Analyzing the risk of each patient based on electronic patients record

A (Thesis/Project/Dissertation)

Presented to the

Faculty of

University of North Carolina,

Greensboro

In Partial Fulfillment

of the Requirements for the Degree

Master of (Arts/Science)

in

Computer Science

by

(Savitha Mamidiyala)

(13th December 2019)

© (2019 and Savitha Mamidiyala)

**ABSTRACT**

Every living being strives for being healthy. Health plays a vital role in humans. If sick, people visit hospitals for the treatment. In current situation, people are being diagnosed by many diseases. Few of diseases are treatable with few drugs and few diseases are not such patients are associated with an increased risk of death.

In the paper, we will be working on the Healthcare data i.e., MIMIC-III (‘Medical Information Mart for Intensive Care’) to analyze the mortality of each patient using high risk medications based on the electronic patients record. To analyze the mortality of each patient we will be using machine learning algorithms such as logistic regression, random forest model and Xgboost model.

In this paper, we will discuss how these models are used to predict the high risk drugs with respective to mortality and also show the features that cause high risk in a generalized way. The primary aim of this study was to develop a mortality prediction model for identifying the high risk medication from health care data.

TABLE OF CONTENTS

[**ABSTRACT** 3](#_Toc25607767)

[introduction 6](#_Toc25607768)

[DATASET 7](#_Toc25607769)

[MIMIC DATASET 7](#_Toc25607770)

[Demographic Data: 8](#_Toc25607771)

[Drugs: 8](#_Toc25607772)

[ICD9 Codes: 9](#_Toc25607773)

[Machine learning algorithms 9](#_Toc25607774)

[Logistic Regression 10](#_Toc25607775)

[Random Forest Model 10](#_Toc25607776)

[XGBoost Model 10](#_Toc25607777)

[results 11](#_Toc25607778)

[Performance Evaluation Metrics 11](#_Toc25607779)

[Logistic Regression: 13](#_Toc25607780)

[Random Forest Model: 14](#_Toc25607781)

[XGBoost Model: 15](#_Toc25607782)

[Feature Importance of XGBoost 16](#_Toc25607783)

[**SHAP (SHapley Additive exPlanations)** 17](#_Toc25607784)

[Conclusion 20](#_Toc25607788)

[APPENDIX A References 21](#_Toc25607789)

introduction

In developed countries, patient records are mostly electronically saved and with patient’s approval, this data could be used for research purposes. Patient electronic records will contain a lot of information and this information could be used in analyzing the risk to mortality using Machine Learning techniques. Several models can be established using the data to determine which factors have a greater impact on individual’s mortality.

In this paper, we use data from about 60000 patients and develop machine learning models such as Logistic Regression, Random Forest Model and XGBoost model to assess various factors and their impact on individuals health.

DATASET

# MIMIC DATASET

The dataset used in this paper is the latest version of MIMIC i.e., MIMIC-III v1.4, which consists of 61,532 intensive care unit stays including 53,432 stays for adult patients and 8,100 for neonatal patients. The data obtained is from June 2001 - October 2012. The dataset is unstructured where the data needs to be cleaned and organized.

In this paper, we mainly concentrate on demographics, high-risk medicated drugs and ICD9 codes (International Classification of Diseases).

## Demographic and Admission Data:

This data consists of hospital information of each patient such as number of diagnosis, lab works, procedures, etc.

Table 1. Demographic Data

|  |
| --- |
| ETHNICITY |
| ADMISSION\_TYPE |
| ADMISSION\_LOCATION |
| INSURANCE |
| DIAGNOSIS |
| DISCHARGE\_WARDID |
| HOSPITAL\_EXPIRE\_FLAG |
| EXPIRE\_FLAG |
| Age |
| GENDER |
| PATIENTWEIGHT |
| num\_emergency |
| num\_inpatient |
| LOS |
| num\_lab\_procedures |
| num\_procedures |
| num\_medications |
| number\_diagnoses |
| Readmit |

## Drugs:

From AGS Beers Criteria we use lists of medications to be avoided in older adults and drugs that need the dose adjusted based on the individual's kidney function and select drug–drug interactions that are associated with harms in individuals.

## ICD9 Codes:

The International Classification of Diseases (ICD) is the international standard [diagnostic](https://en.m.wikipedia.org/wiki/Diagnosis) tool for [epidemiology](https://en.m.wikipedia.org/wiki/Epidemiology), [health management](https://en.m.wikipedia.org/wiki/Health_management) and clinical purposes and it is maintained by is maintained by the [World Health Organization](https://en.m.wikipedia.org/wiki/World_Health_Organization) (WHO).

Table 2. ICD9 Codes - Diseases

|  |
| --- |
| Diseases\_ certain conditions originating in the perinatal period |
| Diseases\_ complications of pregnancy, childbirth, and the puerperium |
| Diseases\_ congenital anomalies |
| Diseases\_ diseases of the blood and blood-forming organs |
| Diseases\_ diseases of the circulatory system |
| Diseases\_ diseases of the digestive system |
| Diseases\_ diseases of the genitourinary system |
| Diseases\_ diseases of the musculoskeletal system and connective tissue |
| Diseases\_ diseases of the nervous system and sense organs |
| Diseases\_ diseases of the respiratory system |
| Diseases\_ diseases of the skin and subcutaneous tissue |
| Diseases\_ endocrine, nutritional and metabolic diseases, and immunity disorders |
| Diseases\_ infectious and parasitic diseases |
| Diseases\_ injury and poisoning |
| Diseases\_ mental disorders |
| Diseases\_ neoplasms |
| Diseases\_ symptoms, signs, and ill-defined conditions |
| Diseases\_external causes of injury and supplemental classification |

Machine learning algorithms

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

Supervised machine learning builds a model that makes predictions based on evidence in the presence of uncertainty. A supervised learning algorithm takes a known set of input data and known responses to the data (output) and trains a model to generate reasonable predictions for the response to new data.

In this paper, we will be discussing 3 machine learning techniques i.e., Logistic Regression, Random Forest Model and XGBoost model.

## Logistic Regression

Logistic regression is used to estimate the probability that an event will occur or that a patient will have a particular outcome using information or characteristics which are related to or influence such events.  In this paper we focus on the use of logistic regression to create models for predicting patient outcomes.

The outcome of each patient can only have 2 values (e.g., alive vs die) are called binary or dichotomous.

## Random Forest Model

A decision tree is a decision support tool where a tree-like graph or model of decisions and their possible consequences, such as, chance event outcomes, resource costs, and utility are used.

Random forest builds multiple decision trees and combines them thereby getting a more accurate and stable prediction. Random forest can be used for both classification and regression problems, that form the majority of current machine learning systems.

## XGBoost Model

XGBoost stands for eXtreme Gradient Boosting. It is a software library that can be accessed from a variety of interfaces. It is an implementation of gradient boosting machines. The library is focused on computational speed and model performance. XGBoost dominates structured or tabular datasets on classification and regression predictive modeling problems.

results

## Performance Evaluation Metrics

The metrics used to evaluate the performance of a model are Accuracy, Precision, Recall & F1 Score. Confusion metrics include true positive, true negative, false positive and false negative. True positive and true negative are the observations that are correctly predicted in Confusion metrics. We want to minimize false positives and false negatives.

True Positives (TP) - These are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also yes.

True Negatives (TN) - These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no.

False Positives (FP) – When actual class is no and predicted class is yes.

False Negatives (FN) – When actual class is yes but predicted class in no.

**Accuracy:**

Accuracy is a performance measure that is a ratio of correctly predicted observation to the total observations. Accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost same. So high accuracy for a model does not mean that the model is the best. Other parameters need to be determined to evaluate the performance of a model.

Accuracy = (TP+TN)/(TP+FP+FN+TN)

**Precision:**

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.

Precision = TP/(TP+FP)

**Recall:**

Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes.

Recall = TP/(TP+FN)

**F1 score:**

F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. F1 is usually more useful than accuracy, especially if you have an uneven class distribution.

F1 Score = 2\*(Recall \* Precision) / (Recall + Precision)

**AUC – ROC Curve:**

An **ROC curve** (**receiver operating characteristic curve**) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters that are TP and FP. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives.

**AUC** ( Area under the ROC Curve) measures the entire two-dimensional area underneath the entire ROC curve. AUC provides an aggregate measure of performance across all possible classification thresholds.

## Logistic Regression:

Logistic Regression model is created on training data with the ratio of 80 to 20 i.e., 25208 records and tested on 6302 records resulting in an accuracy of 76.21%. In this case we can notice that precision and recall are unequal but f1- score is nearly equal which means we are working on the balanced data which indicates that the accuracy obtained in correct.

![A screenshot of a cell phone

Description automatically generated]()

Figure 1. Performance metrics by Logistic regression

The AUC- ROC curve of logistic regression is

A screenshot of a map

Description automatically generated

Figure 2. AUC-ROC curve of Logistic regression

## Random Forest Model:

Random forest model is created on training data with the ratio of 80 to 20 i.e., 25208 records and tested on 6302 records resulting in an accuracy of 73.70%.

A screenshot of a cell phone

Description automatically generated

Figure 3. Performance metrics by Random Forest model

The AUC-ROC curve of the random forest model is

A screenshot of a map

Description automatically generated

Figure 4. AUC-ROC curve of Random Forest model

## XGBoost Model:

Random forest model is created on training data with the ratio of 807 to 20 i.e., 25208 records and tested on 6302 records resulting in an accuracy of 76.40%. In this case we can notice that precision and recall are unequal but f1- score is nearly equal which means we are working on the balanced data which indicates that the accuracy obtained in correct.

A screenshot of a cell phone

Description automatically generated

Figure 5. Performance metrics by Xgboost model

The AUC-ROC curve of XGBoost is

A screenshot of a cell phone

Description automatically generated

Figure 6. AUC-ROC curve of XGBoost model

## Feature Importance of XGBoost

 Feature importance provides a score that indicates how useful or valuable each feature is in the construction of the boosted decision trees within the model. The more an attribute is used to make key decisions with decision trees, the higher its relative importance.

A screenshot of a cell phone

Description automatically generated

Figure 7. Feature importance of XGBoost

## **SHAP (SHapley Additive exPlanations)**

 SHAP explains the output of any machine learning model. The below image shows the fuctionality of SHAP.

A close up of a sign

Description automatically generated

Figure 8. Architecture of SHAP

The below image shows features that are contributing to push the model output from the base value to the model output. Features contributing the prediction higher are shown in red, the prediction lower are in blue. From the image, we can state that Age, num\_lab\_procedures, number\_diagnoses and patientweight are contributing higher prediction to push the model. Discharge\_wardid, Drug\_metoclopramide and Admission\_type are contributing lower prediction to push the model.

A screenshot of a cell phone

Description automatically generated

Figure 9. Features contributing to push the model

We can plot the SHAP values of every feature for every sample and thereby getting an overview of which feature is more important. The plot below sorts features by the sum of SHAP value magnitudes over all samples, and uses SHAP values to show the distribution of the impacts each feature has on the model output. From the below image, it can be observed that higher the age group, the higher is the Drug\_scopolamine\_patch consumed by the people .The lower the patients weight(PATIENTWEIGHT) is the lower they are diagnosed with diseases(number\_diagnoses).

A screenshot of a social media post

Description automatically generated

Figure 10. Overview of SHAP values of features

To understand how a single feature effects the output of the model we can plot the SHAP value of that feature vs. the value of the feature for all the examples in a dataset. From the below image we can observe that the average number of people with lower age has less impact with the number of diagnoses

A screenshot of a cell phone

Description automatically generated

Figure 11. SHAP values of Age vs number of diagnoses

The below image shows the mean absolute value of the SHAP values for each feature to get a standard bar plot.

A screenshot of a social media post

Description automatically generated

Figure 12. The mean absolute value of the SHAP values

Conclusion

In this paper, we have worked on the HealthCare dataset that is obtained from MIMIC data. The unstructured data was organized and cleaned for applying machine learning algorithms. The main motive of the study was to analyze the risk of each patient towards mortality based on the high risk medications from electronic patients records. We have applied Logistic Regression, Random Forest Model and XGBoost Model to analyze the data and noticed that XGBoost is performing better among the other two.

APPENDIX A  
References

Using electronic health record collected clinical variables to predict medical intensive care unit mortality; Jacob Calvert Ann Med Sur (Lon), 2016 Nov;11:52-57

1. <https://medium.com/berkeleyischool/how-to-use-machine-learning-to-predict-hospital-readmissions-part-1-bd137cbdba07>
2. <https://www.guidelinecentral.com/summaries/american-geriatrics-society-2015-updated-beers-criteria-for-potentially-inappropriate-medication-use-in-older-adults/#section-society>
3. <https://en.m.wikipedia.org/wiki/List_of_ICD-9_codes>
4. <https://jamanetwork-com.libproxy.uncg.edu/journals/jama/fullarticle/2540383?resultClick=1>
5. <https://www.researchgate.net/publication/321375022_Predicting_readmission_risk_from_doctors%27_notes>
6. <https://ieeexplore.ieee.org/document/7737956>