Predicting Hospital Readmission for Diabetes Patients Using machine learning

[1] D.Sathyavathi, [2] Dr.A.MARY SOWJANYA
[1,2]Department of CS&SE, Andhra University College of engineering (A), Andhra University, Visakhapatnam, AP, India.

ABSTRACT: One of the most critical problems in healthcare is predicting the likelihood of hospital readmission in case of chronic diseases such as diabetes. Finding readmission in primary stage, allows the hospitals to give special care for those patients, and then can reduce the rate of readmission. In this work we have developed a new model using machine learning. Since hospital readmissions increase the healthcare costs and negatively influence hospitals' reputation, predicting readmissions in early stages allows prompting great attention to patients with high risk of readmission, which in turn leverages the healthcare system and saves healthcare expenditures. Machine learning helps in providing more accurate predictions than current practices. In this work, we try to predict hospital readmission rate of diabetic patients using standard scaler for preprocessing, decision tree for checking train accuracy and test accuracy, random forest for classification, CATboost that deals with categorical feature and XGBoost classifier .A combination of Machine learning and data engineering were found to outperform other machine learning algorithms when employed and evaluated against real life data. We carry out this process by a number of modules that include feature extraction which improves the process of analysis and helps in the formation of an efficient prediction summary.

KEYWORDS: Classification, Diabetes Readmission, Feature Selection, Healthcare Analytics, Machine Learning.

I. INTRODUCTION

Diabetes is a wide spread chronic disease that is accompanied with irregularities of blood glucose levels due to problems related to insulin. The number of people with diabetes in the world has risen from 108 million in 1980 to 422 million in 2014. The prevalence of diabetes is growing most rapidly in low- and middle-income countries [1]. In Jordan for example, the prevalence of Type 2 diabetes was around 17.1% in 2004 with 30% increase in a decade which is a dramatic increase [2].We should be alarmed at the increasing and aggressive growth rate of diabetes related cases and the staggering diabetes related costs. As a possible direct consequence of diabetes increasing, the number of hospital inpatients readmissions continues to rise.

A hospital readmission is when a patient who has been discharged from hospital is readmitted again within a certain time period. Hospital readmissions are now a metric for hospital quality (CMS, 2019). Centers for Medicare & Medicaid Services created the Hospital Readmissions Reduction Program with an aim to improve quality of care and reduce healthcare spending. As hospital readmissions have increased inline with the prevalence of diabetes, it is likely that it may continue to do so compounding the problem (Rubin et al., 2014).

Hospital readmission is expressed by the time that a patient takes before getting back to the hospital. Readmission is considered a quality measure of hospital performance as well as a mean to reduce healthcare costs. Hospitals are financially penalized when the permitted rate of 30-day readmissions is exceeded. The Medicare Payment Advisory Commission in the US estimated that 12% of readmissions can be avoided. Preventing 10% of readmissions would save Medicare in the US more than $1 billion [3]. For diabetes; the cost analysis estimates that $250 million can be saved across 98,000 diabetic patients by incorporating predictive modeling and prompting greater attention to those who were predicted to get readmitted [4]. Current practices to identify at-risk diabetic patients are subjective, a clinician will assess the patient and decide what the appropriate care plan for that individual is. Research has shown that these methods for determining readmission are slightly better than random guessing [5]. On the other side, machine learning plays a vital role in many predicting tasks. Hence, predicting hospital readmissions using machine learning sounds a worth implementing approach. This work shows machine learning as an better approach for predicting diabetic patients' readmissions.

II. MOTIVATION AND BACKGROUND

Early detection of diabetes can reduce and delay diabetes. This can be achieved with exercising, healthy eating, not smoking and by maintaining a healthy body weight. The later the detection of the disease, the worse the diagnosis outcome. The contributing factors such as inactivity and obesity are non-genetic contributing factors. Diabetes can be a trade-off between healthy living comprising of healthy eating, exercising versus convenient, demanding and hectic lifestyles. Undiagnosed diabetes can overtime damage the heart, blood vessels, eyes, kidneys, feet and nerves and increase the risk of heart disease and stroke. Uncontrolled diabetes in pregnancy can have a detrimental effect on mother and baby, with increased chance of fatal loss, malformations, still birth, perinatal death and complications. Gestational diabetes increases the risks of complications before, during and after delivery.

The beneficiaries of this project are twofold, the patient themselves who will benefit in terms of disease management, overall health and early detection. The health service providers will gain, they will have a better understanding of the data where action can be taken to reduce early readmissions associated with the patient diagnosis. Early detection and treatment are essential in order to provide better treatment to patients and potentially saving lives and reducing readmitted patients treatment healthcare costs. The Diabetic patients Datasets of 130-US hospitals for years 1999-2008 datasets have been selected as they are more contemporary with several cultures and age profiles ranging from 0-100. This research focuses on datamining techniques to develop predictive models for classifying diabetes patients by predicting diabetes readmission, short term (within 30days) or long term(after 30 days)and predicting diabetes diagnosis. This work seeks to incorporate a higher recall (sensitivity) as this is suited to the healthcare industry. In healthcare it is important to predict a result but more so to have the correct patient result when a patient is suffering from diabetes (true positives). This will ensure no patient is left untreated. Thus, the research metrics include recall and accuracy

III. RELATED WORK

In most previous papers, manual data extraction and neural networks are the most common method of classification. Piyush Jain et al. made use of Electronic Health Records, applying Naive Bayes classifier on patient data [1] to discover that parallel computing reduces time while also maintaining the overall model performance. They used measures such as recall, precision, and cluster time. Christopher Baechle et al. performed a clinical analysis on the same dataset earlier in 2017, using Naive Bayes and clinical NLP [5] to get a result of 51.93% average savings. Haishuai Wang et al. obtained two datasets from the Barnes Jewish Hospital's general hospital ward and operating room data to predict hospital readmission using cost-sensitive deep learning [2].

In general, Neural networks (NNs) are common in medical classification research. A review by Dreiseitl and Ohno-Machado [6] to some implemented models in medical classification tasks showed that logistic regression (LR) followed by NNs are the most popular classifiers in medicine. Authors in [6] reported that both LR and shallow NNsperform on about the same level more often than not. However, the superior popularity of LR was attributed to the interpretability of model parameters and the ease of use. An obstacle for neural networks is the black-box property that hinders the model interpretability.

Various published papers studied readmission rates of diabetic patients [7-10]. Some studies used machine learning models to predict the risk of all-cause readmissions among patients with diabetes. Bhuvan et al.[4] compared different classifiers that were applied to this problem (same dataset) such as Naïve Bayes, Bayes Network, Random Forest, Adaboost Trees, and shallow NNs.Bhuvan et al.[4] showed that the performance of Random Forest and shallow NNs outperforms other classifiers with a slight preponderance in Random Forest's favor. However, only a single hidden layer of neural network was used.

IV. MACHINE LEARNING

Machine learning is a branch of Artificial intelligence that is concerned with the design and development of algorithms and it enables today's computers to have the property of learning.Machine learning is gradually growing and becoming a critical approach in many domains such as health,education and business.

DECISION TREE

One class of popular machine learning models is tree-based methods. The simplest tree-based method is known as a decision tree.Decision tree is the most powerful and popular tool for classification and prediction.A tree can be "learned" by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The final prediction is then the fraction of positive samples in the final leaf (final split) of the treeDecision trees can handle high dimensional data. In general decision tree classifier has good accuracy. Decision tree induction is a typical inductive approach to learn knowledge on classification

RANDOM FOREST

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction

The fundamental concept behind random forest is a simple but powerful one — the wisdom of crowds. In data science speak, the reason that the random forest model works so well is:

A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.

GRADIENT BOOST CLASSIFIER

Gradient Boosting involves creating and adding decision trees to an ensemble model sequentially. New trees are created to correct the residual errors in the predictions from the existing ensemble.Due to the nature of an ensemble, i.e having several models put together to form what is essentially a very large complicated one, makes this technique prone to overfitting. The **eta** parameter gives us a chance to prevent this overfitting

CATBOOST CLASSIFIER

CatBoost is an algorithm for gradient boosting on decision trees. It would deal with categorical features automatically. It can easily integrate with deep learning frameworks like Google's TensorFlow and Apple's Core ML. It can work with diverse data types to help solve a wide range of problems .

IV. METHODOLOGY

DATASET DESCRIPTION

The dataset of this work [14, 15] consists of 100,000 medical records for 70,000 patients with diabetes collected from 130 hospitals in the USA over 10-years' period from 1999 to 2008. Medical records in the dataset include 50 attributes that are the risk factors, in addition to a label indicating the readmission status of a patient indicates whether a patient was readmitted to the hospital in 30 days or not. The dataset encounters satisfy the following conditions: It is an inpatient encounter (a hospital admission) It is a diabetic encounter, that is, one during which any kind of diabetes was entered to the system as a diagnosis.

- The length of stay was at least 1 day and at most 14 days.
- Laboratory tests were performed during the encounter.
- Medications were administered during the encounter

DATA ENGINEERING

Machine learning approach allows a machine to learn from raw data. However, learning directly from raw data entails a large number of training examples. When data are not sufficient for representation learning, data engineering turns out to be essential to overcome the shortage of data. In this work, we balance between learning capabilities, and the size of the data.

Feature selection: The proposed approach in this work assumes feature selection as part of the learning model. Feature selection is a process where you automatically or manually select those features which contribute most to

your prediction variable or output in which you are interested in .Feature selection reduces overfitting, improves accuracy and reduces training time

Feature creation: To compensate for the lost information from dropped records, additional features were created before eliminating the useless attributes. The first feature is the number of medications and the second feature is the number of changes in the medications. Both features are extracted from the drug attributes.

Feature transformation: Categorical data were transformed into binary encoding or One-hot encoding. And since model stability and parameter estimate convergence are influenced when multi-scaled variables are used.

Imbalanced data: The problem of imbalanced data is one of the obstacles for many Machine Learning (ML) algorithms, it arises when the data are dominated by a majority class and a minority class is rarely detected. As a result, the classifier performance on the minority may be insufficient when compared to the majority. For example, a dumb classifier that always predicts the majority class can achieve high accuracy. There are two main methods to deal with imbalanced data .They are Under-sampling and Over-Sampling methods. Under-sampling methods balances the classing by eliminating great portion of the majority class. Over-sampling methods such as Synthetic Minority

Handling duplicate records: Out of multiple records for the same patient, a single record is kept and the remaining were deleted which led to reducing the number of records to ~70,000. The first record is chosen because it has the highest probability of readmission which helps in balancing the data. For example, all last encounters are labeled by the minority class (not readmitted) which would aggravate the problem of Imbalanced data.

In order to make sure that the model generalizes and does not over-fit, the data were split into two parts: 20% for testing and 80% for training and development. The latter is divided repeatedly into 80% for training and 20% for cross validation. The test set is hold out to evaluate the performance of the model. Early stopping technique was used to avoid overfitting, when the validation error increases for a specified number of iterations, the training is stopped.

V. RESULTS

 # corr plot for numerical columns

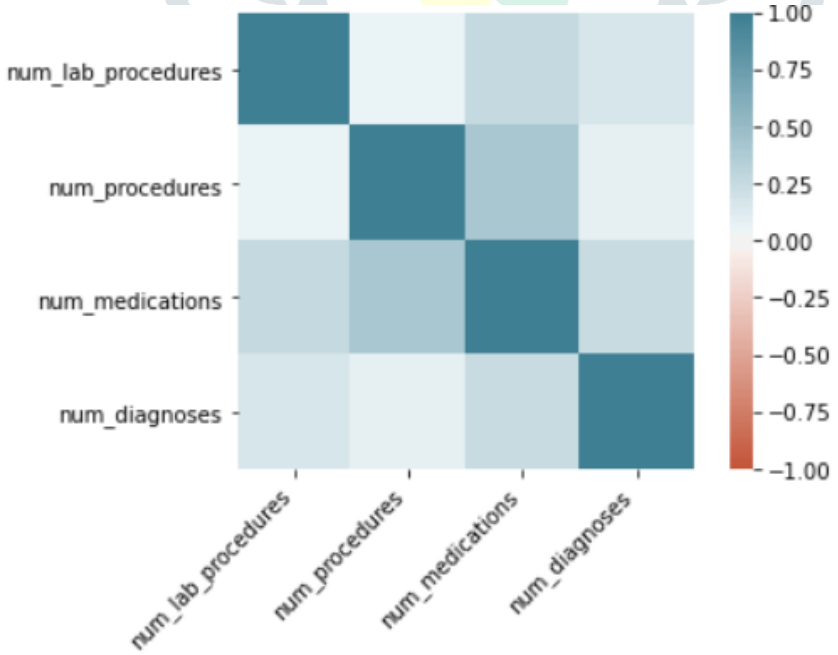

Fig : 1

 # Bar plot for admissions age wise

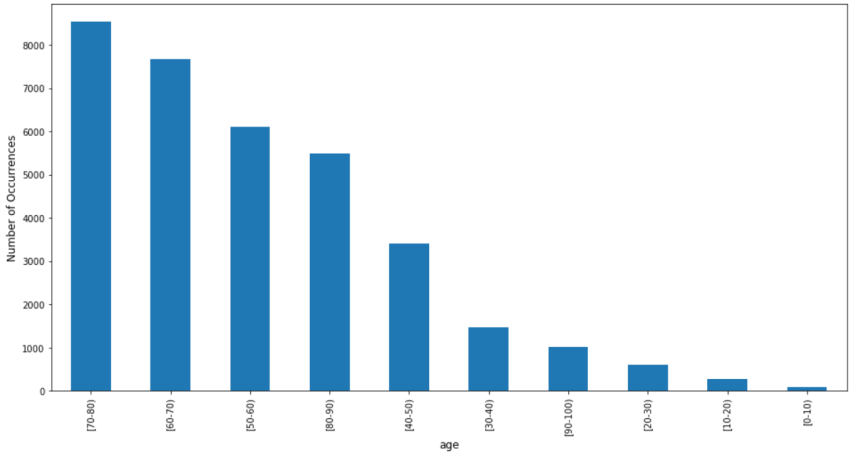

Fig : 2

 # Re-admitted vs Insulin

insulin	Down	No	Steady	Up
readmitted				
NO	2904	15036	9233	2718
Within30days	580	2132	1554	493

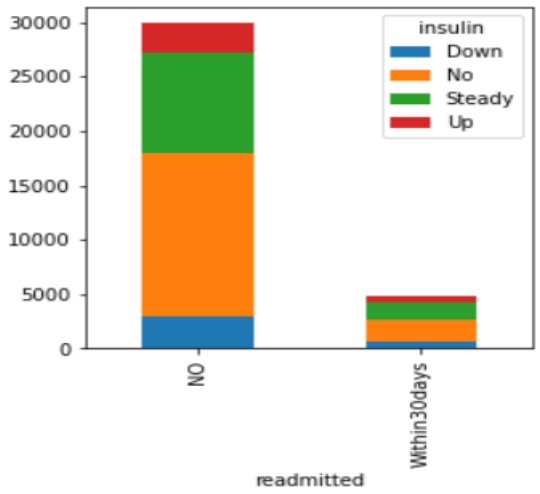

Fig : 3

 # Decision tree

```
Text(0, 0.5, 'Recall/Accuracy')
```

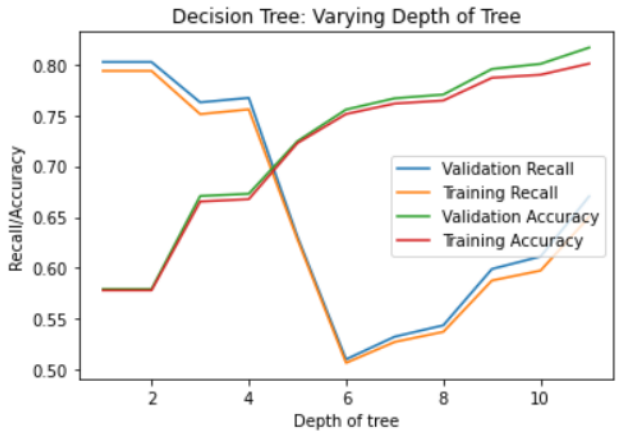

Fig : 4

 # Random forest

```
Metrics on Train Data:::::::
Confusion Matrix:
 [[15346  2626]
 [ 1935 15960]]
Accuracy: 0.872835754314551
Precision: 0.8587108576347788
Recall: 0.8918692372170998
Classification Report:
              precision    recall  f1-score   support

           0     0.8880    0.8539    0.8706     17972
           1     0.8587    0.8919    0.8750     17895

    accuracy                         0.8728     35867
   macro avg     0.8734    0.8729    0.8728     35867
weighted avg     0.8734    0.8728    0.8728     35867


Metrics on Validation Data:::::::
Confusion Matrix:
 [[3692  753]
 [ 628 3894]]
Accuracy: 0.8459908553585369
Precision: 0.8379599741768883
Recall: 0.8611233967271119
Classification Report:
              precision    recall  f1-score   support

           0     0.8546    0.8306    0.8424      4445
           1     0.8380    0.8611    0.8494      4522

    accuracy                         0.8460      8967
   macro avg     0.8463    0.8459    0.8459      8967
weighted avg     0.8462    0.8460    0.8459      8967
```

Fig : 5

 # Readmitted vs no of medications

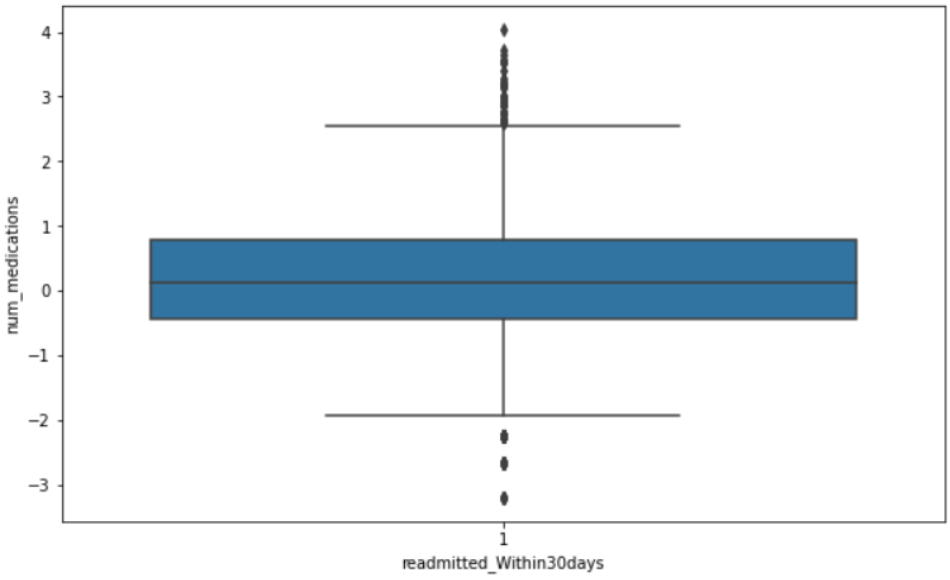

Fig : 6

 # Feature importance

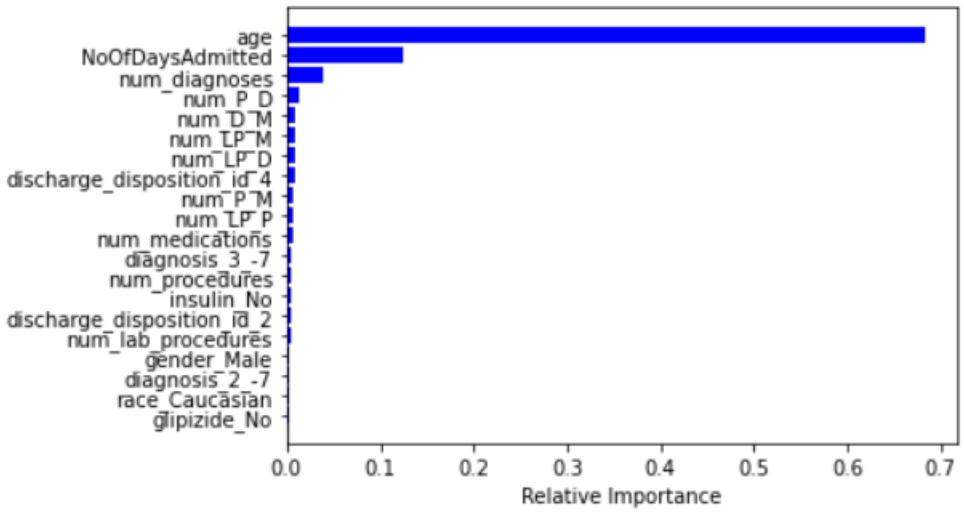

Fig : 7

VI. CONCLUSION

Hospital readmissions raise health care costs and negatively influence hospitals' reputation. Hence, predicting hospital readmissions among diabetics is of great interest. This paper presented deep learning as an effective approach in predicting hospital readmissions among diabetic patients. A combination of machine learning and data engineering were found to outperform other machine learning algorithms when employed and evaluated against real life data. It is clear that reduction in the number of features has substantially helped in improving the accuracy of the machine learning models in this scenario of readmission prediction.

Apart from reducing the features with filter methods of feature selection and consultation, wrapper methods i.e. forward selection, backward elimination, and bi-directional elimination techniques could be utilized in Scikit-learn and applied on the best performing model from the previous module to test if accuracy improves further. Consulting with diabetic specialists helped us realize that some other features such as family history of the patient, and date of admission, may additionally be worth collecting. It was found that the dosage amount of drugs would have been helpful data for predicting readmission. Also, almost 98% of patients are not administered drugs according to the dataset; a possible reason for this could be that the data only records drug administration while admitted in the hospital, whereas most drugs are usually prescribed to be taken at home. In conclusion, this method of analysis could be applied to different clinical datasets other than diabetes, for the purpose of readmission prediction

REFERENCES

[1] World Health Organisation. (2016). Global Report on Diabetes.
[2] Ajlouni, Kamel, Yousef S. Khader, Anwar Batieha, Haitham Ajlouni, and Mohammed El-Khateeb. (2008). "An increase in prevalence of diabetes mellitus in Jordan over 10 years." Journal of Diabetes and its Complications, 22(5): 317–324.
[3] Medicare Payment Advisory Commission. (2007). Report to the Congress promoting greater efficiency in Medicare. Washington, DC.
[4] Bhuvan, Malladihalli S., Ankit Kumar, Adil Zafar, and Vinith Kishore. (2016). "Identifying diabetic patients with high risk of readmission."
[5] Allaudeen, Nazima, Jeffrey L. Schnipper, E. John Orav, Robert M. Wachter, and Arpana R. Vidyarthi. (2011). "Inability of Providers to Predict Unplanned Readmissions." Journal of General Internal Medicine 26(7):771–76.
[6] Dreiseitl, Stephan and Lucila Ohno-Machado. (2002). "Logistic Regression and Artificial Neural Network Classification Models: A Methodology Review." Journal of Biomedical Informatics 35(5):352–59.
[7] Silverstein, Marc D, Huanying Qin, S Quay Mercer, Jaclyn Fong and Ziad Haydar.(2008). "Risk Factors for 30-Day Hospital Readmission in Patients 65 Years of Age." Baylor University Medical Center Proceedings21(4):363–72.
[8] Jiang, H.Joanna, Daniel Stryer, Bernard Friedman, and Roxanne Andrews. (2003). "Multiple Hospitalizations for Patients with Diabetes." Diabetes Care 26(5):1421–26.
[9] Strack Beata, DeShazo Jonathan, Gennings Chris, Olmo Juan, Ventura Sebastian, Cios Krzysztof, and John N. Clore. (2014). "Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records." BioMed Research International.
[10] Yifan, Xing and Jai Sharma. (2016). Diabetes Patient Readmission Prediction Using Big Data Analytic Tools. [11] Mingle, Damian. (2017). "Predicting Diabetic Readmission Rates: Moving Beyond Hba1c." Current Trends in Biomedical Engineering & Biosciences 7(3):555707. 007.
[12] Lecun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep Learning." Nature 521(7553):436–444.