Predicting Elderly Patient Behaviour in Rural Healthcare Using Machine Learning

Prince Appiah[1], Thierry Oscar Edoh [2] [0000-0002-7390-3396] and Jules Degila[3][0000-0003-4688-9178]

[1]University of Education Winneba-Kumasi, Ghana
Princeappiah35@gmail.com
[2] RFW-Universität Bonn, Bonn, Germany
oscar.edoh@gmail.com
[3]Institute of Mathematics & Physical Science, Porto-Novo, Benin
Jules.degila@imsp-uac.org

Abstract. The digitization of modern health care data in a rural community has produced a vast amount of patient data stored in health care record systems. Together with the rise of computing power this data could produce effective insight through advanced analysis of this data and include it in medical applications for use in daily operations. This is the case in which structured, semi-structured and unstructured dataset from emergency room admissions is used for machine learning, in order to develop models that predict the possibility of an elderly patient returning to an emergency room within 96 hours. Logistic regression was the selected algorithm since it commonly used in the healthcare data set. The results from the model had a recall of 73% and a precision of 78%. This paper discusses the implementation of such a model in daily operations with a new approach to cost benefits. In other instances, the study is a proof of the concept of predictive modeling in a health care context in rural communities.

Keywords: Machine Learning, Rural healthcare, unstructured dataset, Logistic regression

1. Introduction

The invention of the computer and even long before that, people are trying to predict or interpret different outcomes from data. This also applies to the health care sector notwithstanding this sector is extremely sensitive to errors because of varied reasons [1]. Rural healthcare centers are exponentially increasing the amount of data. Which has become opportune for machine learning [2]. Data collected in a rural hospital containing vitals, lab results, metadata, etc. can be combined into individual records for every patient. Through machine learning techniques we are able to make use of all this data. Using these algorithms we can predict various outcomes and form recommendations to support medical professionals or do predictions regarding the patient's health. For example, algorithms can help assign medicines and treatments to patients, they can support medical professionals in making diagnoses and present new origins of certain diseases, the possibilities are endless. This knowledge can also be used to act proactively on different issues. It is, for example, possible to set computerized alerts when specific thresholds are exceeded to prevent unwanted consequences. However, despite the volume of data stored, the potential of using the data for strategic and operational decisions through the means of data analysis is, especially in rural health care, rarely acknowledge. It is obvious that, with this aforementioned data of patients, rural healthcare centers are not using data mining techniques to predict the behavior of elderly patients.

The purpose of this study is to apply machine learning to rural health care data, in order to predict elderly patient behavior, providing a basis for medical decisions or risk stratification. The study focus on the prominent property of emergency patients. The fact that ten percent of elderly patients sent from emergency room return within 96 hours. This will help elderly patents to be identified before they return, decisions concerning their care could be taken in

order to reduce the risk of them returning. And also, more importantly, improve patient safety and cut costs, and gain new knowledge about what is causing the elderly patient to return.

2. Problem Statement

Ever since the introduction of information technology in rural health care, the amount of data generated has increased exponentially. However, despite the volume of data stored, the potential of using the data for strategic and operational decisions through the means of data analysis is, especially in rural health care, rarely acknowledge. It is obvious that, with this aforementioned data of patients, rural healthcare centers are not using data mining techniques to predict the behavior of elderly patients.

3. Related Work

Machine learning technique Logistic regression is often used when predicting outcomes especially in combination with clinical trials. For instance, Baan [3] uses logistic regression to identify undiagnosed diabetes. Also [4] used a random forest model together with data collected from publicly available health register. Using only readily available health record data in electronic health registers, Razavian et al. [5] present a retrospective study predicting type 2 diabetes, [6] makes a prediction of risk of hospitalization or death; and Kontio et al. [7] predict patient acuity. These studies are good examples of performing data analysis in a health care setting. But using machine learning to perform data analysis in rural communities is very limited. Furthermore, when it comes to the prediction of elderly patient behavior involving patient's no study has been done concerning rural areas. An example, however, [8]who predicts patient's decisions about end-of-life care, by looking at the characteristics of patient's physicians. This is important in relation to this study, as the decision to revisit an emergency room is indeed taken by the patient. According to Hillestad [9], enhanced health data record systems could facilitate predictive modeling and thus decrease cost due to the insights of these predictions. However, these cost benefits are not assessed in the other studies as well as the setting in rural communities, mentioned in this section. This study, however, does indeed try to approach the cost benefits of implementing predictive models in daily decision making in rural healthcare.

4. Background

4.1 Electronic Health Record's (HER's)

HERs is a real-time digital version of a patient's clinical information that makes it possible to provide dynamic clinical patient information captured in structured records which can be consulted at every moment by medical professionals [10]. Electronic health records have introduced many merits for handling modern healthcare-related data. Its first merit is that healthcare professionals have improved access to the entire medical history of a patient. Electronic health records enable faster data retrieval and facilitate reporting of key healthcare quality indicators to the organizations, and also improve public health surveillance by immediate reporting of disease outbreaks. The electronic health records and the internet together help provide access to millions of health-related medical information critical for patient life [2]. Many rural hospitals and clinics are now using EHRs to keep records of their patents.

4.2 Machine Learning

According to Baan et. al., [3], "machine learning is about making computers modify or adapt their actions so that these actions get more accurate, where accuracy is measured by how well the chosen actions reflect the correct ones".

There two different types of learning supervised and unsupervised learning. Supervised learning is the most common approach to machine learning. The data provided to train the machine learning algorithm is well labeled with the correct outcome. Based on that information, the algorithm generalizes to respond correctly to all possible inputs. Examples of supervised models are k-nearest neighbors, regression model, Bayesian network and support vector machine. Unsupervised learning is a contrast to the supervised learning method. The algorithm is trained on data that is not labeled, due to that we don't know the outcome of. It mainly uses for finding patterns and detecting relevant information in data to form clusters [10]. An example of unsupervised learning is adaptive resonance theory and self-organizing map. This study used a supervised machine learning model known as logistic regression.

4.3 Logistic Regression

Logistic regression is a generalized linear model, sharing some similarities with linear regression, with the exception of the predicted value is binary (0, 1). In [11] stated that "logistic regression uses a logistic function to map the outcome of a linear model to 0 or 1, hence the name *Logistic regression* ".

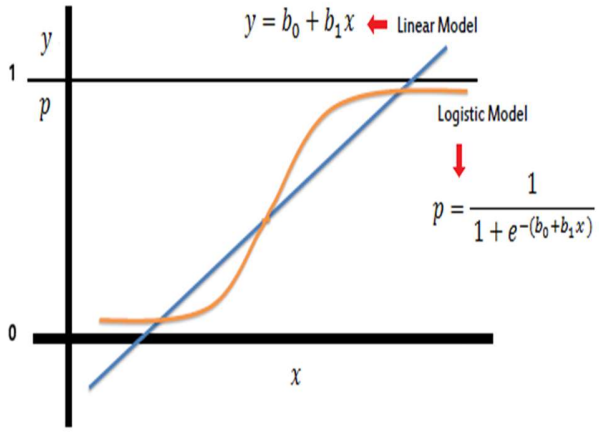

Fig. 1. Logistic Regression Model

The formula (1) for a simple linear regression model, (tries to minimize error in linear function)

$$y(x) = \beta_0 + \beta_1 x \qquad (1)$$

Where *y* is the response variable, β0 is the intercept, β1 is the coefficient and x is the input variable. In the case of a logistic regression model, the linear regression model is mapped to the logistic function from;

$$y(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \qquad (2)$$

But when more features (variables) are introduced the from change to;

$$y(x) = \frac{1}{1 + e^{-(w^T x)}} \qquad (3)$$

Where *w* is the vector weights applied to each variable contained in the vector *x*. It can be trained on one or many independent continuous features.

5. Methods and Data Source

5.1 Study Data

The data was collected from a rural hospital, from the analysis unit. Emergency admissions data was selected as the cohort, with the number of visits counted and previous diagnoses. The data collected were stored in a table and prepared by dichotomizing the categorical features, added as columns and assigned a 1 or a 0. The data was regarded as possible to have an effect on whether a patient would return or not was extracted, which resulted in Table 1.

5.2 Study Design

This study was in the form of an exploratory case study described by [12]. The data used is archival and quantitative. It is done in order to generate new ideas and insights, despite its exploratory. The case outline follows 3 steps. These are; (1) data collection and preparation, (2) model training and testing with adjustable hyper-parameters and (3) result of the study was evaluated including implementation analysis and conclusion.

5.3 Study Methods

The study choice of data mining technique for the study was implemented using Jupiter Notebook. The healthcare data collected from the rural hospital was saved to .csv extension for easy loading into Jupiter Notebook. SciKit Learn was imported to deployed Logistic Regression using Python codes. All the required analysis was done using Jupiter Notebook.

6. Implementation

The first stage is pre-processing the data. Selection was made during the pre-processing of the data to get only patients above the age of 60. Data in this phase partition into training and testing. In the next step, we applied the logistic regression model on the training dataset in order to build and train the model. With Jupiter Notebook, SciKit Learn was used for the analysis which contains Logistic regression. The training data set consists of 9 features, shown in Table 1.

Table 1. Feature Description

Feature	Type	Range
Age	Continuous	N ≥ 60
Sex	Categorical	(f /m)
Emergency time	Continuous	……..
Emergency cause	Categorical	--------
Mode of arrival	Categorical	--------
Admission hour	Categorical	(0,23)
Admission day	Categorical	(1,7)
Admission month	Categorical	(1,12)
Next 96 hours	Target (Binary)	(0,1)

The logistic regression model has been optimized on Area under the curve (AUC) over a hyper-parameters using stratified 5-fold cross-validation.

Table 2. Model Settings Summary

Classifier	Dataset	Hyper-parameter Set (Threshold)	Optimization
Logistic Regression	Basic	λ = (10 to the power of) 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0	Area Under Curve

7. Experiment Results

This section describes the results discovered during the study. The accuracy of the model performance was high, 76%. This was expected since more elderly patients are likely to return within 96 hours. Since accuracy thus not convey much information on model performance, we look interested in precision and recall.

The recall of the model was 73%, this interpreted that more than 50% of the elderly patients returning are identified within 96 hours. Precision range from 50% to 80%. A precision of 78% means that of the elderly patients predicted to return, 50% of them actually did. The precision of 50% was considered high as the data collected is very unbalanced.

Regardless of the confusion matrix, a 50% precision means that a predicted positive outcome will be wrong 50% of the time. Figure 2, shows the scores of the calculated by extracting a small sub-set dataset from the training data. Thus they are just an indication of how the scores are distributed across the different hyper-parameters.

Confusion Matrix:

Threshold 0.5	Actual Positive	Actual Negative
Predicted Positive	64 (TP)	18 (FP)
Predicted Negative	23 (FN)	34(TN)

Where: *True Positive (TP), True Negative (TN) False Positive (FP), False Negative (FN)*

Calculation of precision and Recall:

$$Precision = \frac{TP}{TP+} = \frac{64}{64+18} = 0.78 \quad (4)$$

$$Recall = \frac{TP}{TP+FN} = \frac{64}{64+2} = 0.73 \quad (5)$$

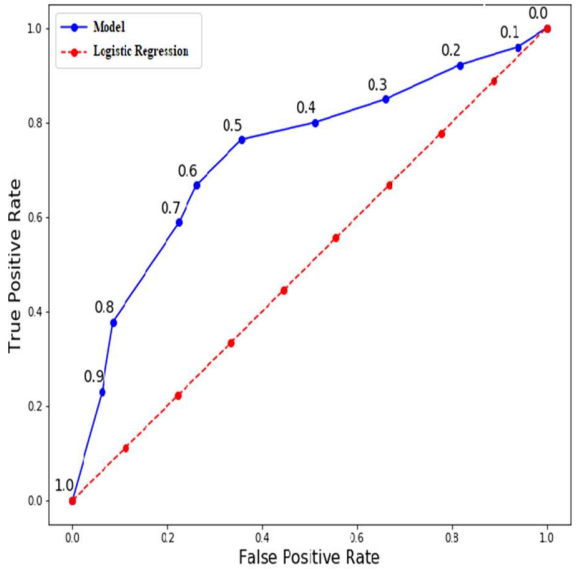

Fig. 2. ROC Curve for Different Hyper-parameters.

From the ROC curve, at a threshold of 1.0, we classify no elderly patients is returning within 96 hours and hence have a recall and precision of 0.0. As the threshold decreases, the recall increases because we identify more elderly patients returning to the emergency unit. However, as our recall increases, our precision decreases because, in addition to increasing the true positives, we increase the false positives. At a threshold of 0.0, our recall is perfect, we could find all patients would return within 96 hours.

8. Conclusion and Future Work

The study found that age affects the result positively, that is patients over 70 years are more likely is it for him/her to return within 96 hours. An intuitive notation is that older patients with heart diseases and sickle cell disease need to be treated with more caution. From the study, the following features have a good impact on elderly patient return, i.e., Admission last 12 months, the number of previous primary diagnoses and emergency cause. Emergence time has the

smallest impact on all the features. After predicting that an elderly patient will return to the hospital within 96 hours, an action would be taken to prevent that from happening. This will lead to an extra, unknown cost above the normal visit cost.

The study opens up some different paths in relation to future work. Firstly other behavior could be studied in rural communities EHR. Secondly, deploy different algorithms or machine learning techniques to come out with the best one instead of using only one model for predictions. Lastly more patient's features could be used for model training and testing.

References

[1] A. A. Fuss, "The Prevention Of Depression : A Machine Learning Approach The Prevention Of Depression : A Machine Learning Approach," 2019.

[2] J. Kallio and M. Juhola, "Support Vector Machine and Deep Learning in Medical Application," 2017.

[3] C. A. Baan *et al.*, "Performance of a predictive model to identify undiagnosed diabetes in a health care setting," *Diabetes Care*, vol. 22, no. 2, pp. 213–219, 1999.

[4] M. Khalilia, S. Chakraborty, and M. Popescu, "Predicting disease risks from highly imbalanced data using random forest," *BMC Med. Inform. Decis. Mak.*, vol. 11, no. 1, 2017.

[5] N. Razavian, S. Blecker, A. M. Schmidt, A. Smith-Mclallen, S. Nigam, and D. Sontag, "Population-level prediction of type 2 diabetes from claims data and analysis of risk factors," *Big Data*, vol. 3, no. 4, pp. 277–287, 2017.

[6] D. Z. Louis *et al.*, "Predicting risk of hospitalisation or death: A retrospective population-based analysis," *BMJ Open*, vol. 4, no. 9, pp. 1–8, 2017.

[7] E. Kontio *et al.*, "Predicting patient acuity from electronic patient records," *J. Biomed. Inform.*, vol. 51, pp. 35–40, 2017.

[8] eljko Ivezic, "Statistics: A Practical Python Guide for the Analysis of Survey Data," 2019.

[9] R. Hillestad *et al.*, "Can electronic medical record systems transform health care? Potential health benefits, savings, and costs," *Health Aff.*, vol. 24, no. 5, pp. 1103–1117, 2017.

[10] D. Boonen, "The impact of bias on the predictive value of EHR driven machine learning models Dries Boonen," 2019.

[11] M. W. Attia, T. Zaoutis, J. D. Klein, and F. A. Meier, "Performance of a predictive model for streptococcal pharyngitis in children," *Arch. Pediatr. Adolesc. Med.*, vol. 155, no. 6, pp. 687–691, 2018.

[12] P. Runeson and M. Höst, "Tutorial: Case studies in software engineering," in *Lecture Notes in Business Information Processing*, 2018.