Data Science Bootcamp

Capstone project report

Prediction of Hospital Occupancy by Department

Ignacio Medina Fernández

LONDON, NOVEMBER 2022

Contents

1 Introduction

Unexpected patient loads and associated workload changes in hospitals can have a negative impact over the quality of care for patients and the occupational health of clinicians. Machine learning techniques are an increasingly attractive tool to solve these issues through data. Fortunately there is now widespread patient data thanks to the widespread use of Electronic Health Records (EHR). Thanks to the relatively structured and standardised nature of these data, such models can be adapted to better understand different environments than that of the original data source and make accurate predictions [1]. For this work, we have used the MIMIC-IV dataset [2], which consists of EHR-like retrospectively collected patient data from the Beth Israel Deaconess Medical Center (Boston, MA, USA). Data were collected from digital bedside monitors and contain records from 2008 to 2019. Prediction of patient outcome (mortality, readmission... etc) is fairly common, particularly using MIMIC datasets, but we have focused on prediction of hospital transfers with the aim of predicting occupancy for the different hospital care units. To the best of our knowledge there are no equivalent works in the literature using this approach. This type of predictions can help to improve patient care and better manage clinician workload spikes.

2 Data sourcing

The MIMIC-IV dataset is made-up of several tables corresponding to patient data obtained from the Emergency Department (ED), Intensive Care Units (ICU) and for the whole hospital. The "base" table for our study is the "transfers" table, where each row is a transfer containing mainly the subject (patient) ID, a hadm ID (hospital admission ID, a stay can be considered a series of consecutive transfers), a transfer start time and a stop time, and finally the care unit where the patient stayed during the transfer. Our goal is to predict the following transfer which is the care unit of the following row (after sorting the dataset) as an first step to predicting patient care unit in real time.

We have extracted features from several MIMIC-IV tables (prescriptions, medrecon, admissions, patients and edstays) to incorporate them into the transfers table. The data and the features were created, transformed and selected in an iterative way, testing different models to select the optimal combination of features and model hyperparameters.

2.1 Prescriptions data cleaning and feature extraction

Prescriptions data were extracted from two tables, "prescriptions" (hospital patients) and "medrecon" (ED patients). These two tables were combined and each row was assigned an Anatomical Therapeutic Classification (ATC) code [3]. This way we could decrease the number of features (there were 9000 unique medication), relating medication with similar therapeutic use and potentially extracting more relevant information. This classification system is hierarchical and two different sets of features were extracted: One corresponding to level 1 ATC (14 unique features) and another corresponding to level 2 ATC (86 unique features in dataset). These were One-Hot Encoded (OHE) and added to the table. Level 1 features were added first. Level 2 features were not added as they were poorly correlated to the event classes. These ATC categories were added using an NDC mapping tool on R [4] to RxNorm (a collection of normalised names for US medication). The NDC column, as well as the other drug identifier columns (drug name, formulary code and GSN code) were missing, so the categories were filled recursively using all drug columns. Essentially, all drugs with either the same NDC, formulary code, GSN code were given the same category. By repeating this step a few times, along with some manual filling, the missing values decreased from 18% to 2% (for level 1 ATCs) and 0.2% (for level 4 ATCs).

2.2 Transfers data cleaning and merging

The transfers base table was modified and cleaned. As the original table had the origin care unit and the destination care unit of a transfer in different rows, the destination event was added to the rows with the origin care unit. This was done by sorting by subject ID and start time and shifting back the event (destination) series. A similar approach was followed for data cleaning, which involved merging any transfers that had the same event as origin care unit or any short transfers with the following and preceding transfers. This was done because predictions for short transfers would be inaccurate (low number of prescriptions) and because 3 h before a patient was transferred, their destination was likely to have been planned already. Joining the patients and admissions table consisted of left joining these to transfers using the subject and HAdm IDs. Prescriptions ID was joined to transfers on subject ID (there were null values for HAdm ID column), before the rows were filtered to only keep those rows were the start time of the prescription was between the start and stop times of the transfer. This involved a loop using data chunks to avoid running out of memory.

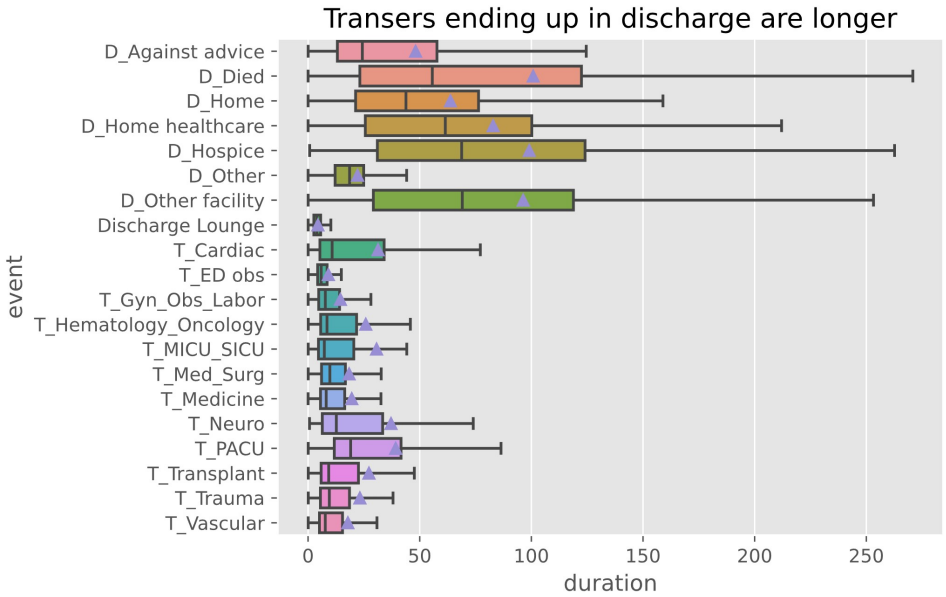


Figure 1: Distribution (box plots) of duration for transfers with different events. The average is shown as a purple triangle. Outliers not shown.

3 Exploratory Data Analysis

After completing the data sourcing stage, an EDA was carried out on 70% of the dataset (30% was kept hidden for evaluation). The first step was simplifying the classes. Similar care units were merged to decrease the number of classes (predicted outcomes) from 40 to 20. The relationship of the extracted features with these classes was then investigated. Differences in feature distributions for different classes were particularly evident for the duration and prescription features. Interestingly, duration, as seen in Figure 1, was much higher for rows with a discharge event. It seems like patients may be transferred to different departments before they end up in the department where they will spend most of their stay before being discharged. Understandably, the number of prescriptions was highly correlated with durations, although this relationship was different for discharge and transfer classes. Figure 2 shows how for transfers ending up in a discharge, prescription number increased more slowly with duration, while for the other transfers this relationship was steeper. These differences were also

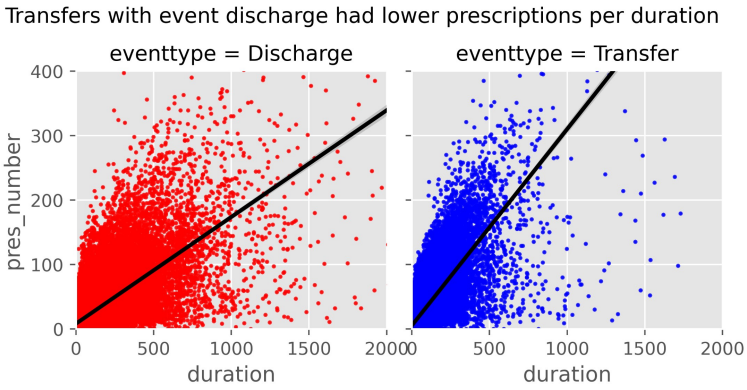


Figure 2: Scatter plot of prescription number and transfer duration for grouped events Discharge and Transfer. Least squared linear regression (using seaborn.regplot on python) fitted and shown as a black line.

clear when comparing the distribution of the two for all the classes. Demographic features proved to have less predictive power, but we could make some interesting observations. For example, distribution of gender for the different classes showed how men were sent to the transplants care unit more often than women, confirming statistics for the general population [5]. Other interesting curiosities was that young men were the most likely to leave care against advice and that up to 10% of patients in the gynecology/obstetrics care units were men, as gynecologists are apparently specialised in treating pelvic pain.

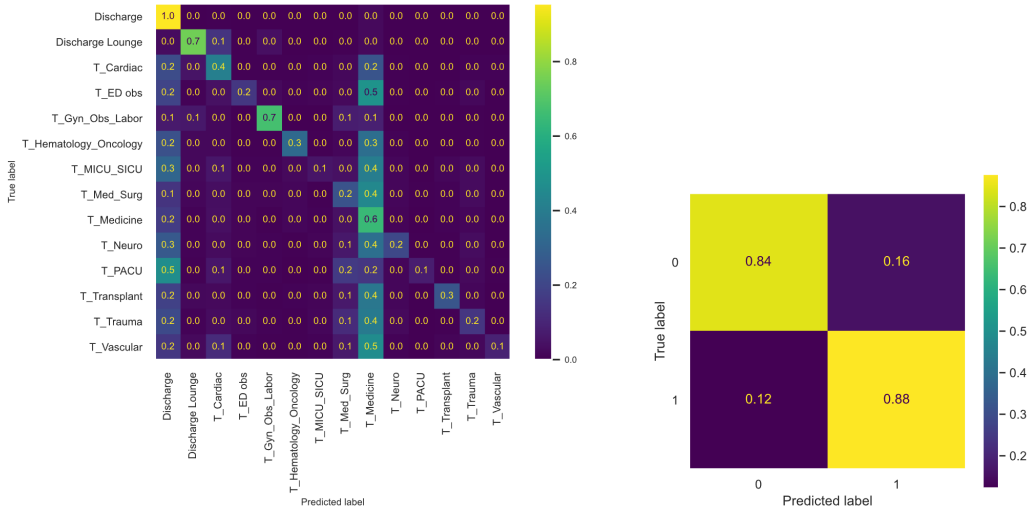
4 Model optimization

Once the initial feature had been prepared and characterised, we fitted several models on the training dataset. We used logistic regression and random forest models because they allow us to explore the data and infer relationships. The initial logistic regression models were fitted with the main purpose of data exploration. By plotting confusion matrices it seemed clear that the model tended to mix the predictions for the different discharge types, so we merged these (we are not interested in predicting discharge type anyway). This allowed us to decrease the number of classes. Partly due to the high class imbalance, the models tended to have a high recall for the discharge event (majoritary class), while failing to predict for the others. The coefficients of logistic regression were obtained and plotted to be used in future work during feature expansion.

PCA was avoided for feature selection as feature variance did not appear to explain well the variance in the events (the target). For this reason, random forest was used to try and select a subset of features to simplify our model. Before we reached this step, new features were added (past transfers) and random forest hyperparameters were tuned. Feature selection was tuned together with the models hyperparameter in a grid, although it appeared that any decrease in features had a significant effect on precission. This, coupled with low scores for most classes (Figure), suggests that the current set of features is not the most relevant for this problem, so for future work new features should be added from the ohter tables' columns (POE, procedures...). However, even if prediction of transfers was not completely possible, we could fit a binary model that predicted Discharge or Transfer with an accuracy of . The confusion matrices for these models are shown on Figure 3

5 Future work and conclusions

In this work we have extracted a number of features from an EHR-like database with the goal of predicting patient destination in a transfer. This is a preliminary work where the problem has been simplified to produce a



(a) Multiclass classification 1

(b) Binary classification 2

Figure 3: Confusion matrix for multiclass classification problems fitted with random forest (a) and binary classification problem (Discharge vs Transfer, b). Same features and similar modes were used for both.

proof-of-concept model. Future work will focus on reformulating the problem to predict the location of a patient in the hospital after a certain time. New features must be extracted from the dataset to improve the accuracy of the model and to adapt it to this problem. At the moment, we have produced a limited classifier for transfer rows with an accuracy score of 0.66 and a binary model (to predict transfer or discharge) with an accuracy score of 0.86.

References

[1] Alicia Curth et al. "Transferring Clinical Prediction Models Across Hospitals and Electronic Health Record Systems". In: *Machine Learning and Knowledge Discovery in Databases*. Ed. by Peggy Cellier and Kurt Driessens. Cham: Springer International Publishing, 2020, pp. 605–621. ISBN: 978-3-030-43823-4.

[2] Alistair Johnson et al. *MIMIC-IV*. 2021. DOI: 10.13026/s6n6-xd98. URL: https://physionet.org/content/mimiciv/1.0/.

[3] *ATC structure and principles*. https://www.whocc.no/atc/structure_and_principles/. Accessed: 2022-11-06.

[4] Fabricio SP Kury and Olivier Bodenreider. "Mapping US FDA National Drug Codes to Anatomical-Therapeutic-Chemical Classes using RxNorm." In: *AMIA*. 2017. URL: https://github.com/fabkury/ndc_map.

[5] F. Puoti et al. "Organ transplantation and gender differences: a paradigmatic example of intertwining between biological and sociocultural determinants". In: *Biol Sex Differ* 7 (2016), p. 35.