Projet data mining (SD 3A)

Envoi du rapport et du code : 10/01/2024 à 23h59

Date de soutenance : 12/01/2024

Projet : Triage des patients aux urgences

Remarque : vous devez faire en sorte d'utiliser TOUTES les connaissances pertinentes acquises pendant les cours de data mining du BUT Sciences des Données (2A et 3A). L'utilisation de connaissances pertinentes acquises dans d'autres matières sera très appréciée.

Contexte

Bien que la plupart des visites aux urgences se terminent par une sortie, les urgences représentent la source la plus importante d'admissions à l'hôpital. À leur arrivée aux urgences, les patients sont d'abord triés en fonction de leur état de santé afin de donner la priorité aux personnes nécessitant une intervention médicale urgente. Ce processus appelé "triage", est généralement effectué par un membre du personnel infirmier sur la base des données démographiques, du principal symptôme et des signes vitaux du patient. Par la suite, le patient est examiné par un autre personnel médical qui élabore le plan de soins initial et recommande finalement une décision : l'admission du patient ou sa sortie de l'hôpital.

Le jeu de données à utiliser dans ce projet est le résultat d'une étude rétrospective qui a inclus toutes les visites de personnes adultes aux urgences (3 en tout, une salle d'urgence universitaire et deux salles d'urgence communautaires) entre mars 2014 et juillet 2017 qui ont abouti à une admission ou à une sortie. Au total, 972 variables ont été extraites par patient.

Objectif du travail

Explorer et extraire de la connaissance à partir du jeu de données disponible sur le lien ci-dessous.

Données disponibles : les données à extraire sont contenues dans un fichier Rdata se trouvant à **https://www.kaggle.com/datasets/maalona/hospital-triage-and-patient-history-data et vous pouvez disposer sur github des scripts R pour manipuler les données (https://github.com/yaleemmlc/admissionprediction)**

Des données rétrospectives ont été obtenues auprès de **trois services d'urgence (Emergence Department)**, couvrant la période de mars 2013 à juillet 2017.

Les services d'urgence représentés comprennent un centre de traumatologie de niveau I avec un recensement annuel d'environ 85 000 patients, un service hospitalier communautaire avec un recensement annuel d'environ 75 000 patients, et un service autonome de banlieue avec un recensement annuel d'environ 30 000 patients. Les trois services d'urgence font partie d'un système hospitalier unique utilisant le système Epic EHR et l'indice de gravité des urgences (ESI) pour le triage.

Les différents niveaux de l'indice ESI sont numérotés de un à cinq, le niveau 1 indiquant la plus grande urgence. Cependant, les niveaux 3, 4 et 5 ne sont pas déterminés par l'urgence, mais par le nombre de

ressources qui devraient être utilisées. Ce triage est déterminé par un/e infirmièr/e expérimenté/e. Pour plus d'information sur l'ESI visitez la page https://en.wikipedia.org/wiki/Emergency_Severity_Index.

L'étude a porté sur toutes les visites de patients adultes pour lesquels la décision d'admission ou de sortie était clairement consignée. Les personnes ayant une autre disposition, comme le transfert, qui nétaient pas d'accord avec l'avis médical (AMA – Against-Medical-Advice) et la fugue, ont été exclues.

Comme déjà souligné ci-dessus, pour chaque visite de patient, 972 variables au total ont été collectées réparties dans les grandes catégories présentées dans la table suivante :

Category	Number of Variables	Only Triage	Only History	Full
Response variable (Disposition)	1	X	X	X
Demographics	9	X	X	X
Triage evaluation	13	X		X
Chief complaint	200	X		X
Hospital usage statistic	4		X	X
Past medical history	281		X	X
Outpatient medications	48		X	X
Historical vitals	28		X	X
Historical labs	379		X	X
Imaging/EKG counts	9		X	X
Total	972	223	759	972

Variable de réponse (Response variable). La principale variable de réponse était le devenir du patient, encodé dans une variable binaire (1 = admission, 0 = sortie).

Évaluation du triage (Triage evaluation). L'évaluation du triage comprenait les variables recueillies systématiquement au triage, telles que le nom de l'hôpital, l'heure d'arrivée (mois, jour, tranche de 4 heures), la méthode d'arrivée, les signes vitaux au triage et le niveau ESI attribué par l'infirmière de triage. Les signes vitaux de triage comprenaient la pression artérielle systolique et diastolique, le pouls, la fréquence respiratoire, la saturation en oxygène, la présence d'un dispositif d'oxygénothérapie et la température. Les valeurs au-delà des limites physiologiques ont été remplacées par des "valeurs manquantes".

Plainte/symptôme principal(e) (Chief complaint). Étant donné le nombre élevé de valeurs uniques (> 1 000) pour la plainte/symptôme principal(e), les 200 valeurs les plus fréquentes, qui représentaient > 90 % de toutes les visites, ont été retenues comme catégories uniques et toutes les autres valeurs ont été classées dans la catégorie "Autres".

Statistiques d'utilisation de l'hôpital (Hospital usage statistic). Le nombre de visites aux urgences en l'espace d'un an, le nombre d'admissions en l'espace d'un an, l'issue de la précédente visite du patient aux urgences et le nombre d'interventions et de chirurgies figurant dans le dossier du patient au moment du rendez-vous ont été considérés comme des indicateurs de l'utilisation antérieure de l'hôpital.

Antécédents médicaux (Past medical history). Les codes ICD-9 pour les antécédents médicaux (PMH) ont été mis en correspondance avec 281 catégories cliniquement significatives à l'aide du logiciel de classification clinique (CCS) de l'Agency for Healthcare Research and Quality (AHRQ), de

sorte que chaque catégorie CCS est devenue une variable binaire avec la valeur 1 si le PMH du patient contenait un ou plusieurs codes ICD-9 appartenant à cette catégorie et 0 dans le cas contraire.

Médicaments (Outpatient medications). Les médicaments utilisés en ambulatoire et répertoriés dans le EHR comme actifs au moment de la rencontre avec le patient ont été classés en 48 sous-groupes thérapeutiques (par exemple, cardiovasculaires, analgésiques) utilisés en interne par le système EHR Epic, chaque variable correspondante représentant le nombre de médicaments dans ce sous-groupe.

Données vitales historiques (Historical vitals). Un délai d'un an à compter de la date de la rencontre avec le patient a été utilisé pour calculer les informations historiques, qui comprenaient les signes vitaux, les examens de laboratoire et les examens d'imagerie commandés précédemment dans l'un des trois services d'urgence (Emergency Department – EDs). Les signes vitaux historiques étaient représentés par le minimum, le maximum, la médiane et la dernière valeur enregistrée de la tension artérielle systolique, de la tension artérielle diastolique, du pouls, de la fréquence respiratoire, de la saturation en oxygène, de la présence d'un dispositif d'oxygénothérapie et de la température. Les valeurs au-delà des limites physiologiques ont été remplacées par des “valeurs manquantes”.

Laboratoires (Historical labs). Compte tenu de la diversité des analyses demandées au sein du service d'urgence (ED), les 150 analyses les plus fréquentes, représentant 94 % de toutes les demandes, ont été extraites, puis divisées en analyses avec des valeurs numériques et en analyses avec des valeurs catégorielles. Le seuil de 150 a été choisi pour inclure des examens suffisamment fréquents pour être significatifs dans la prise en charge de la plupart des patients (par exemple, Troponin T, BNP, CK, D-Dimer), même s'ils n'étaient pas aussi fréquents que les examens de routine tels que la NFS, la BMP et l'analyse d'urine (CBC, BMP, and urinalysis).

Le minimum, le maximum, la médiane et la dernière valeur enregistrée de chaque laboratoire ont été inclus comme caractéristiques. Les résultats des analyses d'urine et des cultures, ont été recodés en variables binaires avec 1 pour toute valeur positive et 0 dans le cas contraire. Toute croissance dans les hémocultures a été qualifiée de positive, de même que les cultures d'urine avec > 49 000 colonies/mL. Le nombre de tests, le nombre de positifs et la dernière valeur enregistrée ont été pris en compte.

Imagerie et électrocardiogramme (Imaging and EKG counts). Le nombre d'ordonnances a été compté pour chacune des catégories suivantes : électrocardiogramme (EKG), radiographie du thorax, autres radiographies, échocardiographie, autres échographies, tomodensitométrie de la tête, autres tomodensitométries (CT), Magnetic resonance imaging (MRI) et toutes les autres imageries.

Langages de programmation : R (les données à extraire sont disponibles dans un fichier Rdata) et Python (le traitement des données sera effectué dans le langage Python).

Rendu : Rapport (20 pages)

Evaluation : Rendu + soutenance