Predicting Future Frequent Users of Emergency Departments in California State

Mayana Pereira
Center for Data Science
Institute of Technology
University of Washington
Tacoma
mayanaw@gmail.com

Vikhyati Singh
Center for Data Science
Institute of Technology
University of Washington
Tacoma
singhv3@uw.edu

Chun Pan Hon
Center for Data Science
Institute of Technology
University of Washington
Tacoma
darrencp@uw.edu

T. Greg McKelvey*
KenSci
Seattle
greg@kensci.com

Shanu Sushmita
Center for Data Science
Institute of Technology
University of Washington
Tacoma
sshanu@uw.edu

Martine De Cock†
Center for Data Science
Institute of Technology
University of Washington
Tacoma
mdecock@uw.edu

ABSTRACT

A large percentage of emergency department (ED) visits originates from a small percentage of patients who keep returning to the ED. Being able to flag these frequent users in advance can help clinicians to take appropriate interventions to reduce the number of ED visits, thereby reducing cost and improving quality of care. In this paper we present machine learning models that can predict future ED utilization of individual patients, using only information from the present and the past. We train decision trees (DT), boosted decision trees (AdaBoost) and logistic regression (LR) models on discharge records from California-licensed hospitals from the years 2009 and 2010, and evaluate their predictive accuracy for the years 2011-2013. We also study the impact of including different groups of demographic, frequency of ED visits, distance to emergency department, and clinical features on the accuracy of our predictive models. Overall there are three key findings of this study. First, all three techniques (LR, DT and AdaBoost) have a strong predictive ability to discriminate frequent ED users (number of visits $\geq$ 5). Second, our models show consistent outcomes across all three test years in our dataset, which is a desired property when a predictive tool that is stable and consistent year over year is required. Third, *least* and *most* frequent ED users are comparatively easier to predict when compared to *moderate* ED users (with higher sensitivity and AUC scores).

1. INTRODUCTION

According to a recent National Health Statistics Report, every year around 20% of adults in the U.S. seek healthcare at an emergency department (ED) [7]. A significant fraction of the frequent ED user population is comprised of patients with complex unmet needs for interventions, including mental health, substance abuse, transportation and housing, and chronic disease management services [1]. Being able to identify these frequent ED visitors in advance is valuable as it can enable targeted interventions that can keep the frequent ED visitors out of the ED, thereby increasing quality of care and, given the high costs of ED care relative to office-based care, potentially reducing cost [10, 18].

A variety of studies have been done to analyze reasons for emergency room use. These are typically retrospective studies where logistic regression models are fit on historic patient data [1, 12, 16]. In a recent study, Wu et al. trained logistic regression models that classify patients as frequent or non-frequent ED visitors in 2009-2010 given their registration data from 2008 [20]. Their approach assumes that the number of ED visits in 2009-2010 is already known for some of the patients, so that it can be used to train classifiers that infer the number of ED visits in 2009-2010 for the other patients. Inspired by this work, in this paper we train and test machine learning models that make predictions about future ED utilization, using *only* information from the present and the past. To this end we use a California Office of Statewide Health Planning and Development (OSHPD) dataset with discharge records from California-licensed hospitals for the years 2009-2013. We use the data from 2009-2010 to train our models, and we evaluate their predictive accuracy for the years 2011-2013. Unlike in the study by [20], no data from the years for which we make predictions is used in training the models that make the predictions.

We train and test predictive models that, for a given patient, predict whether the patient will be a low frequency user ($\leq$ 1 ED visit), a medium frequency user (2-4 ED vis-

*Occupational & Environmental Medicine Fellow
University of Washington

†Guest professor at Ghent University

BCB '16, October 02-05, 2016, Seattle, WA, USA

DOI: http://dx.doi.org/10.1145/2975167.2985845

its), or a high frequency user ($\geq$ 5 ED visits) in the coming year, hence treating the problem as a 3-class classification task. We consider four groups of features: (1) number of hospital admissions and ED visits of the patient in the previous year; (2) age, gender and race of the patient; (3) distance between the patients' home and the EDs visited in the previous year; and (4) comorbidities of the patient and severity of the diagnoses. Gradually expanding our feature set with these groups of features allows us to investigate their impact on the accuracy of the predictive models.

This paper is structured as follows: after presenting related work in Section 2, in Section 3 we describe the datasets that we created for our study. This includes a description of the cohort of patients with ED visits that we selected from the original OSHPD data, as well as a description of the features we constructed. In Section 4 we describe our predictive models (decision trees, boosted decision trees, logistic regression), while in Section 5 we present an extensive evaluation of their predictive accuracy and an analysis of the impact of different feature groups. For completeness, we also include a comparative analysis with a series of logistic regression models trained as in [20] for the binary classification task of inferring whether the number of ED visits of a patient will be below or above a varying threshold. We conclude in Section 6 with interesting directions for future work.

2. RELATED WORK

Medical facilities like EDs are designed to provide episodic care to patients suffering from serious injuries and illnesses. These facilities also cater to patients who are experiencing a sudden increase of underlying chronic medical conditions which require immediate attention [12]. Emergency care constitutes a significant and growing proportion of the practice of emergency medicine. ED overcrowding represents an emerging threat to patient safety and could have a significant impact on the critically ill [4]. For instance, ambulance diversion is a frequent reaction to ED crowding, which may carry consequences including delayed patient transport and lost hospital revenue [9]. In addition, the overall percentage of the population that has an ED visit is decreasing due to the occurrence of frequent and repeated use by a portion of the population [1]. The ED overcrowding problem results in calls for ways to identify avoidable ED visits [8, 9].

A variety of research efforts have focused on understanding the utilization and factors leading to ED crowding by frequent users (see e.g. [17]). The majority of these efforts were either retrospective cohort studies [1, 16], or regression estimations [12]. Recently attention has turned towards the use of machine learning techniques to predict ED usage. For instance, [11] used optimization techniques for feature selection, and combined them with an optimization-based discriminant analysis model (DAMIP) to identify a classification rule with relatively small subsets of discriminatory factors to predict return of patients within 72 hours to an ED. Logistic regression is a prominent and preferred method in the literature on predicting frequent ED users, including [13, 19, 20]. [19] used logistic regression to predict the need for immediate hospital admission that is likely to follow an episode of ED care. The utility of early prediction of hospital admission among ED patients is that it may help identify patients deserving of early admission planning and resource allocation and thus potentially reduce ED overcrowding. [20] focused on identifying frequent ED users in the subsequent two years, and [13] predicted frequent ED use among rural older adults. Lastly, [2] made efforts on predicting the number of ED admissions (for those patients that require a bed and thus represent a demand on bed management), on any given day of the year, taking into account peak periods such as public holidays.

Our research differs from existing work in that we predict the frequency with which patients will visit the ED in the coming 12 months, using *only* information from the previous 12 months. While existing work focuses on the use of logistic regression for the binary classification task of distinguishing frequent from non-frequent users, we study the 3-class classification problem of "bucketing" patients into low, medium and high frequency users. In addition to logistic regression, we use non-linear methods, namely decision trees and AdaBoost to this end.

3. COHORT

We requested non public data for the years 2009-2013 from the California Office of Statewide Health Planning and Development (OSHPD)[1]. The data consists of hospital discharge records that include patient demographic information, such as age, gender, and race/ethnicity, and clinical and administrative information, such as diagnosis and the route by which the patient was admitted (e.g. the admitting hospital's Emergency Room).

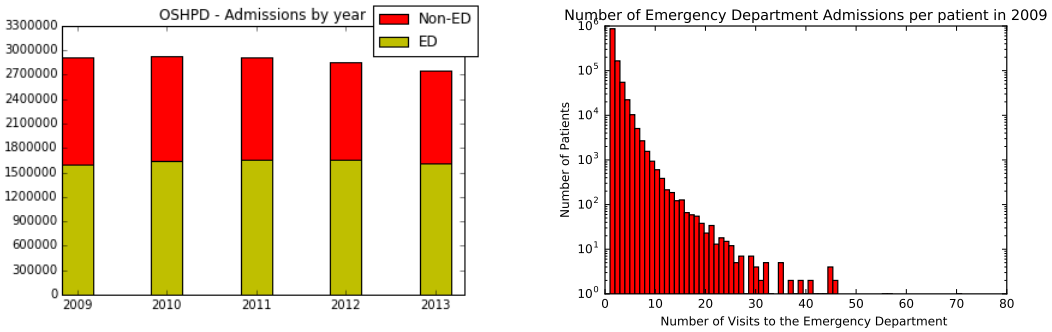
3.1 Data Pre-processing and Cohort Selection

The original 2009-2013 OSHPD dataset is a collection of patient admission records in tabular format. Each row corresponds to one discharge record of one patient. There are a total of 15,140,658 records and 7,355,726 unique patients, with discharge dates from 2009 to 2013. 18,634 patients have an inconsistent or missing gender value in at least one of their records. Similarly, 372,852 patients have an inconsistent or missing value for their race group in at least one of their records. We cleaned inconsistencies in gender and race across different records for the same patient by assigning the most frequently occurring gender and/or race for that patient to all his records. We filled in missing gender and race values in a similar way. This affected 22,260 rows for gender, and 526,972 rows for race group.

Invalid zip codes were found in patients and hospitals. Out of the 465 hospitals, 2 have invalid zip codes. We retrieved the correct zip codes by looking up the hospital names by the hospital ids from OSHPD's website[2], and then looking up the correct zip codes on the hospitals' websites. There are 67,887 patients with invalid zip codes in at least one of their admissions. We replaced an invalid patient zip code in a given record with the patient zip code from the previous record of the same patient if it exists and is valid. If there is no previous record or the zip code of the previous record is invalid, then we consulted the next record instead. However, there are 24,886 patients having no valid zip code in all of their records. The records for these patients were fixed by replacing the zip code with the most frequently occurring hospital zip codes across all admissions

[1]http://www.oshpd.ca.gov/HID/Data_Request_Center/documents/DataDictionary_Nonpublic_PDD.pdf/

[2]www.oshpd.ca.gov/hid/data_request_center/documents/app_d_facility-status.xlsx



(a) (b)

Figure 1: (a) Number of admissions per year; (b) Histogram of number of ED admissions per patient in 2009

of the patient.

Some records for the same patient have overlapping admission periods. Some of these are "same day readmissions", i.e. the next admission date is on the same day as the current discharge date. Others can be described as "readmission before discharge", i.e. the next admission date is before the current discharge date. Both same day readmissions and readmissions before discharge are indicative of a patient being transferred to a different care facility within the same hospital or to a different hospital. We consider two records with overlapping admission periods part of the same hospitalization, and merged them into a single record. This brings the total number of discharge records down to 14,406,870. Our assumption that all same day admissions are planned transfers may lead to under counting of the true rate of ED utilization. The effect is minimal though since less than 4% of the records are affected.

Figure 1 (a) shows the breakdown of the hospitalizations according to admission year, as well as how many of them were ED visits. 40,895 records had an admission date in 2008 and are not included in the further study.

We selected four cohorts, corresponding to the patients with at least one ED admission in 2009, 2010, 2011, and 2012 respectively. Table 1 shows the number of patients in each cohort, as well as the percentage of male/female patients, and the percentage of patients in each age group (see Section 3.2). Figure 1(b) shows a histogram of the number of ED visits per patient in 2009. There were 869,157 patients with exactly 1 ED admission in 2009 (not displayed in the figure), and 265,344 patients with more than 1 ED admission. The histograms for the other years are similar.

3.2 Feature Construction

For each cohort (year) we created a dataset with one row per patient. Besides the patient id or "record linkage number" (RLN) we also included ED_ADMIT: the total number of ED visits of the patient in that year; ALL_ADMIT: the total number of hospital admissions of the patient in that year; GENDER: male, female; AGE: age at the first ED visit of the year, discretized into six subgroups 0-4, 5-14, 15-24, 25-44, 45-64, 65+; and RACE-GRP: White, Black, Hispanic, Asian/Pacific Islander, Native American/Eskimo/Aleut, Other, Unknown/Invalid/Blank.

In addition, as in [20], we constructed features to capture the proximity of an ED. The original OSHPD dataset contains the patient zipcode (patzip) and the hospital zipcode (hspzip). To compute the distance between the patient's home and the hospital, we converted all zipcodes into latitudes and longitudes[3]. For each patient we include DIST1: the percentage of ED visits within a distance of 5 miles ($dist \leq 5$), DIST2: the percentage of ED visits in a distance between 5 and 20 miles ($5 < dist \leq 20$), and DIST3: the percentage of ED visits at a distance greater than 20 miles ($dist > 20$). To train our models (see Section 4) we only use DIST1 and DIST3 as DIST2 = 100 − DIST1 − DIST3.

The original OSHPD data contains an MS-DRG (Medicare Severity Diagnosis Related Groups) based feature that for each hospital admission indicates the presence/absence of a complication/comorbidity (CC) or a major complication/comorbidity (MCC). Similar as above, for each patient we include the following three features in our dataset: SEV0: the percentage of ED visits with no presence of CC or MCC; SEV1: the percentage of ED visits with presence of CC; and SEV2: the percentage of ED visits with presence of MCC.

Finally, we mapped the ICD9 diagnosis codes from the ED discharge records into 30 categories corresponding to the Elixhauser comorbidities [5]: `CHF` (Congestive heart failure), `VALVE` (Valvular disease), `PULMCIRC` (Pulmonary circulation disorder), `PERIVASC` (Peripheral vascular disorder), `HTN` (Hypertension, uncomplicated), `HTNCX` (Hypertension, complicated), `PARA` (Paralysis), `NEURO` (Other Neurological), `CHRNLUNG` (Chronic pulmonary disease); `DM` (Diabetes w/o chronic complications), `DMCX` (Diabetes w/ chronic complications), `Hypothy` (Hypothyrodism), `RENFAIL` (Renal Failure), `LIVER` (Liver disease), `ULCER` (Chronic peptic ulcer disease[4]), `AIDS` (HIV and AIDS), `LYMPH` (Lymphoma), `METS` (Metastatic cancer), `TUMOR` (Solid tumor without metastasis), `ARTH` (Rheumatoid arthritis[5]), `COAG` (Coagulation deficiency), `OBESE` (Obesity), `WGHTLOSS` (Weight loss), `LYTES` (Fluid and electrolyte disorders), `BLDLOSS` (Blood loss anemia), `ANEMDEF` (Deficiency anemia), `ALCOHOL` (Alcohol Abuse), `DRUG` (Drug abuse), `PSYCH` (Psychoses), `DEPRESS` (Depression), and an additional category `UNCLASS` for ICD9 codes that don't correspond to any comorbidity. Next, for each comorbidity, we computed the percentage of ED visits of the patient in which an ICD9 code for that comorbidity was observed, and included it as a feature. Note that the sum of

[3]using postal code data from GeoNames geographical database. http://www.geonames.org/

[4]includes bleeding only if obstruction is also present

[5]includes also collagen vascular diseases

COHORT		GENDER		AGE					
	#patients	%male	%female	%0-4	%5-14	%15-24	%25-44	%45-64	%65+
ED2009	1,134,501	46.99	53.01	1.31	1.52	4.08	13.95	27.72	51.38
ED2010	1,156,901	47.07	52.93	1.31	1.45	4.11	13.99	27.97	51.19
ED2011	1,159,662	47.31	52.69	1.15	1.31	4.16	13.92	28.23	51.20
ED2012	1,153,188	47.57	52.43	0.92	1.27	4.25	14.32	28.58	50.63

Table 1: Statistics about the four selected cohorts. The cohort for a given year corresponds to the patients who had at least one ED visit during that year.

name	description	features
	patient id	RLN
FREQ	frequency	ED_ADMIT, ALL_ADMIT
DEM	demographics	GENDER, AGE, RACE-GRP
DIST	distance	DIST1, DIST3
MED	medical	SEV0, SEV1, SEV2, 31 Elixhauser comorbidity features
	response variable	ED_ADMIT_NEXT

Table 2: Overview of the feature groups

these percentages for the same patient across all comorbidities is not necessarily 100.

Table 2 contains an overview of all the features described in this section, split across four feature groups: "FREQ" refers to the number of hospital admissions and ED visits of the patient in the previous year, "DEM" refers to the patient's demographic characteristics, "DIST" refers to the distance between the patient's home and the EDs visited in the previous year, and "MED" refers to the comorbidities of the patient and severity of the diagnoses.

4. METHODS

We trained supervised machine learning models using the FREQ, DEM, DST, and MED features (as shown in Table 2) constructed from the discharge records from 2009 as input features, and the corresponding number of ED visits in 2010, i.e. ED_ADMIT_NEXT, as the response variable to be predicted. We studied two versions of the prediction task: (1) a *3-class classification task*, where the goal is to classify patients as low frequency (≤ 1 visit), medium frequency (2-4 visits) and high frequency (≥ 5 visits) ED users, (2) *a binary classification task*, with the aim to distinguish between low frequency ($< p$ visits) and high frequency ($\geq p$ visits) ED users as in [20]. We varied the threshold p between 2 and 9 and trained corresponding models for each threshold.

The training data is highly imbalanced: 1,020,325 (i.e. 89.93%) of patients from the 2009 cohort have ≤ 1 ED visit in 2010, 97,219 (i.e. 8.56%) have 2-4 ED visits, and only 16,957 (i.e. 1.49%) have ≥ 5 visits. We therefore undersampled the 2009 cohort by randomly selecting 15,000 patients from each group. The advantage of undersampling highly imbalanced medical data was previously observed in [15]. All 3-class classification models are trained on a sample of 15,000:15,000:15,000 patients from ED2009. The binary classification models are trained on undersampled data as well. For each binary model, we balanced our datasets in order to have an equal number of positive and negative class labels patients (see Table 3). We observed that the choice of sample did not significantly influence the results, so we sampled once for each p (for the binary case) and once for the 3 class case, using the same sample for all models.

p	$< p$ visits	$\geq p$ visits	sample size
2	1,020,325	114,176	100,000:100,000
3	1,078,624	55,877	50,000:50,000
4	1,104,660	29,841	25,000:25,000
5	1,117,544	16,957	15,000:15,000
6	1,124,413	10,088	10,000:10,000
7	1,128,173	6,328	6,000:6,000
8	1,130,278	4,223	4,000:4,000
9	1,131,591	2,910	2,900:2,900

Table 3: Number of patients from the 2009 cohort with $< p$ visits and with $\geq p$ visits in 2010, and sample size of the training data for the binary classification tasks. All 3-class classification models are trained on a sample of 15,000:15,000:15,000 patients from ED2009.

We trained three kinds of models for the binary as well as the 3-class classification tasks, namely decision trees (DT), boosted decision trees (AdaBoost) and logistic regression (LR).

Decision trees are known to be robust and expressive models. The top-down algorithms for growing decision trees can naturally handle binary as well as multi-class classification problems. The leaf nodes can refer to either of the K classes concerned. For this study, we used an implementation of the classification and regression tree algorithm (CART) in R [3]. We built our classification trees using as complexity parameter cp=0.001. The complexity parameter defines how the splits are made in the decision tree; a split is only made if it decreases the overall lack of fit by at least a factor of cp.

Boosted decision trees are ensembles of trees. They are trained using a boosting process in which each subsequent tree is built with weighted instances which were misclassified by the previous tree [6]. Like stand-alone decision trees, these ensembles of trees can naturally handle binary as well as multi-class classification problems. Classification of a new instance with a trained ensemble of trees is based on a simple majority vote of the individual trees. For this study, we used the *adabag*[6] implementation of boosted decision trees in R with 50 rounds of boosting.

Logistic regression is a discriminative classifier that models the posterior probability $P(Y|X)$ of the class Y given the input features X by fitting a logistic curve to the relationship between X and Y. As such, logistic regression model

[6]https://cran.r-project.org/web/packages/adabag/adabag.pdf

outputs can be interpreted as probabilities of the occurrence of a class [14]. The class decision for the given probability is then made based on a threshold value. The threshold is often set to 0.5, i.e. if $P(Y = c^+|X) \geq 0.5$, then we predict that the instance belongs to the positive class, and otherwise we predict the instance belongs to the negative class. For the binary classification tasks, we used R's standard *glm* function. For the 3-class classifications tasks we trained multinomial log-linear models using neural networks with the nnet[7] package in R.

Next, we present and discuss the results obtained using these methods for the 3-class classification and binary classification tasks.

5. RESULTS

Three levels of result analysis are described in this section: (1) the impact of various feature groups on model performance; (2) a comparison of decision trees, boosted decision trees and multinomial log-linear models for the 3-class classification task; (3) a comparative analysis of the effect of varying the threshold for the binary classification tasks. The results are presented in terms of precision, sensitivity (recall) and AUC. For the binary classification tasks, we report the AUC score for the minority class. For the 3-class classification tasks, AUC scores are obtained treating one class as the positive class, and the other two classes combined as the negative class.

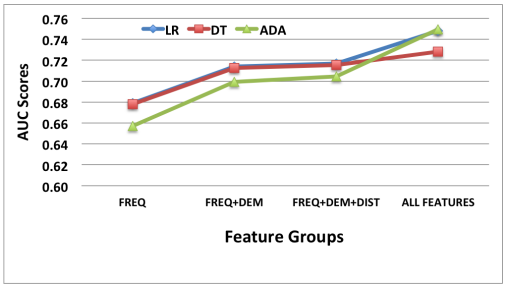
5.1 Performance across feature groups

We first tested the impact of including different feature groups from Table 2 on the overall performance of the models for classifying patients as low frequency ($\leq$ 1 visit), medium frequency (2-4 visits) and high frequency ($\geq$ 5 visits) ED users. To this end, we trained four decision trees, four boosted decision tree models, and four multinomial log-linear models using information from discharge records from 2009, and the number of ED visits in 2010 as the response variable value, i.e. training was done over 2009 $\rightarrow$ 2010 data only (see Section 4). For each technique, we trained a basic model that takes only the frequency features into account (FREQ), a model that takes the frequency features and the demographic information into account (FREQ + DEM), a model that takes the frequency features, the demographic information and the distance features into account (FREQ + DEM + DIST), and finally a model that takes all previous features as well as the medical features into account (ALL FEATURES).

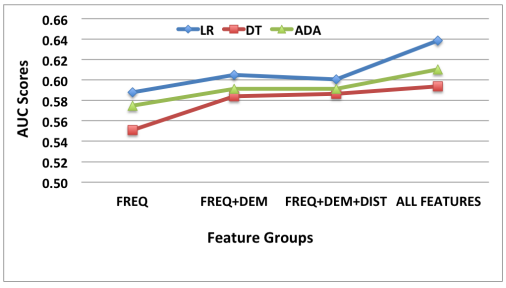
Figure 2 presents the AUC scores of these models treating respectively the low frequency ($\leq$ 1 visit), medium frequency (2-4 visits) and high frequency ($\geq$ 5 visits) group of patients as the positive class. In other words, each of Figure 2 (a), (b) and (c) contains results about the exact same 12 models. The difference results from which class is being treated as the positive class for the AUC score computation. All AUC scores in these figures are averaged across the three test years (2010 $\rightarrow$ 2011, 2011 $\rightarrow$ 2012 and 2012 $\rightarrow$ 2013). More detailed results per year are presented in Section 5.2.

It is interesting to note that (1) the frequency of past ED usage is indicative of the frequency of future ED usage; (2) the addition of demographic information and the severity and comorbidity features is very beneficial; while (3) the distance features are only moderately helpful. This can also be seen in the example trained decision tree shown in Figure 3: the decision tree learning algorithm selected only demographic, frequency and medical features as the split attributes at different levels in the tree in order to make the final class decision.

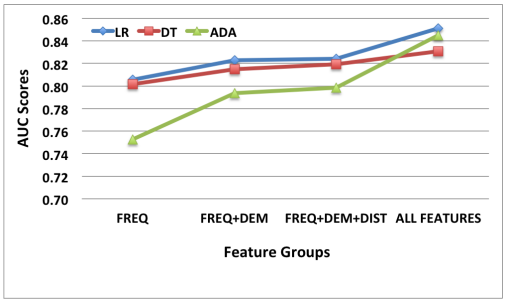
[7]https://cran.r-project.org/web/packages/nnet/nnet.pdf



(a) $\leq$ 1 visit as the positive class



(b) 2-4 visits as the positive class



(c) $\geq$ 5 visits as the positive class

Figure 2: AUC results for different combinations of feature groups from Table 2, treating patients with: (a) $\leq$ 1 visit as the positive class; (b) 2-4 visits as the positive class; (c) $\geq$ 5 visits as the positive class. FREQ= Frequency, DEM = Demographic, and DIST = Distance. The AUC scores are averaged across the three test years.

In general, it can been seen from Figure 2 (a), (b) and (c) that adding more information (more features) improves the overall performance of the models. The highest AUC scores are observed when all features were used. This was observed for all three classes, indicating that the combination of frequency of ED visits, demographic, distance and medical information helps in improving the performance of the models. Therefore, we use all features in the experiments in the remainder of this section.

5.2 3-class classification

Here we describe more detailed results of our models for

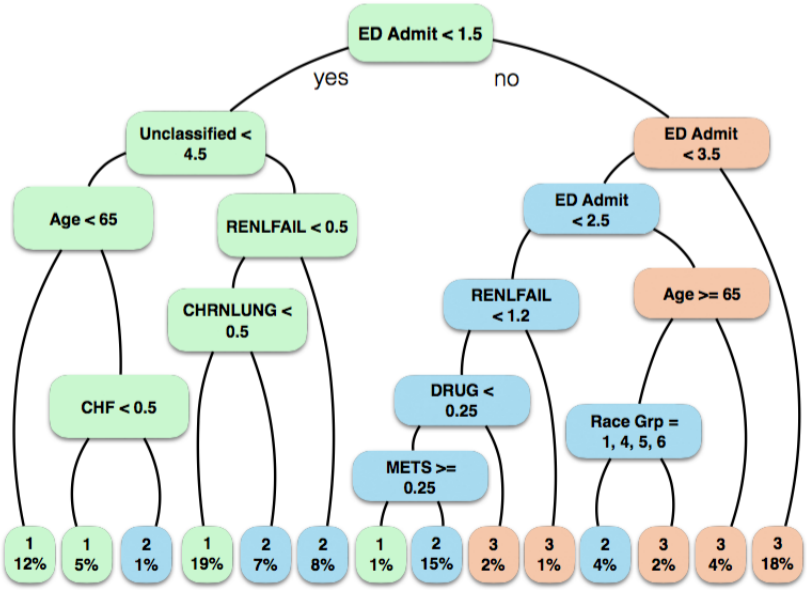


Figure 3: A pruned trained decision tree example for the 3-class classification task. It can be seen how different features related to demographic information (age, race); medical information (comborbidity features), and frequency of past ED visits are used in making the predictions for future ED usage.

Positive Class	Model	2010 → 2011	2011 → 2012	2012 → 2013
≤ 1 visit	Multinomial	0.75	0.75	0.74
	DT	0.73	0.73	0.73
	AdaBoost	0.75	0.75	0.75
2-4 visits	Multinomial	0.65	0.64	0.63
	DT	0.60	0.59	0.59
	AdaBoost	0.62	0.61	0.61
≥ 5 visits	Multinomial	0.85	0.85	0.85
	DT	0.83	0.83	0.83
	AdaBoost	0.84	0.85	0.84

Table 4: AUC results for the 3-class classification problem, using all features. All models were trained over 2009 → 2010 **data only, and subsequently applied to make predictions for** 2010 → 2011**,** 2011 → 2012**, and** 2012 → 2013**.**

performing the 3-class classification task, where the goal is to classify patients as low frequency (≤ 1 visit), medium frequency (2-4 visits) and high frequency (≥ 5 visits) ED users. All models discussed here and in Section 5.3 are trained using all features combined. Table 4 contains an overview of the AUC scores for the 3-class classification problem with all the different models using all feature groups (FRQ-DEM-DST-MED). As is the case throughout Section 5, all training was done over 2009 → 2010 data only. The results in Table 4 show how accurate these 2009 → 2010 trained models are at predicting future ED visits for each of the remaining years for which we have ground truth data available in our dataset.

Three key observations can be made from the results in Table 4. First, all three models (LR, DT and AdaBoost) have strong predictive ability to discriminate high frequency ED users (visit ≥ 5), with AUC ≥ 83%. Second, predicting low and high frequency ED users (AUC ≥ 83% and AUC ≥ 73% respectively), is easier when compared to moderate ED users (AUC ≥ 59%). Third, the performance of all the three models remains the same across all three years for all three classes. Such outcomes are important when consistent performance across all years is desired.

Confusion matrices for each of the models are presented for the year 2012 → 2013 in Table 5. We report results from only one test year here because results for other test years were very similar, hence no new insights could be drawn. Similar to the already mentioned results for the 3-class classification task, the results in Table 5 also show a higher percentage of correctly classified instances for the low and high frequency of ED visit classes. That is, instances from the most frequent (visit ≥ 5) and least frequent (visit ≤ 1) class are often correctly classified as their actual class (≈ 60%). Additionally, the percentage of mis-classification of high frequency users as low or medium frequency ED users is small (≤ 25%). Such outcomes are important because they reduce the chance of missing out patients who are likely to come back several times to ED.

Since our results (AUC, precision and sensitivity) are similar across the three test years (2010 → 2011, 2011 → 2012 and 2012 → 2013), in the remainder of this section, we present averages of the results across the three years. Table 6 presents an overview of averaged precision, sensitivity and AUC scores. The results highlight the following observations: the sensitivity of models for class ≤ 1 and ≥ 5 is higher when compared to the class 2-4 visits, suggesting that among least and most frequent ED users, a large proportion (above 60%) of them can be identified correctly. Next, there is a drastic drop in precision scores for moderate (≈ 13%) and frequent ED (≈ 9%) users when compared to the least frequent users (≈ 95%) class, indicating risk of producing a large number of false positives when making predictions for these two types of ED users. However, the cost (in terms

Decision Trees		Actual Class			
		≤ 1 visit	2-4 visits	≥ 5 visits	Total
Predicted Class	≤ 1 visit	635,607 (61.20%)	29,409 (30.05%)	2,261 (13.67%)	667,277
	2-4 visits	304,405 (29.30%)	39,416 (40.28%)	4,194 (25.36%)	348,015
	≥ 5 visits	98,774 (9.50%)	29,040 (29.67%)	10,082 (60.97%)	137,896
	Total	1,038,786	97,865	16,537	N

AdaBoost		Actual Class			
		≤ 1 visit	2-4 visits	≥ 5 visits	Total
Predicted Class	≤ 1 visit	631,723 (60.82%)	28,571 (29.19%)	2,204 (13.33%)	662,498
	2-4 visits	293,079 (28.21%)	38,461 (39.31%)	4,215 (25.48%)	335,755
	≥ 5 visits	113,984 (10.97%)	30,833 (31.50%)	10,118 (61.19%)	154,935
	Total	1,038,786	97,865	16,537	N

Multinomial model		Actual Class			
		≤ 1 visit	2-4 visits	≥ 5 visits	Total
Predicted Class	≤ 1 visit	711,255 (68.47%)	35,381 (36.14%)	2,874 (17.37%)	749,510
	2-4 visits	248,874 (23.95%)	36,954 (37.78%)	4,014 (24.27%)	289,842
	≥ 5 visits	78,657 (7.58%)	25,530 (26.08%)	9,649 (58.36%)	113,836
	Total	1,038,786	97,865	16,537	N

Table 5: Confusion matrices for 2012 $\rightarrow$ 2013, using all features.

Positive Class	Model	Sensitivity	Precision	AUC
≤ 1 visit	Multinomial	0.69	0.95	0.75
	DT	0.62	0.95	0.73
	AdaBoost	0.61	0.95	0.75
2-4 visits	Multinomial	0.38	0.13	0.64
	DT	0.40	0.12	0.59
	AdaBoost	0.39	0.12	0.61
≥ 5 visits	Multinomial	0.58	0.09	0.85
	DT	0.60	0.08	0.83
	AdaBoost	0.61	0.07	0.84

Table 6: Average results are across the three test years, using all features

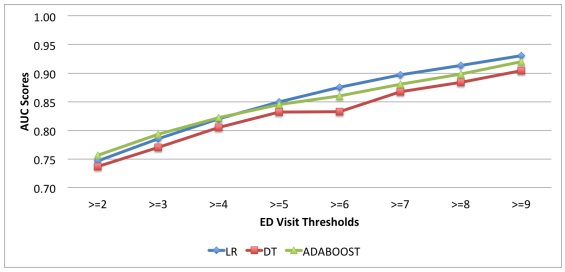


Figure 4: AUC scores of decision trees (DT), AdaBoost, and logistic regression (LR) models for the binary classification task of predicting frequent ED usage. The AUC scores are averaged across the three test years.

of hospital resource utilization and patients' well being) for ED visits of false negatives is higher than for false positives. Therefore, higher sensitivity is more desirable in such scenarios. Nonetheless, in our future research, we aim to improve our existing models and explore some new models to achieve higher precision scores as well. Unlike sensitivity and precision, AUC is much more stable across the three classes.

5.3 Binary Classification

Finally, in Figure 4 we present a comparative analysis of a series of predictive models for the binary classification task of inferring whether the number of future ED visits of a patient will be below or above a varying threshold. The results in Figure 4 show that the AUC scores improve as the threshold for number of visits is increased, for all three methods (LR, DT and AdaBoost). This echoes the findings of our multiclass classification results (Tables 4 and 6), where we already observed that the prediction of high frequency ED users can be done very accurately using machine learning methods. Furthermore, a similar improvement in the performance of the models for the binary classification task was also observed in [20] when the threshold for defining high frequency usage was increased, i.e. the model predicting frequent ED use as defined as 16 or more visits in the subsequent two years showed better discrimination than the models predicting a smaller number of ED visits.

6. CONCLUSION

A large percentage of emergency department (ED) visits originates from a small percentage of patients who keep returning to the ED. Being able to flag these frequent users in advance can help providers to take appropriate interventions to reduce the number of ED visits, thereby reducing cost and improving quality of care. In this paper we presented machine learning models that can predict future ED utilization of individual patients, using only information from the present and the past. We trained decision trees, boosted decision trees and logistic regression models on discharge records from California-licensed hospitals from the years 2009 and 2010, and evaluate their predictive accuracy for the years 2011-2013.

Overall there were three key findings of this study. First, all three kinds of models (LR, DT and AdaBoost) have a strong predictive ability to discriminate frequent ED users (number of visits ≥ 5). Second, all models show consistent outcomes across the three test years in our dataset (2010 $\rightarrow$ 2011, 2011 $\rightarrow$ 2012 and 2012 $\rightarrow$ 2013), which is a desired property when a predictive tool is required that is stable

and consistent year over year is required. Third, the low and high frequency ED users are comparatively easier to single out when compared to moderate ED users (with higher sensitivity and AUC). For completeness, we also did a comparative analysis with a series of logistic regression models trained as in [20] for the binary classification task of inferring whether the number of ED visits of a patient will be below or above a varying threshold. The results suggest that using machine learning techniques, predicting frequent ED users (who are also minority ED users) with high accuracy is possible. Similar outcomes were observed when a single 3-class classification model was used instead of multiple single class models. This is a useful outcome since one model can be used to predict different types of ED users.

In our future research, we aim to investigate additional state-of-the-art machine learning methods for accurately predicting frequent ED users. While the choice of casting the problem as a classification problem in this paper was inspired by [20], approaching it as a regression problem is equally meaningful. Additionally, we aim to find ways to improve the precision so that the number of false positives can be further reduced. In order to further increase the clinical relevance of this problem (predicting frequent ED users), in our future research, we aim to not only accurately predict frequent ED users, but also the unique care needs (for instance, multiple chronic disease management, behavioral health interventions, and palliative care) of the clinically distinct frequent user subpopulations. This will allow for more efficient allocation and precise targeting of evidence-based interventions to reduce ED utilization.

Acknowledgments

The authors would like to thank the staff of the California Office of Statewide Health Planning and Development (OSHPD) for their assistance in facilitating the process of obtaining and understanding the data used in this study.

7. REFERENCES

[1] J.G. Behr and R. Diaz. Emergency department frequent utilization for non-emergent presentments: results from a regional urban trauma center study. *PLoS ONE*, 11(1), 2016.

[2] J. Boyle, M. Jessup, J. Crilly, D. Green, J. Lind, M. Wallis, P. Miller, and G. Fitzgerald. Predicting emergency department admissions. *Emerg Med J*, 29(5):358–365, 2012.

[3] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and regression trees*. Wadsworth Publishing Company, 1984.

[4] R.M. Cowan and S. Trzeciak. Clinical review: Emergency department overcrowding and the potential impact on the critically ill. *Crit Care*, 9(3):291–295, 2005.

[5] A. Elixhauser, C. Steiner, R. Harris, and R.M. Coffey. Comorbidity measures for use with administrative data. *Medical care*, 36(1):8–27, 1998.

[6] Y. Freund and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119–139, 1997.

[7] R. M. Gindi, L. I. Black, and R. A. Cohen. Reasons for emergency room use among U.S. adults aged 18-64: National health interview survey, 2013 and 2014. *Natl Health Stat Report*, (90):1–16, 2016.

[8] I. Higginson. Emergency department crowding. *Emerg Med J*, 29(6):437–443, Jun 2012.

[9] N.R. Hoot and D. Aronsky. Systematic review of emergency department crowding: causes, effects, and solutions. *Ann Emerg Med*, 52(2):126–136, Aug 2008.

[10] G.S. Kumar and R. Klein. Effectiveness of case management strategies in reducing emergency department visits in frequent user patient populations: a systematic review. *J Emerg Med*, 44(3):717–729, 2013.

[11] E.K. Lee, F. Yuan, D.A. Hirsh, M.D. Mallory, and H.K. Simon. A clinical decision tool for predicting patient care characteristics: patients returning within 72 hours in the emergency department. *AMIA Annu Symp Proc*, 2012:495–504, 2012.

[12] R. Moineddin, C. Meaney, M. Agha, B. Zagorski, and R.H. Glazier. Modeling factors influencing the demand for emergency department services in Ontario: a comparison of methods. *BMC Emerg Med*, 11:13, 2011.

[13] E. Neufeld, K.A. Viau, J.P. Hirdes, and W. Warry. Predictors of frequent emergency department visits among rural older adults in Ontario using the resident assessment instrument-home care. *Aust J Rural Health*, 24(2):115–122, 2016.

[14] A.Y. Ng and M.I. Jordan. On discriminative vs. denerative classifiers: A comparison of logistic regression and naive bayes. In *Advances in neural information processing systems 14, Proceedings of the 2001 NIPS conference*, pages 841–848. MIT Press, 2001.

[15] M.M. Rahman and D.N. Davis. Addressing the class imbalance problem in medical datasets. *International Journal of Machine Learning and Computing*, 3(2):224, 2013.

[16] J.P. Ruger, C.J. Richter, E.L. Spitznagel, and L.M. Lewis. Analysis of costs, length of stay, and utilization of emergency department services by frequent users: implications for health policy. *Acad Emerg Med*, 11(12):1311–1317, 2004.

[17] R.G. Solberg, B.L. Edwards, J.P. Chidester, D.G. Perina, W.J. Brady, and M.D. Williams. The prehospital and hospital costs of emergency care for frequent ED patients. *Am J Emerg Med*, 34(3):459–463, 2016.

[18] L.J. Soril, L.E. Leggett, D.L. Lorenzetti, T.W. Noseworthy, and F.M. Clement. Reducing frequent visits to the emergency department: a systematic review of interventions. *PLoS ONE*, 10(4), 2015.

[19] Y. Sun, B.H. Heng, S.Y. Tay, and E. Seow. Predicting hospital admissions at emergency department triage using routine administrative data. *Acad Emerg Med*, 18(8):844–850, Aug 2011.

[20] J. Wu, S.J. Grannis, H. Xu, and J.T. Finnell. A practical method for predicting frequent use of emergency department care using routinely available electronic registration data. *BMC Emergency Medicine*, 16(1):1–9, 2016.