Biostatistics

Predicting the risk of emergency admission with machine learning

Gabriel Singer, MVA 2024

1 Introduction

A high-performance healthcare system requires, among other things, high-performance emergency departments capable of absorbing and effectively managing the influx of patients. The article we are studying [3] reports that in 2017 in the UK there was a 2.6 per cent increase in the number of people admitted to emergency departments. Today, according to the NHS [2], the number of patients having to wait more than 4 hours in emergency has reached historic records. Not to mention the case of emergency departments in France, which suffer terribly from staff shortages and are often (if not always) overwhelmed. Note also that over the last 20 years, the number of admissions to emergency departments in the USA has risen steadily [1].

With the emergence of artificial intelligence on a large scale, both in terms of computing capacity (Nvidia GPUs) and know-how (new MLops Software engenieer professions), the improvement of prediction models can be "scaled up" and lead to operational tools such as the ones used in [6]. In the literature, models have struggled to determine which specific risk might lead to admission to the emergency department. This suggests that the problem is multi-factorial, involving non-linear relationships between different predictor variables.

The studied article [3] investigates the following question:

Problematic

What is the probability that a given person will be admitted to the emergency room for the first time?

One specificity of the article is that it's focuses on **the first visit** to the emergency department. The aim is to compare a model classically used in the risk prediction sector with two machine learning models.

2 Modelisation

The very general framework of this problem belongs to the class of **supervised learning.**[1] Roughly speaking, the associated problem can be formulated as follows:

Mathematical formulation

Let $n \in \mathbb{N}^\star$ and $P = (p_1, ..., p_n) \in \mathbb{R}^n$ be a patient modelised by n variables. The dataset is made of q points $(P_k, \phi(P_k))_{1 \leq k \leq q}$ where $\phi : \mathbb{R}^n \mapsto [0, 1]$ who, to a patient associates the probability of that patient being admitted to the emergency department for the first time. Here $q = 4,637,29$ and $n \in \left\{ \underbrace{58}_{QA}, \underbrace{80}_{QA+}, \underbrace{121}_{T} \right\}$.

The goal is to approximate ϕ.

As benchmark model they use the Cox proportional hazards modelisation [6]. This reference model was then compared with the two very well knonw models: random forest (RF) [5] and gradient boosting classifier (GBC) [4]. Both GBC and RF models were used as ensemble models based on decision trees.

Since the model doesn't do everything in machine learning, it's important to pay attention to the data entered into the model.

The initial set of features for all models included 43 variables, such as patient demographics, lifestyle factors, laboratory tests, currently prescribed medications, selected morbidities, and previous emergency admissions. The authors then added 13 more variables such as marital status, prior general practice visits, and 11 additional morbidities. They also add a temportal dimension like time since first diagnosis, for instance.

The study analyzed linked EHRs from the UK's CPRD from 1985 to 2015, involving records from 674 practices covering 7 percents of the UK population, linked to hospital and mortality records. Utilizing a CPRD-approved subset without needing patient consent, it focused on **4.6 million eligible patients** out of 7.6 million, based on age, registration duration, and availability of NHS numbers and socioeconomic status.

To predict emergency admission risks, three sets of predictors were employed: QA with 43 variables from the QAdmissions model, QA+ adding 13 variables including marital status and new comorbidities, and T, introducing temporal elements to some QA+ predictors. Despite potential biases due to non-random missing data, particularly in BMI, smoking status, and alcohol intake, these were mitigated using binary indicators and multiple imputation, leading to **58, 80, and 121 variables** in the QA, QA+, and T sets respectively. The study confirmed the minimal bias from imputation through calibration plots (see plot 2), ensuring accurate prediction outcomes.

Note that RF and GBC are **robust** to missing data, making them good model choices, [12].

[1] Theoritically it is a supervised learning problem, however in practice, as we will see later it is closer to semi-supervised learning probleme. Since the considered models are able to handle missings values, it's "practically" not a problem.

3 Results

GBC consistently outperformed RF and CPH models on both, calibration and accuracy.

Let's start by analyzing the study of bias, and then we'll talk about calibration and performance.

The authors followed [8], they stratified AUC by practice and present the findings in a funnel plot by practice-level rate of emergency admissions. One of the main assumptions in [8] to be able to analyze the Funnel is that the sample size must be large enough. This is the case here. By the symmetric aspect of 1, there is no **bias** by practice population and admission rate for all practices in both the derivation and validation cohorts.

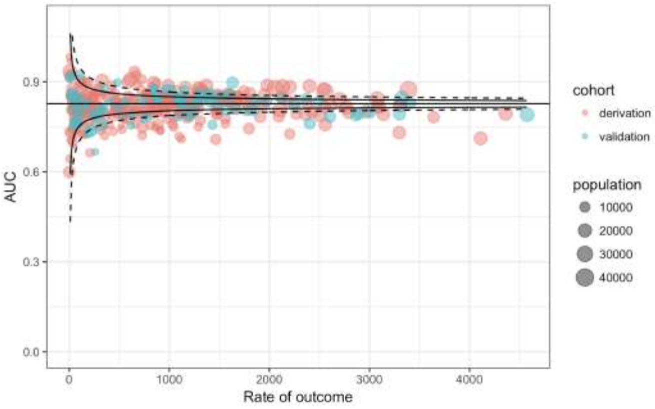


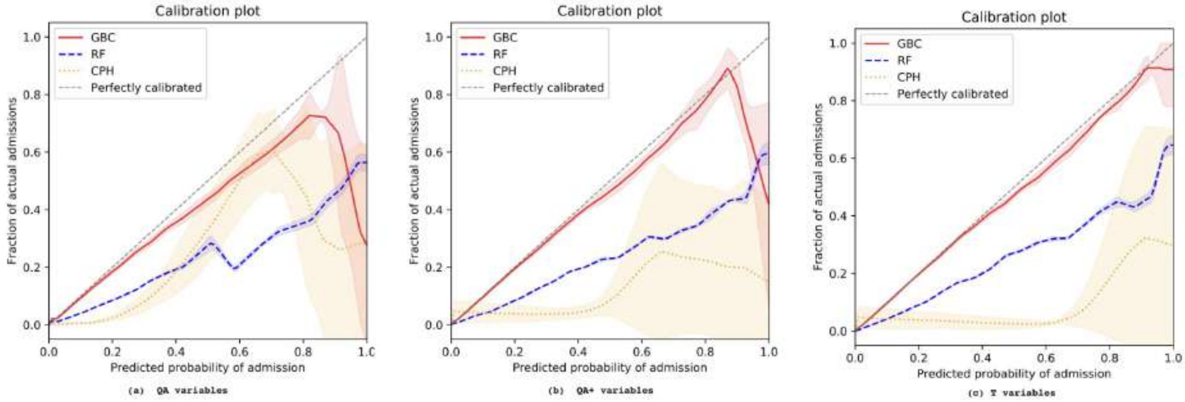
Figure 1: Funnel plot

Calibration is an often forgiven and missunderstood metric. In this hospital context calibration takes its full meaning. Calibration indicates how confident the model is in its prediction. In fact, the hospital operator enters certain patient characteristics into the software and the software displays a probability, (as in [6]).

Calibration Graph

As a reminder, the calibration graph is calculated as follows: first, the $[0, 1]$ axis is partitioned into sub-intervals of size 0.1. Suppose we have $N_1 \geq 0$ predictions with probabilities between 0 and 0.1. We take the average $x_1 = \frac{\sum Prob\in[0,0.1]}{N_1}$. Then we calculate the associated empirical probability y_1. This gives us a point $C_1 := (x_1, y_1)$. The calibration graph is defined as the set of those points.

Note that in all three cases and for all three models, their calibration curves are below the $y = x$ axis, so they're very sure of themselves. In the health sector, it's better to be cautious. So this is a good point. However, we can see that GBC (red curve) is by far the best calibrated. The addition of temporal variables improves model calibration, especially in areas of high admission probability(i.e $[0.8, 1]$). Note also that the funnel plot is corroborated by the calibration analyse of the models.



(a) QA variables

(b) QA+ variables

(c) T variables

Figure 2: Calibration des modèles BGC, RF, CPH sur les datasets QA, QA+,T.

In term of calibration, the GBC is the best, since it's calibration plot is the closest to $y = x$ and adding time variable really improove calibration.

Let's analyse the area under the receiver operating characteristic curve (AUC) score.

Initial results showed GBC with (AUC) of 0.779, compared to 0.752 for RF and 0.740 for CPH. External validation further confirmed GBC's superiority, with AUCs of 0.796 for GBC, versus 0.736 for both RF and CPH. The inclusion of temporal information notably enhanced the AUC for all models in both internal and external validations. Remark that the respeciv AUC scores lives in an non over/under-fitting range (i.e $[0.7, 0.85]$).

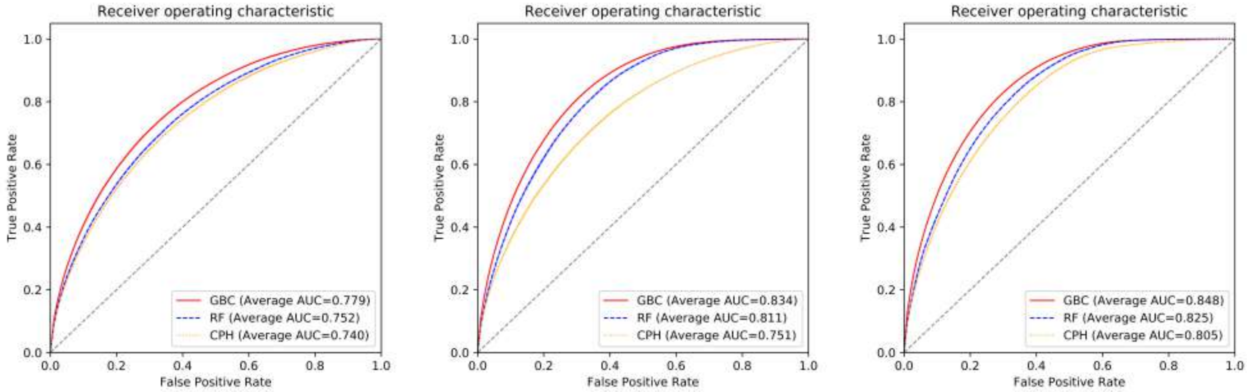
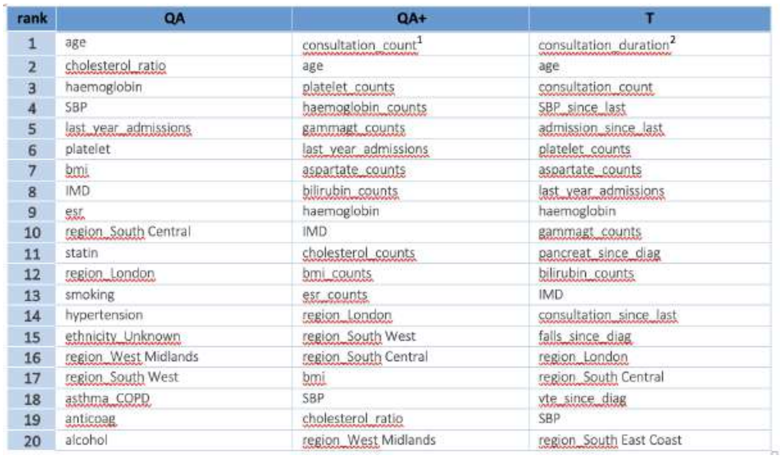


Figure 3: ROC curves respectively for QA, QA+ and T datasets, [3]

GBC outperforms also in term of AUC score.



rank	QA	QA+	T
1	age	consultation_count[1]	consultation_duration[2]
2	cholesterol_ratio	age	age
3	haemoglobin	platelet_counts	consultation_count
4	SBP	haemoglobin_counts	SBP_since_last
5	last_year_admissions	gammagt_counts	admission_since_last
6	platelet	last_year_admissions	platelet_counts
7	bmi	aspartate_counts	aspartate_counts
8	IMD	bilirubin_counts	last_year_admissions
9	esr	haemoglobin	haemoglobin
10	region_South Central	IMD	gammagt_counts
11	statin	cholesterol_counts	pancreat_since_diag
12	region_London	bmi_counts	bilirubin_counts
13	smoking	esr_counts	IMD
14	hypertension	region_London	consultation_since_last
15	ethnicity_Unknown	region_South West	falls_since_diag
16	region_West Midlands	region_South Central	region_London
17	region_South West	bmi	region_South Central
18	asthma_COPD	SBP	vte_since_diag
19	anticoag	cholesterol_ratio	SBP
20	alcohol	region_West Midlands	region_South East Coast

Figure 4: TOP 20 features for GBC

The increase in performance by incorporating temporal data is confirmed by table 4. The variables that imported the most during predictions are shown on the left for RF and on the right for GBC. In each table, there are three columns, each corresponding to one of the three datasets. We can see that, even in the QA case, the temporal variable (first admission) is the most important, and that in the T dataset, the top 5 is only made of temporal variables (e.g. consultation duration, last year admissions).

4 Strengths and weaknesses

Let's start with the strengths.

This modelisation differs from others in that there are a few studies concerning the modeling of admissions of people knowing that they have already been to hospital [7, 9] but none concerning **first admissions** to the emergency department. In addition, the database used is incomparably larger than those used in other articles. In particular, it contains data on people who have never been admitted to hospital, which is not the case with previous studies using inpatient data. This makes it possible to model a wider variety of cases. It also differs form the other since the authors inclueded **temporal variable**, which improoved classification results in the **BGC** case and even more important, they show (see top three lines of 4) that these predicators are more important than their binary counterparts.

BGC and RF were selected due to their proven superiority over other machine learning algorithms across **tabular datasets**. Their robustness and scalability make them suitable for large datasets (here the dataset lives into $\mathbb{R}^{120\times 4.10^6}$ for T dataset for instance). There is no big effort to do for the finetuning. These methods handles both **categorical and numerical data** of any scale without necessitating the conversion or normalization of feature values.

This is an **operational advantage.** What's more, they are relatively easy to **interpret**, enabling us to understand any decisions made by the model. Espacially in the case of a (little RF). The interpretability of machine learning models is a very important point when it comes to making decisions such as triaging patients in the emergency department.

A final positive point, which is particularly important in view of the growing number of cyber-attacks, is that, unlike deep-learning, GBC is robust in the face of numerous attacks.

The first point to note about the weaknesses is that it's not clear what the practical purpose is of predicting the risk of emergency room admissions on very large scales ($1 - 2 - 4 - 5$ years), for example with the figure 3 in the article studied.

The biggest weak point is about the sparsity of the data base. In fact, the **datasets are missing data** on certain variables that are sometimes important (both from a model point of view according to 4 and from an intuitive point of view), such as somking status, 30% of people are missing their smoking status.

They say that: limiting the number of variables taken into account in their models limits their performance. This is not entirely true, especially when it comes to operational production. Indeed, the addition of variables inevitably requires greater computational capacity (often exponential in the number of variables), a phenomenon known as the "**curse of dimensionnality**", [13]. And thus the hole process may become non-operational.

Bibliographie

[1] Emergency Admissions: Why Are They Growing So Fast? The Health Foundation, [Consulted: 14/02/2024]. Available on: https://www.health.org.uk/blogs/emergency-admissions-why-are-they-growing-so-fast.

[2] NHS England, A&E Attendances and Emergency Admissions, Monthly Time Series.

[3] Fatemeh Rahimian, Gholamreza Salimi-Khorshidi, Amir H. Payberah, Jenny Tran, Roberto Ayala Solares, Francesca Raimondi, Milad Nazarzadeh, Dexter Canoy, et Kazem Rahimi.

[4] Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001; 29(5):1189–232.

[5] Liaw A, Wiener M. Classification and regression by randomForest. R News. 2002; 2/3:18–22.

Predicting the risk of emergency admission with machine learning: Development and validation using linked electronic health records. *Deep Medicine, Oxford Martin School, Oxford, United Kingdom; The George Institute for Global Health, University of Oxford, Oxford, United Kingdom; Oxford University Hospitals NHS Foundation Trust, Oxford, United Kingdom*, [November 20, 2018].

[6] J. Hippisley-Cox et C. Coupland. Predicting risk of emergency admission to hospital using primary care data: derivation and validation of QAdmissions score. *BMJ Open*, 3(8):e003482, 2013. 10.1136/bmjopen-2013-003482 PMID: 23959760.

[7] Risk prediction models for hospital readmission: a systematic review. Kansagara D, Englander H, Salanitro A, Kagen D, Theobald C, Freeman M, et al. JAMA. 2011; 306(15):1688–98. DOI: 10.1001/jama.2011.1515, PMID: 22009101.

[8] Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ. 1997 Sep 13;315(7109):629-34. doi: 10.1136/bmj.315.7109.629. PMID: 9310563; PMCID: PMC2127453.

[9] Predicting the likelihood of emergency admission to hospital of older people: development and validation of the Emergency Admission Risk Likelihood Index (EARLI). Lyon D, Lancaster GA, Taylor S, Dowrick C, Chellaswamy H. Fam Pract. 2007; 24(2):158–67. DOI: 10.1093/fampra/cml069, PMID: 17210987.

[10] Doctor AI: predicting clinical events via recurrent neural networks. Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. arXiv. 2015 Nov 18 [cited 2017 Jul 28]. Available on: http://arxiv.org/abs/1511.05942.

[11] Children's AE Attendances Available on: https://fingertips.phe.org.uk/indicator-list/view/iYi2ex7my0.

[12] scikit-learn: Machine Learning in Python

[13] Bellman R.E. Adaptive Control Processes. Princeton University Press, Princeton, NJ, 1961.

[14] Xingjun Ma, Yuhao Niu, Lin Gu, Yisen Wang, Yitian Zhao, James Bailey, Feng Lu. "Understanding Adversarial Attacks on Deep Learning Based Medical Image Analysis Systems." *Journal of Medical Imaging and Health Informatics*, vol. 13, no. 5, 2023, pp. 1082-1091.