International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 02 | Feb 2025 www.irjet.net p-ISSN: 2395-0072

Predicting the Impact on Re-admission Rates for Hospitalized Diabetic Patient

Achala Harsha[1], Chethan Nazre S[2], Chethana M[3], Hithaishi M[4], Nithin K[5]

[1] *Department of Information Science, Malnad College of Engineering, Hassan, Karnataka, India*
[2] *Department of Information Science, Malnad College of Engineering, Hassan, Karnataka, India*
[3] *Department of Information Science, Malnad College of Engineering, Hassan, Karnataka, India*
[4] *Department of Information Science, Malnad College of Engineering, Hassan, Karnataka, India*
[5]*Assistant Professor, Department of Information Science, Malnad College of Engineering, Hassan, Karnataka, India*

Abstract — *Hospital readmissions among diabetic patients pose serious health risks and financial burdens. This study uses machine learning to predict 30-day readmissions, with XGBoost achieving the highest accuracy (94%). Key factors like inpatient visits, hospital stay duration, and diagnoses play a crucial role in readmission risk. These insights can help improve patient care and reduce unnecessary hospital visits.*

Keywords— Hospital readmission, diabetes, machine learning, XGBoost, predictive modeling, inpatient visits, hospital stay duration, medication changes, healthcare analytics.

1. INTRODUCTION

In the past decades, readmissions to the hospitals have become an aspect of the retrospectives and prospective research that sought to eliminate it from the hospitals [1]. A patient who gets readmitted in a hospital within a specified period after he or she was discharged from the same hospital is referred to as a hospital readmission. The occurrence of readmission to the hospital for some selected diseases in particular shows the standard of the hospital. In other word, it shows that the first admission did not give adequate care to the patient and hence the life of the patient is at risk. Furthermore, the cost of care is negatively affected by the increased rate of hospital revisits. More specifically, 30-day hospital readmission rates were relatively high among older and higher risk patients [2]. It stated that venting factors would cost the American hospitals more than $26 billion on average on each patient. Instead of Americans, patients with diabetes, they have a greater risk of incurring more costs. Out of all the expenses the US had on diabetic patients in 2011, $41 billion was incurred by patients who were readmitted within 30 days 4. Advocating for higher standards and minimizing unnecessary expenses, the United States congress enacted the Hospital Readmission Reduction Program (HRRP). Consequently, beginning in October 2012, the Centers for Medicare, and Medicaid Services (CMS) initiated policies that financially minimize repayment incapability hospitals.

Addressing this critical issue involves the intensive data analysis throughout the research process. This study is a secondary analysis using machine learning methods. Our goal of the analysis is to find the determining factors that lead to higher readmission and correspondingly being able to predict which patients will get readmitted. Therefore, we proposed two research questions:

1) What approaches can we utilize to effectively predict hospital readmission within this dataset?

2) Which factors are the most significant indicators of hospital readmission among diabetic patients?

The remainder of this paper is structured as follows: In Section 2, we provide a concise summary of previous research and highlight the existing gaps in the literature. Section 3 will detail the methodology employed in this study, encompassing the description of the dataset and the analytical procedures. This entails data processing, exploratory analysis, feature engineering, as well as modeling and evaluation. Section 4 presents the results and discussion in relation to each research question, followed by the conclusion and recommendations for future work in Section 5.

2. RELATED WORK

Numerous prior investigations have examined the risk factors associated with readmission rates across various disease types. For instance, one study [6] conducted a broad analysis aimed at predicting hospital readmissions without concentrating on a specific illness.

In the context of diabetic patients, other research efforts [7] [8] [9] have concentrated on subsets of the diabetic population and utilized smaller datasets. When assessing readmission rates, certain studies have emphasized the role of demographic and socioeconomic variables that may affect these rates [10]. For example, research by [11] highlighted age as a significant factor, revealing that both acute and chronic glycemic control impacted readmission risk for individuals aged 65 and older, based on data from 29,000 patients. Additionally, [12] investigated the correlation

between the likelihood of readmission and the primary diagnosis by measuring HbA1c levels.

Among the recent studies, [13] predicted diabetes with high risk of readmission through modeling multivariate patient medical records using machine learning classifiers such as Naïve Bayes, Bayesian Networks, Random Forest, Adaboost and Neural Networks. To contribute to the implementation of work into the real world, a cost analysis is used to determine the effective cost.

Similarly, [Mingle] addressed the previous research gap that no typical performance metrics of machine learning classifies is documented. This research contributes to the field in several significant ways:

1) It advances the identification and validation of risk factors associated with readmission rates. Previous literature suggests that understanding these factors can be instrumental in formulating protocols aimed at enhancing inpatient care.

2) It investigates previously unexplored machine learning algorithms to enhance the precision of predictive performance

3. METHODOLOGY

In this section, we will provide a description of the dataset, the exploratory data analysis, feature engineering, modeling, and evaluation.

3.1 Data Set

To explore this problem, we used a secondary dataset from UCI machine learning repository [14] dataset. The dataset includes 101,766 instances, representing 10 years (1999- 2008) of clinical care at 130 US hospitals and integrated delivery networks across the Midwest (18 hospitals), Northeast (58), South (28), and West (16). Most of the hospitals (78) have bed size between 100 and 499, 38 hospitals have bed size less than 100, and bed size of 14 hospitals is greater than 500. The features collected in the dataset are related to patient's demographic information such as race, gender, age, weight; the information related to their hospital diagnosis and treatment, such as num_lab_procedures, num_medications, num_outpatient, diagnosis, and medication prescription. The dataset is just an extracted subset set of Health fact dataset. Given this is an open dataset that include the longitudinal and cross-sectional data, and with relatively complete attributes (55 attributes), and released in the recent year (2014), we chose the dataset for exploring the questions.

3.2 Exploratory Analysis

Before undertaking any formal analysis, we engaged in exploratory data analysis to examine the data types, attributes, and overarching patterns present within the dataset. Our primary focus was on the class label "Readmitted" (refer to Fig 1), prompting us to investigate the distribution of readmissions alongside various categorical variables. To explore the relationships among numerical variables, we employed scatter plots to illustrate their interconnections and distributions (refer to Fig 2).

Figure 1: Bar plot for the class label

Figure 2: Scatter plot for numeric features.

3.3 Data Pre-Processing

Following the exploratory analysis, we identified multiple challenges present in the original dataset. Consequently, it is essential to undertake various data wrangling tasks, including data cleaning, addressing missing values, generating new variables, and performing data transformation prior to modeling. The tools employed for this purpose include Python packages such as Numpy, Pandas,

Matplotlib, and Seaborn. We executed several pre-data processing procedures.

3.3.1 Dealing with Missing Data

We discovered many missing values coded as "?" across nominal variables. As Table 1 shows, this dataset has 8 variables which contain missing values. Since weight, medical_specialty, and payer_code contains over 35% values, and because of the irrelevancy toward our study, we decide to drop all of them. Race only includes 2.23% missing values, so we only drop the missing values and keep the rest. Primary (diag_1), secondary (diag_2) and additional. (diag_3) diagnoses each has less than 2% missing values, but compared to the total number of instances, we still need to clean them. Technically, our goal is to maintain the most information of the dataset, especially the diagnosis is an important variable related to the diabetes patients. Therefore, we adopted a strategy to drop the missing values when all three diagnoses were missing. We then only drop 3 unknown and invalid instances in our dataset.

column	count_missing	percent_missing
weight	98569	96.86
Medical_specialty	49949	49.08
Payer_code	40256	39.56
race	2273	2.23
Diag_3	1423	1.4
Diag_2	358	0.35
Diag_1	21	0.02
gender	3	.00003

Table1: Variables with missing values.

3.3.2 Dropping Attributes

After a quick view of the current dataset, we found some patients died during the hospital admission who do not have any probability of being readmitted, so we removed those tuples, as the discharge_disposition_id=11. We also drop two variables (drugs named citoglipton and examide) in which all records have the exactly same value. By noticing that two variables called encounter_id and patient_nbr has no relevance with the class label readmission, so we also drop those two variables.

3.3.3 Creation of New Features

1) patient_service: We created a new feature called patient_service, which measures the total number of hospital/clinician services a patient used in the past year. This feature is the sum of original variables for number of inpatient visits, emergency room visits, and outpatient visits. We did not apply weighting for these three variables. The reason for the creation is to lower the dimension of our data and try to make the dataset simpler.

2) med_change: The dataset contains 23 medications of the medicine use for a patent during the stay in hospital. Each of the tuple records when a change was made in this medication or not during the current stay as No-for no medication, Up-for increasing the dose, Down-for decreasing the dose and Steady-for keeping the current dose. Instead of counting changes for each medication, we decide to combine them and count changes for all of them.

3) We define No and Steady as no change, while up and down for change. Doing this step will simplify the model and we can try to find out if the readmission is related with medication changes.

4) num_med: Not only medication changes can be related with readmission, the total number of medications used can also be a key feature, since the number of the medicine reflected the severity of certain disease. And thus, we created a variable called num_med to store the total number of medications a patient used during the stay of hospital.

3.3.4 Recoding Existing Variables

1) Recode Diagnoses: The dataset includes three diagnostic variables (`diag_1`, `diag_2`, and `diag_3`) that were encoded using the ICD-9 system, which is the International Statistical Classification of Diseases and Related Health Problems. This system is designed to map health conditions to corresponding generic categories together with specific variations, assigning for these a designated code, up to six characters long This classification organizes diseases, symptoms, and external factors contributing to injury or illness into specific diagnostic codes, which can be up to six characters in length. First, we replaced the unknown value "?" into 1. We then recode the diagnoses into Circulatory-1, Respiratory-2, Digestive-3, Diabetes-4, Injury-5, Musculoskeletal-6, Genitourinary-7, Neoplasms-8, and Others-0. If ICD code is between 390 and 460, or it equals to 785, it belongs to category 1 (circulatory). If ICD code is between 460 and 520 or it equals to 786, it belongs to category 2 (respiratory). If ICD code is between 520 and 580 or it equals to 787, it belongs to category 3 (digestive). If ICD code equals to 250, it belongs to category 4 (diabetes). If ICD code is between 800 and 1000, it belongs to category 5 (injury). If ICD code is between 710 and 740, it belongs to category 6 (musculoskeletal). If ICD code is between 580 and 630 or it equals to 788, it belongs to category 7 (genitourinary). If ICD code is between 140 and 240, it belongs to category 8 (neoplasms). Others belong to category 0 (others). Appendix A shows the details of the

recoding process

2) Recode Age: To analyze the correlation between age and readmission rates, age categories were transformed into numerical values by assigning the midpoint of each age range. For example, the age range of 10-20 years was represented by the value of 15 years. This conversion facilitated a more quantitative examination of the impact of age on readmission rates.
3) Recode Readmission: The research concentrated on readmissions occurring within a 30-day period, recognized as a clinically relevant timeframe. Patients who experienced readmissions within this period were assigned a code of `1`, whereas those with no readmission or readmissions occurring after 30 days were coded as `0`. This binary categorization was consistent with the study's objectives

4) Recode Other Variables: For three variables which related with admission type, discharge disposition and admission source, we decided to encode the dummy variables for these categories. For variable "change", we recoded change into 1 and no change into 0. For gender, we recoded male into 1 and female into 0. For diabetes_Med, we recoded yes into 1 and no into 0. For race, we recoded the categorical variables into dummy variables: Caucasian-1, African American-2, Hispanic-3, Asian-4, and others-0. For A1Cresult, we recoded >7 and >8 into 1, Norm into 0, and None into 99. For max_glu_serum, we used the similar method, namely, we recoded >200 and >300 into 1, Norm into 0, and None into 99.

3.4 Feature Engineering

3.4.1 Data Type Conversion

For nominal features, we converted them into object type, for the later numerical variables processing.

3.4.2 Log Transformation, Standardization, and Correlation

The scatter plot of the distributions, as illustrated in Figure 1, reveals that most numerical features exhibit significant skewness and elevated kurtosis. According to the established criterion of skewness for normal distribution, a value exceeding +1 or falling below -1 indicates a highly skewed distribution. Conversely, if the skewness lies between -1 and -0.5 or between 0.5 and 1, the distribution is classified as moderately skewed. A skewness value within the range of -0.5 to 0.5 suggests that the distribution is approximately symmetric. Regarding kurtosis, a threshold of 3 is indicative of a normal distribution. To address these issues, we applied log transformation to the numerical variables, thereby facilitating their normalization to achieve a Gaussian-like distribution. Given that the numerical variables do not share a common scale, we subsequently employed standardization methods to rescale the data using the appropriate formula.:

$$New\ value = \frac{Value - Mean(Values)}{Standard\ Deviation\ (Values)}$$

After all data are standardized, we checked the correlation between the variables using a heat map to find top 15 correlated variables as Fig. 3 shows. There is not too much correlation between the variables and the correlation listed are self-explainable.

3.4.3 Outliers.

For detecting and processing the outliers, we used the coverage rule for normal distribution to deal with outliers. As Fig. 4 shows, the remaining 0.3% of the data are treated as outliers for this project. And thus, we removed the outliers.

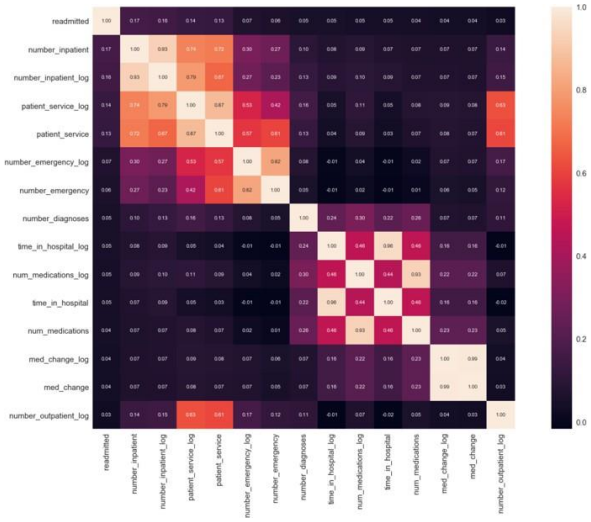

Figure 3: Heat map of top 15 correlated variables.

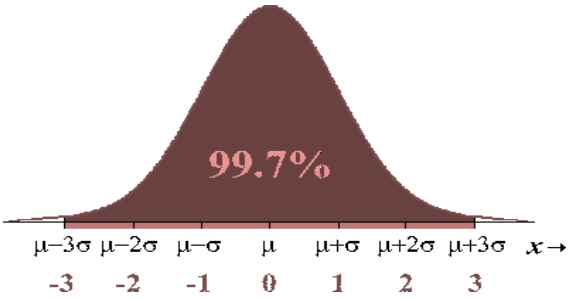

Figure 4: 99.7% of the observations fall within 3 standard deviations of the mean. [8]

3.4.4 Class Imbalance

Prior to the modeling phase, we conducted an analysis to assess the balance of class labels within the dataset. The findings revealed that there are 79,512 instances classified as class 0, which corresponds to patients with no need for admission or those experiencing readmissions beyond 30 days. In contrast, only 9,607 patients fall into the category of readmissions within 30 days. This results in a disproportionate ratio exceeding 8:1, with the proportion threshold set between 10-20%. Such an imbalance indicates that our dataset is significantly skewed, which may enhance accuracy in subsequent modeling efforts. To evaluate the

balance of class labels, we employed a confusion matrix, as illustrated in Figure 5. Our initial benchmark model, utilizing logistic regression, achieved an accuracy of 89%, although both precision and recall rates were recorded as zero. To address the imbalance, we implemented an over-sampling technique known as SMOTE, targeting the underrepresented class of readmissions. Figure 6 provides a visual representation of the mechanisms of over-sampling and under-sampling. Following the application of SMOTE, the dataset will consist of 79,512 patients in both category 0 and category 1. Additionally, Figure 5 presents the confusion matrix before and after the data balancing process.

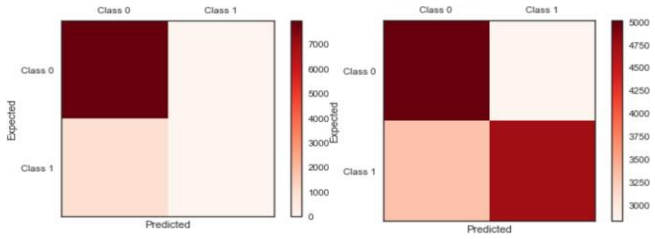

Figure 5: Confusion matrix before data balancing (left) and confusion matrix after data balancing (right).

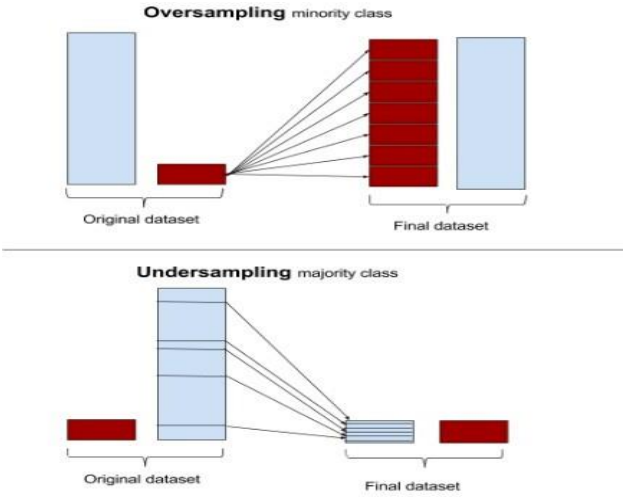

Figure 6: Explanation of over-sampling and under-sampling.

4. EXPERIMENT

Our objective in this modeling experiment is to identify the factors associated with high-risk diabetic patients. This is framed as a classification problem, specifically determining whether a patient will be readmitted within 30 days of discharge, after 30 days, or not at all. To address this, we employed various classification algorithms to ascertain the most effective method for achieving the highest accuracy. We selected and compared four distinct classification algorithms. Before training these algorithms, we divided our dataset into two separate subsets: the training set and the test set, comprising 90% and 10% of the data, respectively. The parameters for each algorithm were selected based on their classification performance, which was assessed using 10-fold cross-validation on the training set. The performance of all algorithms was subsequently evaluated on the test set. The methods we implemented include

4.1 Logistic Regression

Logistic regression is used as a benchmark model for our analysis. Since we assume that our data can be modeled as a log likelihood of outcome for the binary class label readmission, logistic regression can help us to understand the relative impact and significance of each attribute. We test this model by using 90% training and 10% testing data and 10-fold cross-validation. We achieved a cross-validation score: 61.29% and test set score 61.35%. By looking into the confusion matrix, we can calculate several measures of accuracy:

```
Accuracy is 0.61
Precision is 0.63
Recall is 0.59
AUC is 0.61
```

4.2 Decision Trees

Decision trees is a popular tree-based model that is easily to interpret the logic for splitting. Decision trees classify the data by sorting them down the tree from the root to some leaf node, with the leaf node providing the classification to the data. Due to the interactions between variables inherently, we removed the interaction variables from the feature set we did for logistic regression. Similarly, we did 10- fold cross validation score for decision trees too. The score equals to 88.97% and the dev set score is 89.43%, so decision trees look good for this dataset. After checking the score, we analyzed the confusion matrix for decision trees for both entropy and gini methods. As a result, both yielded the same results of measurements:

```
Accuracy is 0.89
Precision is 0.92
Recall is 0.87
AUC is 0.89
```

The result turned out that decision trees performed better than logistic regression based on its accuracy. The following graph showed the splitting process of the tree node. We visualized the trees in first two levels (Fig7).

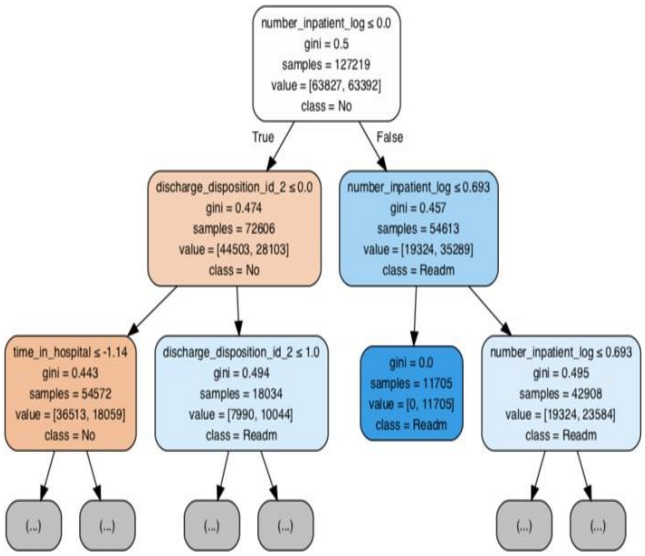

Figure 7: Decision trees for Gini index.

From the graph, it indicated inpatient visits is the first feature this decision tree used in deciding whether a patient will get readmitted.

4.3 Random Forests

Random Forest is composed of a set of decision trees. Each decision tree acts as a weak classifier and pooling the responses from multiple decision trees leads to a strong classifier. Each decision tree is trained independently and determines the class of an input by evaluating a series of greedily learned binary questions. The random forest consisting of 10 trees, with the max_depth of as 25 nodes was used, as it was found to be optimal from the experiment with varying number of trees and depth in the forest. After implementing Random Forest, we achieved similar results of measurements for using gini and entropy methods. Random Forest showed better results than decision tree as regards to prediction accuracy.

```
Accuracy is 0.92
Precision is 0.98
Recall is 0.87
AUC is 0.92
```

4.4 Model Improvement

Following the execution of the random forest algorithm, we opted to enhance our model by employing a boosting technique utilizing the relatively novel algorithm XGBoost. Boosting serves as an ensemble approach that constructs a robust classifier from a series of weaker classifiers, contingent upon the degree of correlation between the learners and the actual target variable. Each subsequent predictor rectifies the errors made by the preceding model, iteratively stacking models until the training data is accurately predicted or a predetermined maximum number of models is reached.

EXtreme Gradient Boosting (XGBoost) is an ensemble machine learning technique that has gained significant traction since its inception in 2014. It represents a scalable and precise implementation of gradient boosting machines, demonstrating remarkable capabilities in maximizing computational efficiency for boosted tree algorithms. XGBoost is designed specifically to enhance model performance and computational speed, accommodating a variety of generic loss functions, and offering a range of customizable parameters.

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

Training loss Complexity of the Trees

We applied and tuned the algorithm for better performance. We tuned the following three parameters

1. eta: learning rate to prevents overfitting (eta=0.01, 0.02,0.05).

2. max_depth: the max depth of the tree (max_depth=3,4,5,6,7,8,9).

3. cols_sample: the percentage of features can be chosen (cols_sample=0.6,0.7,0.8,0.9,1.0).

We tuned the three parameters one by one and iterate the values to find the least test error and highest accuracy. The best iteration we found is with accuracy 0.94, precision 1.0, recall 0.88 and AUC is 0.94.

4.5 Evaluation

In this section, we will discuss the evaluation of classifier performance and answer the second question of identifying the most import factors.

4.5.1 Classifier Comparison

Each algorithm underwent evaluation through a 10-fold stratified cross-validation process. This technique involves partitioning the dataset into several folds in a random yet balanced manner. Stratified cross-validation specifically aims to maintain the class distribution across the folds, ensuring that each fold accurately reflects the overall dataset. In this method, the learning algorithm is trained on nine of the folds while being tested on the remaining fold. By repeating this cross-validation procedure, we mitigate the risk of bias introduced by any random initialization, thereby enhancing the reliability of the results.

The performance of all algorithms is assessed using the area under the curve (AUC), which corresponds to the c-statistic in the context of binary classification. The AUC-ROC curve serves as a performance metric for classification tasks across various threshold settings. The ROC curve represents a probability curve, while the AUC quantifies the model's ability to differentiate between classes. Specifically, it indicates the likelihood that a positive instance, defined as "<30" coded as 1, is ranked higher than a negative instance coded as 0. A higher AUC value signifies a superior model performance in accurately predicting 0s as 0s and 1s as 1s. Prior studies in the domain of readmission have reported AUC values ranging from 0.5 to 0.7.

In the evaluation of four predictive models, Table 2 indicates that XGBoost outperforms the others in forecasting the admission rate, attaining the highest accuracy of 0.94 and an AUC of 0.61. The random forest model follows as the second most effective, achieving an accuracy of 0.92 and an AUC of 0.94. Additionally, Figure 8 illustrates the comparative performance of the models overall.

Classifier	Accuracy	Precision	Recall	AUC
Logistic_ Regression	0.61	0.63	0.59	0.61
Decision Tree	0.89	0.92	0.87	0.89
Random Forest	0.92	0.97	0.87	0.92
XGBoost	0.94	1.0	0.88	0.94

Table2: Comparison between different algorithms

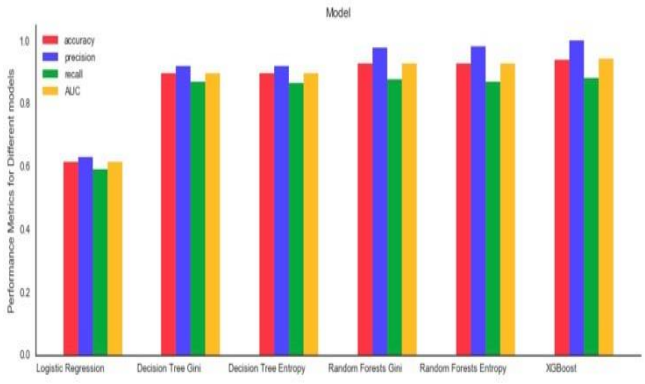

Figure 8: Comparison between models.

4.5.2 Most Important Predictors

For the second question what the strong predictors are contributing to predicting readmission, different algorithms provided different results. Specifically, Fig. 9 illustrated showed the most important variables after the classification for decision tree. We plotted those features whose importance is bigger than 0.01. The most important variables are number_inpatient and time_in hospital, and discharge_disposition_id_2, number_procedures and num_medications are among the top 5 strongest predictors.

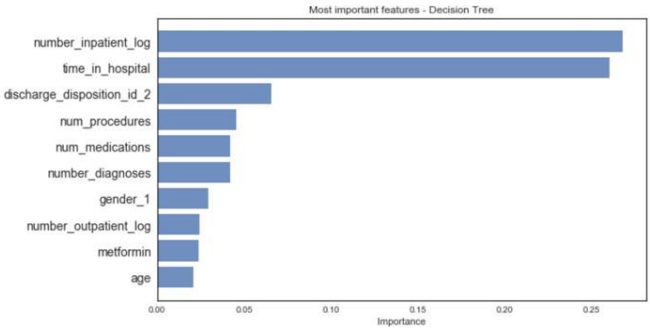

Figure 9: Most important features for decision tree model.

Fig. 10 showed the important features for random forests, which are different from the decision trees, with number_inpatient, time_in _hospital, number_diagnosis, discharge_id_2 and metformin are among the top 5 important predictors.

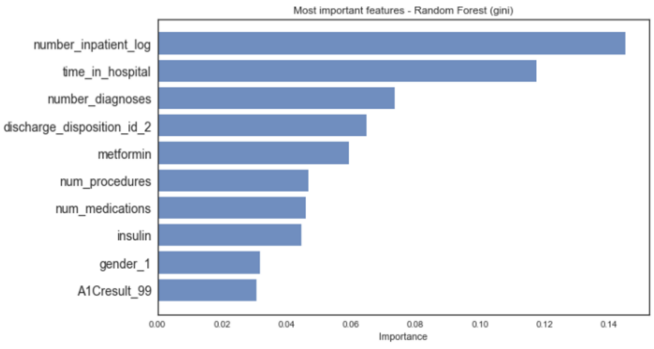

Figure 10: Most important features for random forests.

Fig. 11 indicated the important features for XGBoost which are slightly different than previous with number_medications, time_in_hospital, age, number_procedures, num_diagnosis are among the top 5 important predictors. The results are quite interesting.

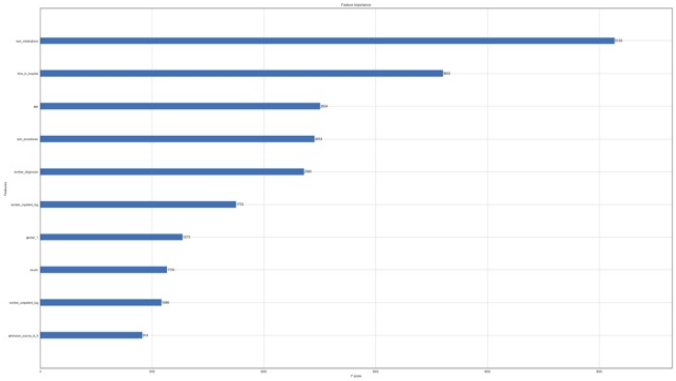

Figure 11: Most important features for XGBoost.

5. CONCLUSIONS

In this work we adopted machine learning methods to identify high risk patients and evaluated different machine learning algorithms. Compared to the previous analysis, our study achieved high accuracy due to the sophisticated pre-processing procedure. The XGBoost method is reported to be the best method for prediction of the readmission rate for diabetes patients.

We identified the most important factors as the time_in_hospital and number of inpatients, number of diagnoses, which appears to associate with the severity of the disease. Further studies could conduct more exploration when analyzing these factors individually.

REFERENCES

[1] Benbassat, J. Taragin, M. 2000. Hospital readmissions as a measure of quality of health care advantages and limitations. *Arch Intern Med.* 160(8):1074–1081.

[2] Leppin, A.L., Gionfriddo, M.R., Kessler, M., Brito, J.P., Mair, F.S., Gallacher,K., Wang, Z., Erwin, P.J.,Sylvester, T., Boehmer, K. and Ting, H.H., 2014.Preventing 30-day hospital readmissions: a systematic review and meta-analysis of randomized trials. *JAMA internal medicine,* 174 (7), 1095-1107. Hines, A.L., Barrett, M.L., Jiang, H.J. and Steiner, C.A., 2006. Conditions with the largest number of adult hospital readmissions by payer, *Statistical Brief.* 172 (2011).

[3] Salerno, A.M., Horwitz, L.I., Kwon, J.Y., Herrin, J., Grady, J.N., Lin, Z., Ross,J.S. and Bernheim, S.M., 2017. Trends in readmission rates for safety net hospitals and non-safety net hospitals in the era of the US Hospital Readmission Reduction Program: a retrospective time series analysis using Medicare administrative claims data from 2008 to 2015. *BMJ open,* 7(7) Dungan, K. M. The effect of diabetes on hospital readmissions.,2012. *Journal of diabetes science and technology,* 6(5), 1045–1052.

[4] Eby, E., Hardwick, C., Yu, M., Gelwicks, S., Deschamps, K., Xie, J. and George, T., 2015. Predictors of 30-day hospital readmission in patients with type 2 diabetes: a retrospective, case–control, database study. *Current medical research and opinion*, 31(1), 107-114.

[5] Howell, S., Coory, M., Martin, J. and Duckett, S., 2009. Using routine inpatient data to identify patients at risk of hospital readmission. *BMC Health Services Research*, 9(1), 96.

[6] Jiang, H.J., Stroyer, D., Friedman, B. and Andrews, R., 2003. Multiple hospitalizations for patients with diabetes. *Diabetes Care,* 26(5), 1421-1426.

[7] Hosseinzadeh, A., Izadi, M.T., Verma, A., Precup, D., and Buckeridge, D.L., 2013. Assessing the predictability of hospital readmission using machine learning. IAAI.

[8] Strack, B., DeShazo, J.P., Gennings, C., Olmo, J.L., Ventura, S., CIS, K.J. and Clore, J.N., 2014. Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed Research International*.

[9] Bhuvan, M.S., Kumar, A., Zafar, A. and Kishore, V., 2016. Identifying diabetic patients with high risk of readmission. *arid preprint* arXiv:1602.04257.

[10] https://archive.ics.uci.edu/ml/datasets/diabetes+130_us+hospitals+for+years+1999-2008.

[11] https://freedium.cfd/how-to-use-machine-learning-to-predict- hospital-readmissions-part-2-616a0c920

[12] Damian M. Predicting Diabetic Readmission Rates: Moving Beyond Hba1c. Curr Trends Biomedical Eng & Biosci. 2017; 7(3): 555707. DOI: 10.19080/CTBEB.2017.07.555715.