Capstone project — Classifying hospital disposition of ER visits using Machine Learning

By: Yash Nagpaul
Data Science Diploma Candidate at BrainStation

Problem Space and Value Proposition

In June 2023, a study[1] was published in the Canadian Medical Association Journal with the title 'Emergency departments are in crisis now and for the foreseeable future'. This project is aimed at answering the following question: How might we use Machine Learning to efficiently decide whether or not a patient visiting the emergency room of a hospital needs to be admitted?

Background

With medical data and public health records being increasingly digitized all over the world, Data Science and Machine Learning has an ever more important role to play in finding efficient solutions for problems by spotting underlying patterns in the vast amounts of medical data that are extremely difficult (if not impossible) for humans to spot.

Details on dataset

The data used for this project was originally collected for a study[2] that was published in the Public Library of Science (PLoS) journal in 2018. The researchers who were trying to solve this problem have provided a de-identified, semi-preprocessed dataset of patient visits to the Yale New Haven Health system. The raw data used in the original study was obtained from electronic health records of hospitals and is not publicly available since health records are protected personal information of the patients. The de-identified data that they have provided was partially pre-processed, and some of the categorical variables had been converted into dummy variables. The data dictionary that was provided has also been included with this project.

The dataset has been broadly categorized into 10 categories by the researchers: Demographics, Triage variables, Hospital Usage, Chief complaint, Past Medical History, Outpatient medications, Imaging, Historical vitals, Historical labs, and the response variable (disposition).

Summary of cleaning and preprocessing:

We started working with a dataset of over 560,000 patient visits (each visit being a single row) and 972 unique features. By the end of the the EDA and Cleaning phase, we had:

- Learnt that there is an imbalance in the distribution of the response variable in our data (see Figure 1)

[1] https://www.cmaj.ca/content/195/24/E851
[2] https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0201016

- Ensured that there were no duplicate entries.
- Dropped all columns where over 10% of the values were missing (to prevent biased imputation without being a healthcare expert).
- Converted all 15 categorical columns into dummy variables.
    - Ultimately we were left with 683 columns to work with.
- Split up the data into training and test sets.
- Imputed the missing values using the median value of the column.
- Scaled the data so that any large variations arising due to different scales of measurements among the variables is normalized.

Value counts of unique entries in the response variable column

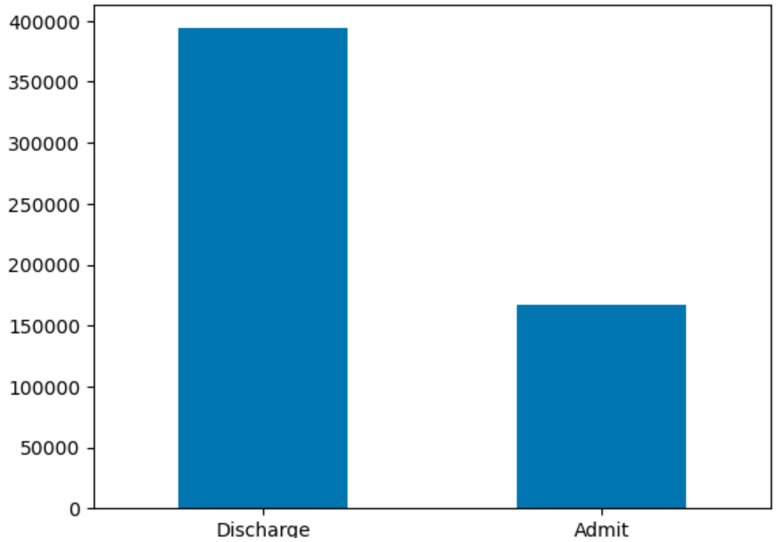

Fig 1. A chart showing the imbalance distribution of the response variable

Insights, modeling, and results:

We decided to go with two of the most popular models used for binary classification problems: Logistic Regression and XGBoost. Our baseline model was a Logistic Regression without hyperparameter optimization which was 86.04% accurate on unseen data. Before evaluating further, we ran a grid search using various hyperparameter options to find the best parameters for the model. The accuracy remained nearly identical (86.03%). Hence, we looked to find different measures of model performance. To account for the imbalance in our original dataset, we resampled our data (downsampled the majority class). The downsampling approach was opted since it would still leave us with over 200,000 data points to work with, and model iterations would be more efficient.

With a balanced dataset, the accuracy of the hyperparameter optimized model fell slightly to 84.21%. Since the difference was not very pronounced, we looked even further for more measures of model evaluation, i.e., Precision and Recall scores.

Ultimately, the model that was trained on the balanced dataset turned out to be the best one as it combined high accuracy (84%) along with **lowest false negative rate when predicting ‘discharge’** and a **high false positive rate when predicting ‘admit’** – which is what we would prefer to see in healthcare-related problems.

After optimizing the Logistic Regression model, we also fitted an XGBoost classifier which outperformed our best Logistic Regression model overall. The XGBoost had a similar overall accuracy of about 85% on unseen data, but we got a slightly better recall score (0.83 compared to 0.82 with the best Logistic Regression model) .

We also uncovered the top 10 predictive variables resulting in a disposition being ‘Admit’ which are discussed in the notebook (see image below).

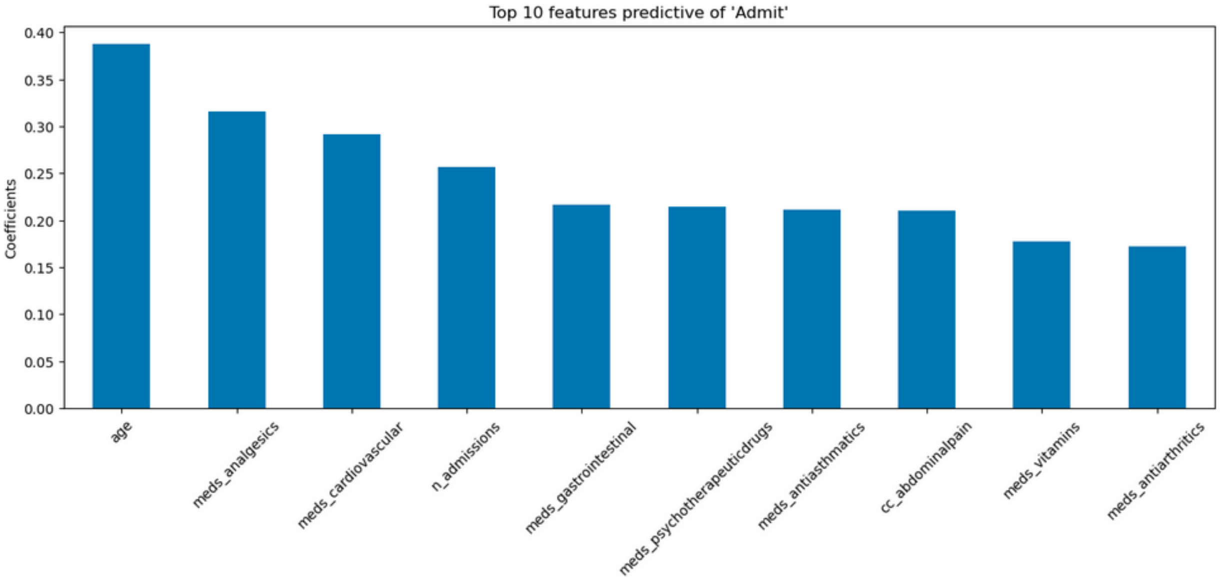

Fig 2. Top 10 features predictive of ‘Admit’ disposition

Findings and conclusions:

As a Data Scientist, I consider a model with an accuracy of > 85% to be a good model. The results we achieved were comparable to similar studies that have been done in this field, including the one[3] that inspired this project. However, the actual limitations of this project would come under light if we had involved a healthcare expert who would likely have been able to ask better questions as we made decisions throughout this project.

Another limitation of this project is that it uses a specific dataset from a healthcare system based in the United States. If there are big inconsistencies between how Canada and the United States store healthcare information, this project would likely need significant changes accordingly (especially in the pre-processing stage) for it to be applied in Canada.

[3] https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0201016