Predicting Patient Readmission for Improved Resource Allocation using Machine Learning

STEM Fellowship Big Data Challenge 2022

Umar Ali
School of Biomedical Engineering
University of British Columbia
Vancouver, Canada
umar@aminabdolkhani.com

Amin Abdolkhani
School of Biomedical Engineering
University of British Columbia
Vancouver, Canada
amin@umarali.ca

Hamza Jimale Rasheed
Department of Integrated Engineering
University of British Columbia
Vancouver, Canada
ham.rashe@gmail.com

Johnny Zhao
School of Biomedical Engineering
University of British Columbia
Vancouver, Canada
jzhao19@ubc.student.ca

***Abstract*—A major reason why current economical models of healthcare are not sustainable is due to the fact that many patients are unnecessarily readmitted within 30 days. In this paper, we use a decision tree based machine learning model to predict the likelihood that a patient is readmitted to the hospital within 30 days. This allows for hospitals to more aptly allocate their resources and create a much more sustainable economical model for healthcare services. Our model has shown to have an accuracy of 94% and its decision making process is consistent with the literature, indicating its high clincal utility.**

I. Introduction

Canada's emergency departments are constantly facing severe overcrowding. There are approximately 11.7 million ED visits per year in Canada which contribute to a shortage of beds, medical supplies and professionals in the emergency room [5]. This problem results in reduced patient quality of care, long wait times leading to ill-treatment of critical patients and burnout for hospital staff [1]. Furthermore, emergency room readmissions contribute to a large cost on the health care system, such as the U.S. where the cost of rehospitalization is upwards of $17.4 billion dollars [4].

A major contributor to this issues the mismanagement of patient throughout [2]. Patient throughput refers to the method of which patients pass through the hospitals and most notably the emergency department. Mainly, standardised triage systems have been developed worldwide to prioritise patients based on urgency of care needed in the typically overcrowded emergency department [7].

The triage process, generally consisting of a visual inspection and checklist, is time-consuming and leads to bottlenecks in patient throughput. Furthermore, commonly used triage systems like the Emergency Severity Index, while being very accurate in identifying urgent care patients with and accuracy of 99.8%, tend to frequently misclassify low-urgency patients which results in needless waste of valueable resources such as beds and staff [7]. These cases of misclassification lead to more readmissions which are a huge bearing on the healthcare economy.

To remedey this resource allocation problem, our team developed a machine learning model to evaluate readmission risk of patients to the emergency room after discharge. We intend for this model to assist hospital administrators in allocating resources and reduce this risk.

II. Methods & Materials

Our research comprised of three major sections: (A.) Data Selection, (B) Data Exploration, and (C) Decision Tree. During the data selection phase, we had focused on understanding the current landscape of precision public health and finding open source data sets available. Afterwards, the data exploration phase was meant as a means of understanding the data at hand as well as to determine a suitable model. Lastly, we had determined two potential machine learning models that were capable of being easily interpreted.

A. Data Selection

During the data selection section, we had conducted a comprehensive literature review of various articles containing combinatorial sets of keywords of: "artificial intelligence", "precision public health", "patient admission", "hospital readmission", "evidence based medicine", and "resource allocation". We had reviewed a total of 33 articles that met the search criteria and further investigated the data availability policies for each paper. A total of 3 papers were available with open source data for analysis. We had chosen the dataset from Robinson and Hudali [6] for its relevance to our research objective and article impact score.

B. Data Exploration

For the data exploration, we had computed the distribution of each feature in the dataset using Python. We had plotted the histogram of all samples in the dataset within a single column feature. Afterwards, we had computed the correlation matrix, to determine possible correlations between features. Lastly, we had computed the scatter and density plots of each feature to gain more insight about cross-feature correlation.

During the data exploration, the features from Table I were noted as the most prominent and were considered for training the model; as well as the feature constituents such as the values used for HOSPITAL and LACE index scores.

TABLE I
TABLE OF FEATURES USED IN TRAINING MODEL; TABLE TAKEN AND ADAPTED FROM ROBINSON AND HUDALI [6]

Characteristic	**Not readmitted n = 397**	**Readmitted n = 35**
Age, mean (SD)	62 (15.7)	56 (14.9)
Female	193 (49%)	15 (43%)
Urgent or emergent admission	397 (100%)	35 (100%)
Discharge from oncology service	41 (10%)	3 (9%)
Length of stay 5 days	246 (62%)	23 (66%)
Hospital admissions in the last year	2.3 (3.0)	5.2 (1.7)
Emergency department visits in last 6 months	2.3 (1.9)	3.3 (3.6)
H - Low hemoglobin level at discharge	26 (6%)	4 (11%)
S - Low sodium level at discharge	86 (22%)	9 (26%)
C - Comorbidity index score (SD)	4.4 (3.0)	5 (3.7)
HOSPITAL score (high risk)	235 (55%)	31 (86%)
LACE index (high risk)	337 (79%)	32 (89%)

C. Model Training

The dataset that we are using in this study is structured data, as such, we have determined that a decision tree would provide an adequate accuracy and allow us to easily interpret the model. Hence, decision tree models are advantageous as we can investigate the decision making process for further insights. Whereas with machine learning algorithms like logistic regression and neural networks, the human interpretation is more difficult.

In this study we had used two different models: (1) Random Forest and (2) Gradient Boosted Decision Tree. For the latter mode, we had configured three different hyperparameter settings to find the optimal performance. We had used a 70-30 data split to predict the "Admission within 30 days" feature.

1) Random Forest: The random forest (RF) model was trained using the TensorFlow python package with the Decision Forests library. We had used the default parameters to train the random forest model and used the first indexed tree that was returned as our primary decision model for testing.

2) Gradient Boosted Decision Tree: The gradient boosted decision tree (GBDT) was trained with three different configurations that alter the hyperparameter settings. Refer to Table II for the specific settings.

TABLE II
TABLE OF CONFIGURATION SETTINGS FOR MACHINE LEARNING MODELS

Model	**Configuration**
Random Forest	Default
Gradient Boosted Decision Tree	Best First Global Growing Strategy
	Best First Global, Sparse Oblique Split Axis, and Random Categorical Algorithm
	Hyperparameter Template: Benchmark Rank 1

III. RESULTS

This study yielded several areas of insight that could prove to be valuable for further investigating and studying the demographics and liklihood of patient readmission to the hospital within 30 days. Moreover, this analysis has shown a strong decision making process using evidence based medicine to guide its screening values.

A. Data Discoveries

We can see that in Figure 1 that there is an even distribution between male and female participants, thus indicating no significant bias. However, the distribution between the number of admissions per year and the admission within 30 days is much more asymmetric.

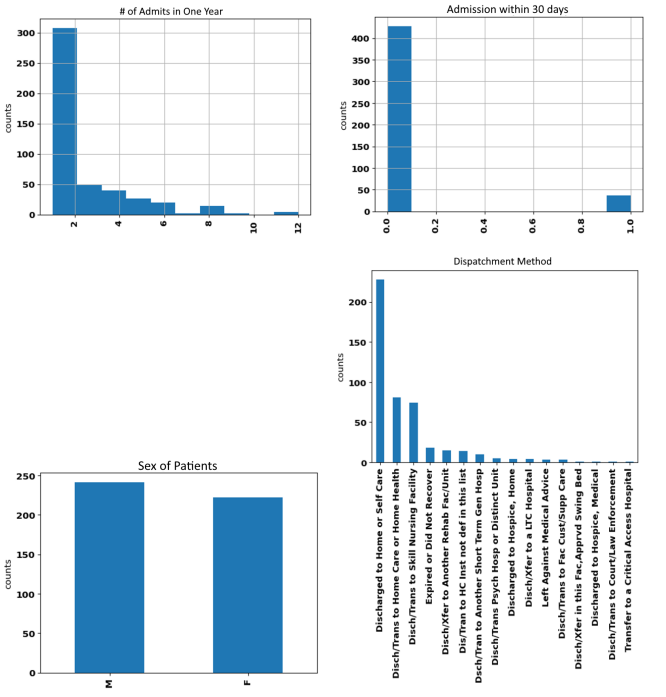

Fig. 1. Histograms of three different data features: Number of patient admissions over the year, number of patients who were readmitted within 30 days, and the count of the sex of patients respectively

Refer to Figure 2 correlation matrix shows that there are few correlations across all the features within the dataset. The correlations that were found, were correlations between index score features and their constituents.

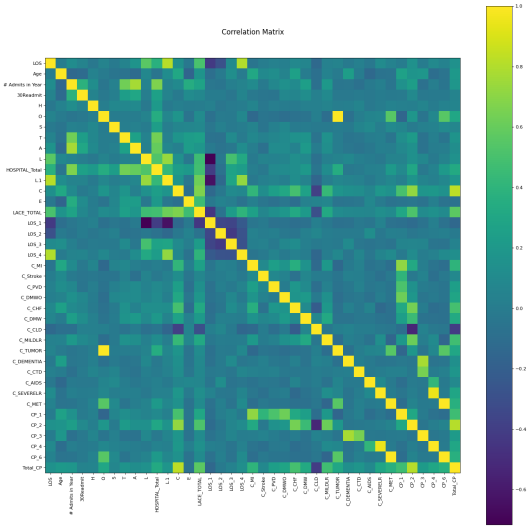

Fig. 2. Correlation matrix visualizing significant correlations between all features in dataset

B. Model Accuracy

The decision tree model has shown to have an on average accuracy of greater than 90% for determining likelihood of patient readmission within 30 days. Refer to Figure 3 to see the training accuracy of the two different machine learning models over the total number of trees. This shows us that the random forest models produces much more in-depth decision trees whereas the gradient boosted decision trees are much more shallow.

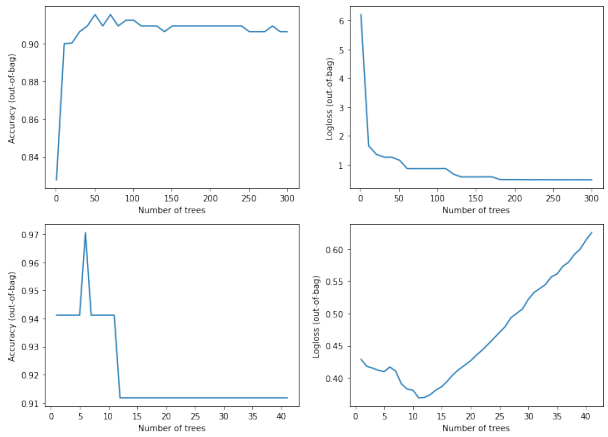

Fig. 3. Plots of accuracy and loss values during traing of decision trees for Random Forest (**Top Row**) and Gradient Boosted trees (**Bottom Row**)

Refer to Table III to see the table of performance metrics across the all models being assessed. We can see that the random forest is the worst performing model and that the gradient boosted decision trees are much better performing with an occasional trade off with area under curve values.

TABLE III
TABLE OF MODEL ACCURACY, AREA UNDER CURVE, AND LOSS

Model	**ACC**	**AUC**	**LOSS**
Random Forest	0.90	0.750	0.470
Gradient Boosted Decision Tree # 1	0.94	0.750	0.360
Gradient Boosted Decision Tree # 2	0.94	0.668	0.414
Gradient Boosted Decision Tree # 3	0.94	0.704	0.388

We can see a more detailed view of the performance metrics of each model by referring to Figure 4. The receiver operating characteristic (ROC) curve shows the rate of true positives and false positives so we can better measure the generalizability of the models.

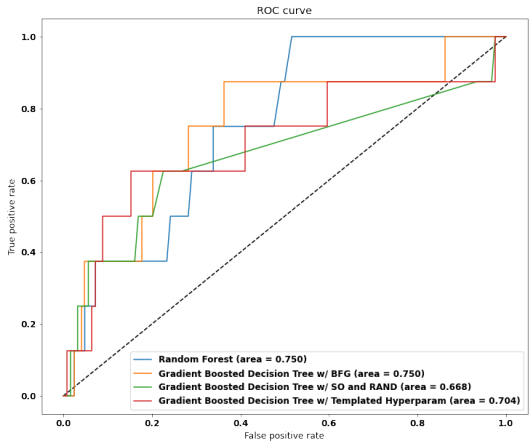

Fig. 4. Receiver operating characteristic curve of all four decision tree models

Lastly, we can refer to Figure 5 to examine the exact decision making process of the best performing machine learning model. We can see that the model places a high priority on determining the number of admissions a patient has had in a year as a significant predictor for readmission within 30 days.

IV. DISCUSSION

In this study we are looking to demonstrate that a decision tree based machine learning model is an adequate model for predicting patient readmission. Moreover, we are wanting to assess this tool as a potential clinical tools for administrators in hospitals.

A. Data Exploration

We can see that during the data exploration phase that the data acquired was fairly symmetric and even in its distributions for patient demographics, in regards to sex and age. However, there is a severe class imbalance between number of patients who were readmitted to the hospital within 30 days. This

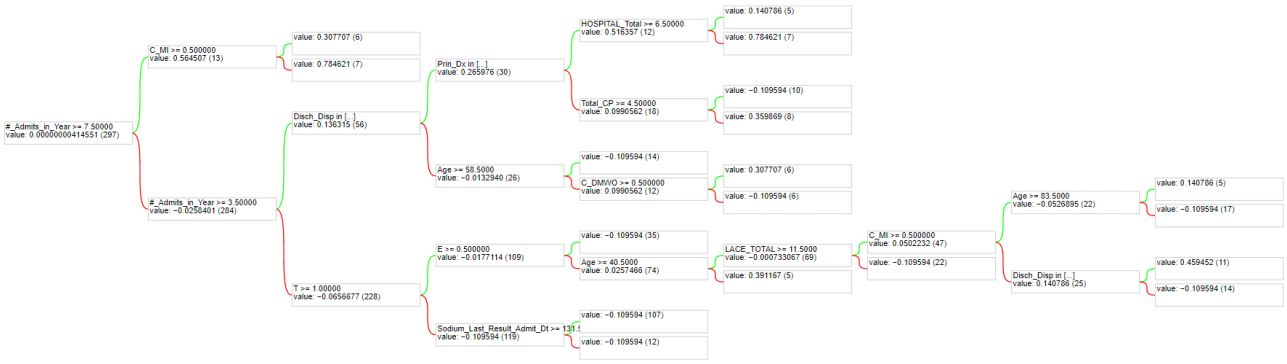

Fig. 5. Architecture of best performing gradient boosted decision tree (with best first global growing strategy) with specific feature values shown for making predictions

means that our model is likely to be overfit towards predicting no readmission within 30 days. A possible solution to this could be finding more diverse data however, finding other datasets with the same features would prove to be difficult and not possible within the time frame of this analysis. Another possibility was to resize the overbalanced class to become proportional with the unbalanced class. The downside with this approach is that the model then does not have sufficient data to be trained to reach an adequate accuracy.

Through the initial data exploration it was found that many of the features have very little correlation with another. This further supports the choice of using decision trees as opposed to regression models since regression models often depend on some correlation between features. Furthermore, this exploration has shown us and affirmed many findings that already exists in literature, such as older patients having a higher likelihood of being readmitted into the hospital.

B. Decision Tree and Clinical Relevance

The features identified in Figure 5 strongly correlate with high risk variables identified in literature. Specifically, we can see that the tree puts a strong emphasis on the number of hospital admissions, age and the Charleston Comorbidity Index. These same variables, in addition to length of stay, have been shown to significantly increase the likelihood of hospital readmission in statistical and retro-spective studies [3].

These variables emphasise the state of health of the patient. The algorithm accurately associates the relationship between overall patient health and these factors to ultimately predict whether a patient will be readmitted within a 30 day period or not.

C. Limitations

A major limitation of our study was a significant class imbalance in patient outcomes. Out of 432 patients in the data set only 8.1% of the patients resulted in a readmission. This bias in the data can results in biases in the algorithm which result in a lower accuracy of the model. Furthermore, this skew may result in testing or validation data splits which do not have and cases of readmission and do not evaluate or validate the model fairly.

This limitation is very serious and has deep implications on the effectiveness of our model. Going forward, our team intends to conduct a very thorough literature search to find diverse sources of data and fix this imbalance. In addition, we will validate our model with clinical and statistical studies to confirm its effectiveness.

D. Future Investigations

For future investigation, our team intends use different models and evaluate their effectiveness. Particularly we wish to implement a deep learning model to garner more insight into readmissions. As a part of this process we will search for larger datasets with balanced classes to address the limitations encountered in this study.

Furthermore, our team plans to implement an Artificial Intelligence medical triage system which can classify patient urgency based on medical history and physiological variables. Ultimately, envision the model being used in a hospital setting to streamline patient throughput and remove or alleviate triage bottlenecks.

V. Conclusion

Emergency department overcrowding is a serious and growing problem in Canada and all over the world. The proper allocation of hospital resources, such as staff and medical supplies, are critical to ensuring optimal patient health and outcomes. Our team identified 30 day hospital readmissions as a major source of resource misallocation due to mismanagement and we created a machine learning learning model to help mitigate it. Specifically, the model accurately identifies patients at risk of readmission and allows hospital administrators to allocate resource accordingly and reduce this risk. In addition, our model sheds light on the factors that predispose patients to being readmitted within 30 day which can help to educate and inform healthcare professionals. Ultimately, we hope that our model can be used as a viable solution to the ever present emergency department overcrowding problem.

VI. Code Availability

The code for this project is available at gist.github.com/gitUmaru/591ded63f021f9f02423e80b91668656.

The code for this analysis is under the GNU General Public License v3.0.

REFERENCES

[1] Andrew Affleck et al. "Emergency department overcrowding and access block". In: *Canadian Journal of Emergency Medicine* 15.6 (2013), pp. 359–370.

[2] Ray DeAnda. "Stop the bottleneck: improving patient throughput in the emergency department". In: *Journal of Emergency Nursing* 44.6 (2018), pp. 582–588.

[3] Maria Glans et al. "Risk factors for hospital readmission in older adults within 30 days of discharge–a comparative retrospective study". In: *BMC geriatrics* 20.1 (2020), pp. 1–12.

[4] Stephen F Jencks, Mark V Williams, and Eric A Coleman. "Rehospitalizations among patients in the Medicare fee-for-service program". In: *New England Journal of Medicine* 360.14 (2009), pp. 1418–1428.

[5] *NACRS emergency department visits and lengths of stay*. URL: https : / / www . cihi . ca / en / nacrs - emergency - department-visits-and-lengths-of-stay.

[6] Robert Robinson and Tamer Hudali. "The HOSPITAL score and LACE index as predictors of 30 day readmission in a retrospective study at a university-affiliated community hospital". In: *PeerJ* 5 (2017), e3137.

[7] Joany M Zachariasse et al. "Performance of triage systems in emergency care: a systematic review and meta-analysis". In: *BMJ open* 9.5 (2019), e026471.