific to the geographical location, and primary activities of ER visits define the boundaries and application of the model, interpretation of the results, and subsequent deployment.

Challenges and Issues

- Limitations of the data and model prohibit the use for predicting the likelihood of more than 4 ER visits per year outside of the USA or for other healthcare services.
- Additionally, pandemic conditions must be considered as there has been a material decrease in ER visits during the COVID-19 pandemic, primarily due to potential patients avoiding the risk of exposure to the virus (6).

Methods

Technical Approach

- A machine learning model with the goal to predict patients who are likely to over-utilize the emergency room or have more than 4 ER visits in a year could be a step toward reaching these patients ahead of time to help them avoid unnecessary costs.
- This could be translated into a machine learning classification solution where algorithms such as a logistic regression, a gradient boosted decision tree, and others can be fit to the data to help determine the model with the best performance.
- The programming languages Python and R were chosen for this project due to their ease of use, modeling capability, and visualizations.
- The JupyterNotebook and RStudio IDEs were chosen because of the open source nature of the software and supportive community of specialists.

Data Overview

- To help build the model, we've acquired healthcare data from the leading government sponsored healthcare provider in the U.S. The available attributes include medical and pharmacy claims as well as demographic variables such as gender, age, and location.
- Overall, the dataset includes 69K records on patients over the previous 12 months, containing 46 features, with "MORE_THAN_4_ER_VISITS" identified as the target.

Methods continued

Handling Null/Missing Values

- The dataset contains 20 Numerical features which contain either NaN or missing values.
- Additionally, there are 15 categorical variables in the dataset.
- We used LabelEncoder to transform categorical features into numeric values.
- After review, there are 11 features with an average 65 null values along with "Member_Months_Pre" with 2 and "ORCA_SCORE" being highest with 3400 null values in it. We have decided to replace null values with their median values instead of deleting the records completely.

Methods continued

Data Exploration: Data exploration started with looking into population and distribution of target feature i.e. “More_Than_4_Er_Visits”. Of the total 69K observations, 32K records indicated patients who had more than 4 ER visits versus 37K with less than 4 ER visits.

Outlier detection : The calculation of a Z score, or how many standard deviations a number is away from the mean, was used to detect outliers in the dataset. As a standard practice our threshold value was “3” standard deviations. Any record with 3+ Z score was marked as an outlier and replaced with the median value of that feature.

Feature selection: We considered “correlation” to identify most suitable features for modeling. We took 37 top correlated features into consideration with scores starting from -0.27 to 0.56.

Model Preparation: Model preparation was done with “More than 4 Er” being a target variable and the remaining 36 being dependent features. The entire dataset had previously been converted in numeric format during preprocessing using Label Encoder and was ready for modeling.

Logistic Regression: The problem statement is focused on predicting whether a patient will either have 4 or more visits to an ER or not. As this is a binary outcome, a Logistic Regression algorithm was chosen for modeling

Revisiting Model: After running the defined model with 100% population, a summary of the model output was used to further fine tune on the basis of p-Value score. Two features: Country_Clean & Reg_Region_Desc were removed from feature list due to significantly higher p_value score.

Results

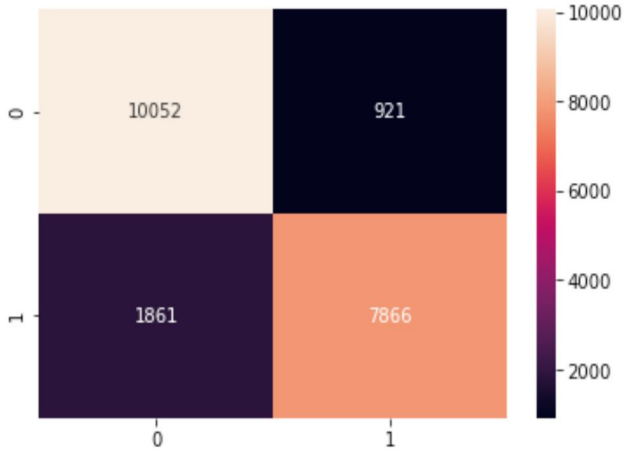
We are considering output of **Confusion Matrix** and **Classification Report** to conclude the exercise.

We have True positive and True negative combined 17918 versus 2782 being False positive and false negative on 20700 test observation.

Accuracy

In order to ensure that the Logistic Regression model performs well on new data, a portion of the initial dataset of 30% was set aside to serve as the testing sample. The remaining 70% of the dataset was used for training purposes. All iterations of the Logistic Regression based on the attributes and methods documented above showed 87% accuracy.

	precision	recall	f1-score	support
0	0.84	0.92	0.88	10973
1	0.90	0.81	0.85	9727
accuracy			0.87	20700
macro avg	0.87	0.86	0.86	20700
weighted avg	0.87	0.87	0.86	20700



Conclusion

Overall, the model developed showed favorable accuracy in the testing and training processes with the dataset available. Other metrics, such as precision, recall and f1 scores, also produced optimistic results towards the capability and potential applicability of the mode. Based on these results, this model could be used to predict the probability patients visiting the ER could return more than 4 times in a year and then potentially become readmitted to the hospital system.

When deployed, healthcare practitioners may input the same information and determine what level of care and remediation steps should be applied on a situational basis to reduce repeat visits and consequently limit impacts to the healthcare system. The only constraints would be on healthcare practitioners ability to collect the information used to create the model in addition to the scope limitations noted above.

Furthermore, after the model is deployed, ongoing monitoring should be put in place to ensure that the level of performance seen at training continues to hold true. This could require tracking actual outcome (or lack thereof) for a certain period of time and then compare these to the predictions made at the time. This will allow the project team to decide when it’s time to revisit the model and potentially re-train it if performance starts to degrade.

Next Steps

Perform the following:

- Final paper
  - Executive Summary
  - Technical Report
- Recording of presentation
- Final revisions and review

References

1. Brennan, J. J., Chan, T. C., Killeen, J. P., & Castillo, E. M. (2015). Inpatient Readmissions and Emergency Department Visits within 30 Days of a Hospital Admission. The western journal of emergency medicine, 16(7), 1025–1029. https://doi.org/10.5811/westjem.2015.8.26157

2. Wikimedia Foundation. (2021, June 8). Hospital readmission. Wikipedia. Retrieved September 10, 2021, from https://en.wikipedia.org/wiki/Hospital_readmission.

3. Hospital readmissions reduction Program (HRRP). CMS. (n.d.). Retrieved September 10, 2021, from https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Readmissions-Reduction-Program.

4. Emergency medical Treatment & Labor act (EMTALA). CMS. (n.d.). Retrieved September 10, 2021, from https://www.cms.gov/Regulations-and-Guidance/Legislation/EMTALA.

5. Morganti, K. G., Bauhoff, S., Blanchard, J. C., Abir, M., Iyer, N., Smith, A., Vesely, J. V., Okeke, E. N., & Kellermann, A. L. (2013). The Evolving Role of Emergency Departments in the United States. Rand health quarterly, 3(2), 3.

6. Hartnett, K. P., Kite-Powell, A., DeVies, J., Coletta, M. A., Boehmer, T. K., Adjemian, J., Gundlapalli, A. V., & National Syndromic Surveillance Program Community of Practice (2020). Impact of the COVID-19 Pandemic on Emergency Department Visits - United States, January 1, 2019-May 30, 2020. MMWR. Morbidity and mortality weekly report, 69(23), 699–704. https://doi.org/10.15585/mmwr.mm6923e1

7. van Baal, P., Morton, A., & Severens, J. L. (2018). Health care input constraints and cost effectiveness analysis decision rules. Social science & medicine (2018), 200, 59–64. https://doi.org/10.1016/j.socscimed.2018.01.026

8. Greenwald, P. W., Estevez, R. M., Clark, S., Stern, M. E., Rosen, T., & Flomenbaum, N. (2016). The ed as the primary source of hospital admission for older (but Not YOUNGER) adults. The American Journal of Emergency Medicine, 34(6), 943–947. https://doi.org/10.1016/j.ajem.2015.05.041

9. Tsai, M. H., Xirasagar, S., Carroll, S., Bryan, C. S., Gallagher, P. J., Davis, K., & Jauch, E. C. (2018). Reducing High-Users' Visits to the Emergency Department by a Primary Care Intervention for the Uninsured: A Retrospective Study. Inquiry : a journal of medical care organization, provision and financing, 55, 46958018763917. https://doi.org/10.1177/0046958018763917

10. Kacprzyk, A., Stefura, T., Chłopaś, K. et al. “Analysis of readmissions to the emergency department among patients presenting with abdominal pain”. BMC Emerg Med 20, 37 (2020). https://doi.org/10.1186/s12873-020-00334-x