Predicting 30-Day Hospital Readmission for

Diabetic Patients Using Machine Learning Methods

Maggie Tang

Summer 2020

**Abstract**

Hospital readmission is defined as a term to explain situations within which patients return to the hospital within 30 days under unplanned of their initial hospitalization, it includes patients readmitted to identical hospital or other hospital for any reasons [[1].](https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/HRRP/Hospital-Readmission-Reduction-Program) This paper aims to see whether machine learning models is informative at predicting diabetic patient hospital readmission. Using 10 years Health Facts data together with several data processing techniques, patients in this dataset were classified into two different groups (YES = readmitted within 30 days, NO = Not readmitted or readmitted after 30 days). Out of total 101766 records, 11194 (11 %) encounters were found as readmitted. A comprehensive methodology with data pre-processing, feature selection, and under size sampling are some methods used to deal with noisy and imbalance data. Predictive models employed in this study include Logistic Regression (LR), Random Forest (RF), Multilayer Perceptron (MLP), and Adaboost. Metrics including accuracy, precision, sensitivity, specificity and AUC were conducted to live the model performance. The comparative analyses result showed that logistic regression achieved the best results with an overall accuracy of 64.7%, while imbalance dataset didn't work in this study. This study found that risk factors like admission type, length of stay, medication change, metformin, number of inpatients are strong predictors of readmission.

**1.Introduction**

Diabetes is a wild spread chronic disease, it is the seventh leading cause of death in the United States, and also one of the most critical healthcare problems today. Diabetes happens when your blood glucose, also called blood sugar, is too high because either the pancreas is not any longer able to make insulin or the body can't make use of insulin as good as it should be. Diabetes has quite customary syndrome due to different age groups and differing types of diabetes. Generally, diabetes is divided into three different forms: type 1, type 2 and Gestational diabetes. Type 1 diabetes occurs at any age but most frequently in children, body produces little or no insulin with type 1 diabetes, then daily insulin injections to maintain blood glucose levels under control is a must; 90% diabetes cases are fall in type 2, it occurs more common in adults, body can not observe good use of the insulin that it produces with type 2 diabetes, treatment for type 2 diabetes is healthy lifestyle and oral drugs, to keep blood glucose levels under control; Gestational diabetes (GDM) occurs during pregnancy and is associated to both mother and child, GDM usually disappears after pregnancy but women affected and their children are at increased risk of developing type 2 diabetes later in life [[2]](https://www.cdc.gov/diabetes/basics/diabetes.html#:~:text=Diabetes%20is%20the%20seventh%20leading,diabetes%20has%20more%20than%20doubled.). Common diabetes complications including heart condition risk increasing, foot ulcers risk increasing, blindness, and kidney failure, it can also cause heart attack and stroke. Overall death risk for people with diabetes is a minimum of double than those without this disease. Report from National Diabetes Statistics Report, 34.2 million people, or 10.5% of the U.S. population have diabetes at 2020, with an annual increase new patient approximately 1.5 million. If the current trend continues, one in three Americans are likely to have diabetes by the year 2050. Diabetes cost the U.S. $327 billion in 2017, that indicates every 1 in 7 health care dollars is spent on treating diabetes and its complications [[3]](https://www.diabetesresearch.org/file/national-diabetes-statistics-report-2020.pdf).

Although diabetes has no cure, special care for diabetic patients can make a great improvement on their survival rate. There’re some steps diabetes patients can take and stay healthy, like take diabetes medicine; make a diabetes meal plan with help from your doctor, avoiding sugar and tobacco, achieve and maintain a healthy body weight; stay physically active; keep track your blood sugar every day.

Hospital readmission is a real-world problem. According to report from the Center for Medicare &Medicaid Service (CMS), there's 1 in 5 readmission rates in the United States. Hospital readmissions cost Medicare $26 billion annually, with about $17 billion are avoidable [[4]](https://www.cms.gov/about-cms/agency-information/omh/downloads/omh_readmissions_guide.pdf). Readmission rate is taken into account as a high-quality measure of a hospital's performance. CMS has created many programs to improve health care providers' quality, hospital Readmission Reduction Program (HRRP) is one of these programs. Beginning on October 1, 2012, HRRP seeks to improve healthcare for Americans by linking payment to the quality of hospital care, federal government cuts payments to hospitals that have high rates of readmission  [[1]](https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/HRRP/Hospital-Readmission-Reduction-Program). Records from 3367 U.S. hospitals shows 71% hospitals were under penalized, with an average penalty rate of 0.28%. For those hospitals, necessary steps have to take to cut back the readmission rate. There're several strategies were proven to scale back hospital readmission, like identify high risk patients, ensure appropriate nursing coverage, improve transitional care, ensure patients understand their post-discharge care instructions, and schedule 7-day follow-up appointments. In recent years, rapid development in computer technologies created a wide adoption with electric medical data system, more and more medical related data and claims are available. Many computational models are developed to create prediction in medical field, they help healthcare to spot high risk readmission patients, identify the risk factors, and provide reasonably additional assistance for healthcare to reduce the readmission rate.

Remainder of this paper is structured as follows. In Section 2 we will review literatures which have been used for hospital readmission prediction. Section 3 details the methodology for this study, covers data pre-processing, feature selection, rebalance data, and modeling. Section 4 discuss the results. Section 5 conclude this study, and Section 6 describes limitation and recommendations for future research.

2**. Literature review**

Machine learning is becoming more and more popular now days in medical classification research. A review by Bojja et al. [[5]](https://scholar.dsu.edu/cgi/viewcontent.cgi?article=1016&context=research-symposium) evaluated several machine learning models to predict high-risk readmit patients. They found Some key features effect readmission are number of inpatients, time in hospital, number of medications and number of diagnosis.

Long et al. focused on using qualitative methods to elucidate readmission risk factors [[6]](https://bmjopen.bmj.com/content/3/9/e003212.long), they found patients went on to emergency department after they experienced health issue rather than contact their primary physician. That showed formal transitional care services are unlikely to be adequate in reducing readmission rate among high-risk patients.

Some published papers worked on predicting hospital readmission rates focused on the impact of A1c test result (A1C percentage positively corresponds to blood glucose levels. Higher A1C level associate to higher risk of developing diabetes, or complications of diabetes. Normal A1C level is below 5.7 percent for somebody who doesn't have diabetes, level of 6.5 percent or higher on two separate occasions shows that you have diabetes. If A1C level equal or less than 7 percent, patients will get common treatments, while doctors will recommend change treatment plan if A1C level is higher than 7 percent.) In related research [[7]](https://www.hindawi.com/journals/bmri/2014/781670/), Strack et al. combined A1c (A blood test for type 2 diabetes and prediabetes, which is also widely applied to measure diabetes care performance) with variable "change of medication", made a replacement named "HbA1c" with four levels including "no test was performed", "result was high and the diabetic medication was changed", "result was high but the diabetic medication was not changed", and "normal results of the test". They found the results of readmission looked as if it would be related to the choice to check the HbA1c, instead of the values of the HbA1c result, they strongly suggested more attention should be paid to diabetes patients during their hospitalization, that may bring greater impact on readmission.

Pujianto et.al also acknowledged the importance of HbA1c [[8]](https://www.researchgate.net/publication/338128718_Comparison_of_Naive_Bayes_Algorithm_and_Decision_Tree_C45_for_Hospital_Readmission_Diabetes_Patients_using_HbA1c_Measurement), they removed 83% of observations from the initial dataset which has a A1C value of "None", indicates patients didn't take the A1c test. The most effective performance from this study was combined SMOTE with decision tree C4.5, with an accuracy value of 82.74%, precision value of 87.1%, and recall value of 82.7% at predicting diabetic patient’s readmission.

Chakraborty et al. applied high performance support vector machine algorithm to predict the probability of a diabetic patient being readmitted [[9]](https://www.researchgate.net/publication/279530614_Predicting_Readmission_of_Diabetic_Patients_using_the_high_performance_Support_Vector_Machine_algorithm_of_SASR_Enterprise_Miner), helped the hospitals design follow-ups. This study split 60% data into training set and 40% into testing set, several feature selection methods were used, like LARS, LASSO, Stepwise Regression and Forward Regression. Among numerous machine learning models including Decision Tree, Logistic Regression, MBR, SVM and others, support vector machine performed the most effective, with the best accuracy of 63.7%, sensitivity of 49.7%, and a specificity of 75.1%. This study also identified key factors for readmission are number of inpatient and number of outpatients, primary diagnosis, admission mode.

Study carried by Glasgow et al. [[10]](https://pediatrics.aappublications.org/content/88/1/98), focused on readmission for children with diabetes, they found readmission rate increased due to the missed insulin injections.

Hammoudeh et al [[11]](https://www.researchgate.net/publication/328887677_Predicting_Hospital_Readmission_among_Diabetics_using_Deep_Learning), used deep learning algorithm tried to resolve the problem that most current practices to spot at-risk diabetic patients readmission are too subjective, and just a very slight better than random guessing. They combined Convolutional neural networks and data engineering, found this mix performed better than other machine learning algorithms at predicting real life data.

Goudjerkan et al. [[12]](https://thesai.org/Downloads/Volume10No2/Paper_36-Predicting_30_Day_Hospital_Readmission_for_Diabetes_Patients.pdf) worked on the diabetes dataset with Multilayer Perceptron. Random Forest was performed for feature selection in their study, SMOTE algorithm was found unravel to solve the target variable unbalanced problem. The proposed MLP model was applied with 80:20 train-test ratio, it achieved a high score (above 90%) on all evaluated metrics: accuracy, recall, precision, and AUC.

Another point of view is presented by Hosseinzadeh et al. [[13],](https://www.academia.edu/4232658/Assessing_the_Predictability_of_Hospital_Readmission_Using_Machine_Learning) they used machine learning to assess the predictability of hospital readmission. Results from this study suggested that prescription medications, diagnostic information and knowledge on procedures during hospital admission can successfully predict hospital readmission.

3. **Methodology**

In this study, we followed a data mining methodology called cross-industry standard process (CRISP-DM) for data processing. It's an [open standard](https://en.wikipedia.org/wiki/Open_standard) process model that describes common approaches employed by data processing experts [[14]](https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining#:~:text=Cross%2Dindustry%20standard%20process%20for%20data%20mining%2C%20known%20as%20CRISP,most%20widely%2Dused%20analytics%20model.). CRISP-DM breaks the tactic of knowledge into six major phases (show in Fig. 1.): (1) Business Understanding: define the target and clarify the goal of this study, (2) Data Understanding: take a close examine at available data, explore the information with tables and graphics, verify data quality (3) Data Preparation: cleaning and remodeling relevant data, it's the foremost important part of data processing, (4) Modeling: develop models using comparable analytical techniques, (5) Evaluation: evaluate the performance of models to estimate the generalization accuracy of a model in future data, (6) Deployment: deploy the models to be employed in decision-making processes. CRISP-DM provides strong guidance for analytics, data processing and data science projects, it is the most widely used data processing methodology. Roughly 80% of the complete project time was spent on the first three steps, the attention paid to the earlier steps increased the model performance and result reliability. Despite its popularity, CRISP-DM has not been revised since its creation at 1996 [[15]](https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html).

To estimate the performance of the estimator models, a 10-fold cross-validation was used in this study. K-fold Cross-validation is a resampling procedure used to evaluate performance of predictive models. It splits the initial data into a training set and a test set, then use training set to train the model, and test set to test it. The advantage of this method is that every observation is used for both training and validation, it helps to utilize data better, and get more information from algorithm performance. Previous research carried by Suliman & Chang [[16]](https://books.google.com/books?id=u3NgDwAAQBAJ&pg=PT44&lpg=PT44&dq=.+Previous+research+showed+that+10+seem+to+be+the+optimal+number+for+folds+(the+optimizes+the+time+it+takes+to+complete+while+minimizing+the+bias+and+variance+associated+with+the+validation+process&source=bl&ots=qW5H60_DTr&sig=ACfU3U2bVWwvHmiyJI7quPH-aDkavydMrw&hl=en&sa=X&ved=2ahUKEwiT6834lu7pAhVSGs0KHZdUBGwQ6AEwAHoECAkQAQ#v=onepage&q&f=false) revealed that 10 seems to be the optimal number for folds (the optimizes the time it takes to complete while minimizing the bias and variance associated related to the validation process)

Fig. 1. CRISP-DM

A close up of text on a white background

Description automatically generated

3. 1. **Data description**

Dataset used for this study contains clinical care from 130 America hospitals, from 1999 to 2008, with 101,766 observations and 50 features representing patient and hospital outcomes [[17]](https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008).

Following information was extracted from the database:

(1) It is a hospital admission inpatient encounter.

(2) It is a diabetic encounter.

(3) The length of days stay in hospital was in the range (1,14).

(4) Laboratory tests were performed during the encounter.

(5) Medications were administered during the encounter.

There're 50 features describe diabetes encounters in this database, including diagnoses, diabetic medications, number of hospital visits, demographic. The full list of features and their description are show in Table 1.

Table 1 Features and descriptions

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Feature Type** | **Description and values** | **% missing** |
| Encounter ID | Numeric | Unique identifier of an encounter | 0% |
| Patient number | Numeric | ID for patients | 0% |
| Race | Nominal | Caucasian, Asian, African American, Hispanic, and other | 2.23% |
| Gender | Nominal | Male, female, and unknown/invalid | 0% |
| Age | Nominal | 10-year intervals in 10 distinct groups, from age 0 -100. | 0% |
| Weight | Numeric | Weight in pounds | 96.9% |
| Admission type id | Nominal Integer | 8 distinct inpatient admission type code  corresponding to 8 admission types, like newborn, emergency, elective, urgent. | 0% |
| Discharge disposition id | Nominal Integer | 29 distinct discharge disposition code corresponding to 29 discharge disposition types. | 0% |
| Admission source id | Nominal Integer | 21 distinct admission source code corresponding to 21 admission source types. | 0% |
| Time in hospital | Numeric | Days between admission and discharge, at least 1 day, at most 14 days. | 0% |
| Payer code | Nominal Integer | Identifier corresponding to distinct payment method. | 39.6% |
| Medical specialty | Nominal Integer | 84 distinct values for admitting physician specialty. | 49.1% |
| Number of lab procedures | Numeric | 0-132 | 0% |
| Number of procedures | Numeric | 1-6 | 0% |
| Number of medications | Numeric | 1-81 | 0% |
| Number of outpatient visits | Numeric | 0-42 | 0% |
| Number of emergency visits | Numeric | 0-76 | 0% |
| Number of inpatient visits | Numeric | 0-21 | 0% |
| Diagnosis 1 | Nominal | The primary diagnosis, ICD9 coded, 848 distinct values | 0.02% |
| Diagnosis 2 | Nominal | Secondary diagnosis, ICD9 coded, 923 distinct values | 0.35% |
| Diagnosis 3 | Nominal | Additional secondary diagnosis, ICD9 coded, 954 distinct values | 1.4% |
| Number of diagnoses | Numeric | Number of diagnoses foe each patient. | 0% |
| Glucose serum test | Nominal | 4 distinct values indicate the range of glucose serum test result, “>200”, “>300”, “normal”, and “none” for no test was taken. | 0% |
| A1c test result | Nominal | 4 distinct values indicate the blood glucose range of A1c result: “>8”, “>7” “normal”, and “none” indicates no test was taken | 0% |
| Change of medications | Nominal | Indicates if there has a change in diabetic medications. 2 levels: “change”, “no change” | 0% |
| Diabetes medications | Nominal | Indicates if patients changed medications. 2 levels: “yes”, “no” | 0% |
| 24 features for medications | Nominal | “up”, “down”, “steady”, “no”. | 0% |
| Readmitted | Nominal | Dependent variable.  ">30", "<30", "No". | 0% |

Readmitted was the dependent variable, it was classified into two categories: "Yes", if the patient was readmitted within 30 days of discharge from hospital; "No", if a patient readmitted after 30 days or no readmission at all. McIlvennan et al. [[18]](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4439931/) mentioned that the time frame was set at 30 days because readmission during this time can be influenced by the quality of care received at the hospital and how well discharges were coordinated. Funding agencies also use 30 days to determine penalty to hospital. Around 89% variables in the original data fall in "Not readmitted" class, thus the readmitted rate is 11% (show in Fig. 2.). This dataset is highly skewed, which is commonly encountered with medical data, where majority of data samples contribute to one class, and only very small minority samples contribute to another class. In general, unbalanced dataset always causes difficulties in machine learning algorithms, it will push the model to the more common class and produce bias. To avoid this, this study worked both on original unbalanced data and well-balanced data, we'll compare the results to see which one can give us better prediction.

Fig. 2. Output Class Distribution

As real-world data, especially medical data are always noisy, data preparation is directly related to the model performance. Data cleaning process in this study includes detecting and repairing missing values and noises. Consequently, the data will be replaced, modified and deleted.

Missing values: There were several features cannot provide correct information since they had a high missing percentage of missing values. These features were weight (96.9% missing values), medical specialty (49.1% missing values), payer code (39.6% missing values). Varying guidance exists at answering what proportion of missing data should be allowed in the use of machine learning models. Statistical guidance articles have stated that bias is likely to produce in analyses with missingness more than 40%, it should be considered excluded from the modeling process [link](https://www.sciencedirect.com/science/article/pii/S0895435618308710). We removed weight, medical specialty and because of the high percentage missing values. Payer code was at the edge of that range, since it just indicated the payment types, it was not relevant to the outcome, we removed it from the dataset. Observations with missing values in columns like race, diag\_1 were removed.

ID variable: Patient\_nbr is a unique identifier of a patient, it has 30248 duplicated rows, some previous researches dropped these duplicates [[7].](https://www.hindawi.com/journals/bmri/2014/781670/) But when we took a deep look inside patient\_nbr, all duplicates received different treatments, some had different outcomes, it's reasonable to consider them as unique samples, we deleted the column but kept all duplicates.

Another irrelevant id variable is encounter\_id, we removed it too.

Nearly zero variance variable: Number\_outpatient and number\_emergency have nearly zero variance, we removed both variables; There’re 25 variables indicated lab test and medications from max\_glu\_serum to metformin pioglitazone, 23 have nearly zero variance, we removed those, kept insulin and metformin.

3 invalid gender values were eliminated.

There're 3 diagnoses records in the original dataset, the primary diagnosis (diag\_1), the secondary diagnose (diag\_2), and the additional secondary diagnose (diag\_3), each has 700 levels. We just focused on the primary diagnosis and removed diag\_2 and diag\_3.

Feature encoding: After having cleaned the data from missing values and other potential bias, it is important to optimize the feature and, mostly in this case, reduce the number of unique values for categorical variables. This study encoded most of the variables into string format. Race has 6 distinct values, but 75% were fall in Caucasian, we kept this category and set "Others" to the rest categories. Admission\_type\_id, discharge\_disposition\_id, and admission\_source\_id each has values skewed to one category with more than 50% proportion. In order to reduce the complexities to a manageable level, based on hospital coding report ([[19]](https://www.resdac.org/cms-data/variables/claim-inpatient-admission-type-code-ffs) [[20]](https://www.cms.gov/Medicare/Medicare-Contracting/ContractorLearningResources/Downloads/JA0801.pdf) [[21]](https://www.resdac.org/cms-data/variables/claim-source-inpatient-admission-code-ffs)), we reduced each variable to two levels.

Num\_procedures column indicates number of procedures (other than lab test) performed during the encounter, it is highly skewed with 46% value fall into 0 category. We assigned binary values to it, 0 stands for patients didn't get any non-lab procedures, 1 stand for patients who got those procedures. Number inpatient was highly skewed to the 0 category, we encoded two values to num\_inpatient, 0 for no inpatient visit, 1 for had inpatient visit. We converted values in number\_diagnoses to three categorical classes: Less than 9; 9; More than 9.

Age was encoded as nominal in 10-year intervals with 10 distinct groups, we encoded age as discrete with two levels, 0-50 and 50-100, while other researchers decided to keep it as numerical type, for example, age 50-60 will be converted to 55.

Diag\_1 was clustered it down to 5 categories based on the ICD9 codes [[22]](https://en.wikipedia.org/wiki/List_of_ICD-9_codes).

After data cleaning process, there's 94945 with 17 features in the dataset, show as Table 2.

Table 2 Variables after data preparation

|  |  |  |
| --- | --- | --- |
| **Name** | **Feature Type** | **Description and values** |
| Race | Nominal | Caucasian; Others |
| Gender | Nominal | Male, Female |
| Age | Nominal | over 50; Under 50 |
| Admission type id | Nominal | Emergency; Urgent; Elective; Reserved; Others |
| Discharge disposition id | Nominal | Discharge to Home; Others |
| Admission source id | Nominal | From Emergency Room; From refer; Others |
| Time in hospital | Numeric | 1-14 |
| Number of lab procedures | Numeric | 1-93 |
| Number of procedures | Nominal | 0; 1 |
| Number of medications | Numeric | 1-35 |
| Number of inpatient visits | Nominal | 0; 1 |
| Diagnosis 1 | Nominal | Diabetes; Circulatory; Digestive; Injury; Others |
| Number of diagnoses | Nominal | Less than 9; 9; More than 9 |
| Change of medications | Nominal | change; no change |
| Diabetes medications | Nominal | yes; no |
| Readmitted | Nominal | yes; no |
| Metformin | Nominal | Up; Steady; No; Down |
| Insulin | Nominal | Up; Steady; No; Down |

3.2 **Feature selection**

Feature selection is one of the core concepts in machine learning, it provides insights into several novel factors that may help to delineate readmission rates that associated with diabetes. Feature selection process will select the most effective subset of attributes that are most important and have high contribution at the time of prediction making. We then use the reduced subset to suit the model. It helps to make models simpler and easier to interpret, shorter training time, and reduce overfitting, it also encompasses a huge influence on the model performance. In general, feature selection fall under four categories: filter methods, wrapped methods, embedded methods and heuristics. The filter method acquiring no feedback from the classifier, it can exclude a number of the features before the training stage. The thought is to get rid of uninformative and redundant features with the help of statistical tests like Chi-Square, ANOVA, and Pearson’s Correlation. The wrapper methods measure the “usefulness” of features supported the classifier performance. General Process of wrapper methods including seek for a subset of features, build a machine learning model, evaluate model performance, and repeat. Embedded method is a combination of filter methods and wrapper methods. Feature selection employed in this study is Forward Feature Selection, one of the wrapper methods. This method starts at no feature and add one at one occasion, evaluate the model performance, stop when reach the optimal performance for the model with a specific subset of feature.

Machine learning model built for feature selection during this study is Naive Bayes. It is a classification technique supported Bayes’ Theorem, which has an assumption that all independent variables are independent of each other in a dataset. Naive Bayes model is straightforward to build and very useful when working with large datasets. Table 3 shows the optimal subset of feature.

Table 3 Features selected for final model

|  |  |  |
| --- | --- | --- |
| **Name** | **Feature Type** | **Description and values** |
| Admission type id | Nominal | Emergency; Urgent; Elective; Reserved; Others |
| Time in hospital | Numeric | 1-14 |
| Number of inpatient visits | Nominal | 0; 1 |
| Change of medications | Nominal | change; no change |
| Metformin | Nominal | Up; Steady; No; Down |

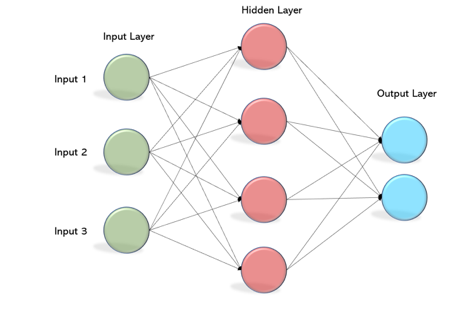
3. 3. **Predictive models**

Identify diabetes patients with high risk to readmit in 3o days is a classification problem. P

art of this study is to make use of several machine learning algorithms. Four popular classification methods, artificial neural network (ANN), logistic regression (LR), random forest (RF), and Adaboost, along with forward feature selection were built and compared each model's performance on the sample dataset. Some studies ([[23]](https://pubmed.ncbi.nlm.nih.gov/15894176/) [[24]](https://personal.utdallas.edu/~ryoung/phdseminar/DataMiningComparison-Melody.pdf) [[25]](https://www.wku.edu/instres/documents/comparison_of_empirical_models.pdf) [[26])](https://www.sciencedirect.com/science/article/pii/S0167923610001041) worked on comparing data mining methods in different settings found machine learning methods (i.e., artificial neural network, random forest, support vector machines, naive bayes) performed better than statistics methods (i.e., logistic regression). However, results got from this study didn't prove it.

Artificial neural network is an interconnected group of nodes, inspired by a simplification of neurons in a brain [link](https://en.wikipedia.org/wiki/Artificial_neural_network#:~:text=An%20artificial%20neural%20network%20is,to%20the%20input%20of%20another.). It is capable of modeling very complex functions, and very powerful with massive datasets. ANN works well with both labeled data (Supervised learning) and unlabeled data (Unsupervised learning). Now days, ANN is used for variety tasks, especially popular for classification. ANN is used in Self Driving cars, Facial Recognition, Text Classification, Stock Market Prediction, and lots of other interesting applications. The best thing for ANN is you don't have to train your model, it will get trained by itself, like a human brain. Multi-layer perceptron (MLP) is one of the most popular neural network architectures (show in Fig.3.).

Fig. 3. MLP type artificial neural network architecture used in this study



Random forest is an ensemble learning method which is flexible and simple to use. Random forest is one of the most popular machine learning algorithms due to its simplicity and diversity (works both for classification problems and regression problems). It builds many trees and adds additional randomness to the model while the tree number growing, then marge them to induce a better performance, it is usually trained with the “bagging” method. One disadvantage of decision trees is that they have an inclination to overfit by memorizing the training data. As a result, random forests were created cut back the overfitting.

Logistic regression is generally used for classification purposes by assigning observations to a discrete set of classes. This statistical model assumes that there is a linear separability between the points and makes prediction based on the concept of probability (show as Fig. 4.). Logistic Regression performs well when the dataset is linearly separable, it not only gives a measure of how relevant a predictor (coefficient size) is, but also its direction of association (positive or negative), it is easier to implement, interpret and very efficient to train. Logistic regression is widely used in a lot of fields, for example: whether an email is a spam or not; whether the tumor is malignant or not; whether an online transformation is fraud or not; whether smoke everyday will increase the probability of getting lung cancer.

Fig. 4. Logistic regression

A screenshot of a cell phone

Description automatically generated

Adaboost constructs a powerful classifier by sequentially combining a collection of weak classifiers. A single classifier is learnt to minimize the classification error at the first iteration. At each consequent iteration, a brand-new classifier is learnt which seeks to reduce the error of the classifier composed of the set of classifiers learnt until the previous iteration.

3.4. **Imbalance data**

The quality of the predictive model performance is challenged with class imbalance issue. Imbalance data is one of the obstacles for many machine learning algorithms, it is a common problem in machine learning classification. machine learning algorithms used for classification are designed with the assumption that each class has equal numbers, imbalance data will push the machine learning model to the majority class and hard to detect the minority class. As a result of imbalance data, the classifier performance on the minority may be insufficient when compared to the majority. Imbalance data widely exit in medical diagnosis, spam filtering, and fraud detection. The imbalance might vary in different dataset, some are slightly, some are severe. Severe imbalance is more challenging to model and may require specialized techniques. Sampling stratifies are often used to overcome the class imbalance problem. There are two main methods to deal with imbalanced data: Under-sampling methods, which balances the classes by randomly selecting examples from the majority class until the remaining number of examples is roughly the same as in the minority class; Over-sampling methods, which balance the classes by increasing the number of the minority class. Sampling stratify used in this studied is under size sampling, it removes rows from the majority class so the values in a categorical column are equally distributed.

Since dependent variable "readmitted" in this dataset is highly skewed with 88.8% value in "Yes" class and rest in "No" class, we experimented with the choices using and comparing the results of the models with original data (biased data) vs the well-balanced data.

Table 4 shows the confusion matrix, and table 5 shows the accuracy from all four classification models employed in this study with imbalance dataset. we found models made with imbalance dataset just detected the patients who didn't readmitted (the "No" class). So, in this study, balance data before modeling was taken into consideration.

Table 4 Confusion Matrix from imbalance data

A picture containing drawing

Description automatically generated

Table 5 Accuracy from imbalance data

A screenshot of a cell phone

Description automatically generated

4**. Results**

With the HRRP carried by the Centers for Medicare and Medicaid (CMS), hospitals have become strongly interested in reducing the readmission rate. During this study, we used the dataset composed of 101766 diabetic patient’s readmission records, created prediction models with several popular machine learning algorithms to spot patients at high readmission risk.

This research demonstrated whether predictive modeling may be accustomed to predict diabetes hospital readmission risk, also to identify high-risk factors that affect diabetes hospital readmission. This study used CRISP-DM methodology with a 10-fold cross validation, compared some popular machine learning models (Logistic regression, Random forest, MLP, Adaboost) in predicting diabetic patients 30 days readmission, under sampling was conducted to solve the data imbalance problem. All the predictive models were tested for accuracy, specifity, sensitivity, precision and AUC. A machine learning model for predicting high-risk readmission patients is only useful if a large fraction of patients at high-risk are correctly identified (i.e. high sensitivity) without raising a large number of false alarms.

Confusion matrix is the primary source to estimate the model performance, we first calculated accuracy from the confusion matrix for each model. Based on the accuracy, logistic regression produced the best results with an overall prediction rate of 64.7%, random forest came out as the runnerup with an overall prediction rate of 64.6%, followed by MLP and Adaboost with an overall prediction rate of 62.9% and 61.4% (show in table 6). A careful examination of those results reveals that the prediction accuracy for the "No" class is significantly higher than the prediction accuracy for the "Yes" class. In fact, all four models predicted diabetic patients who don't seem to be likely to readmit with better than 90% accuracy while they did poorly on predicting the diabetes patients who are likely to readmit with accuracy just around 16%. Since main purpose of this study is to predict the "Yes" class, less than 50% accuracy for this class was considered not acceptable.

Table 6 Prediction results for 10-fold cross validation

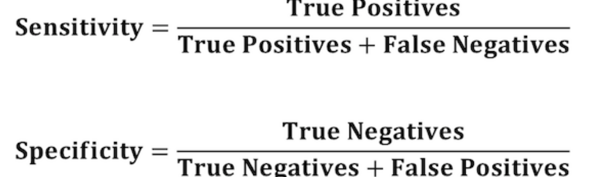
A screenshot of a cell phone

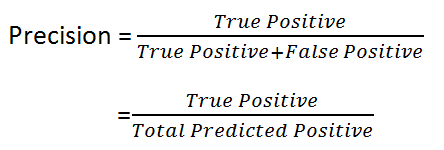
Description automatically generated

There're other measures for evaluating classification models except accuracy, like sensitivity (Fig. 6.), specifity (Fig. 6.), and precision (Fig.5.). Sensitivity is a measure of the proportion of actual positive cases that got predicted as positive (or true positive); Specificity is the proportion of actual negatives, which got predicted as the negative (or true negative); precision is the proportion of positive identifications was actually correct. In this study, specifity, sensitivity, and precision were employed to evaluate the performance of the models.

Fig. 6. Sensitivity and Specificity

Fig. 5. Precision

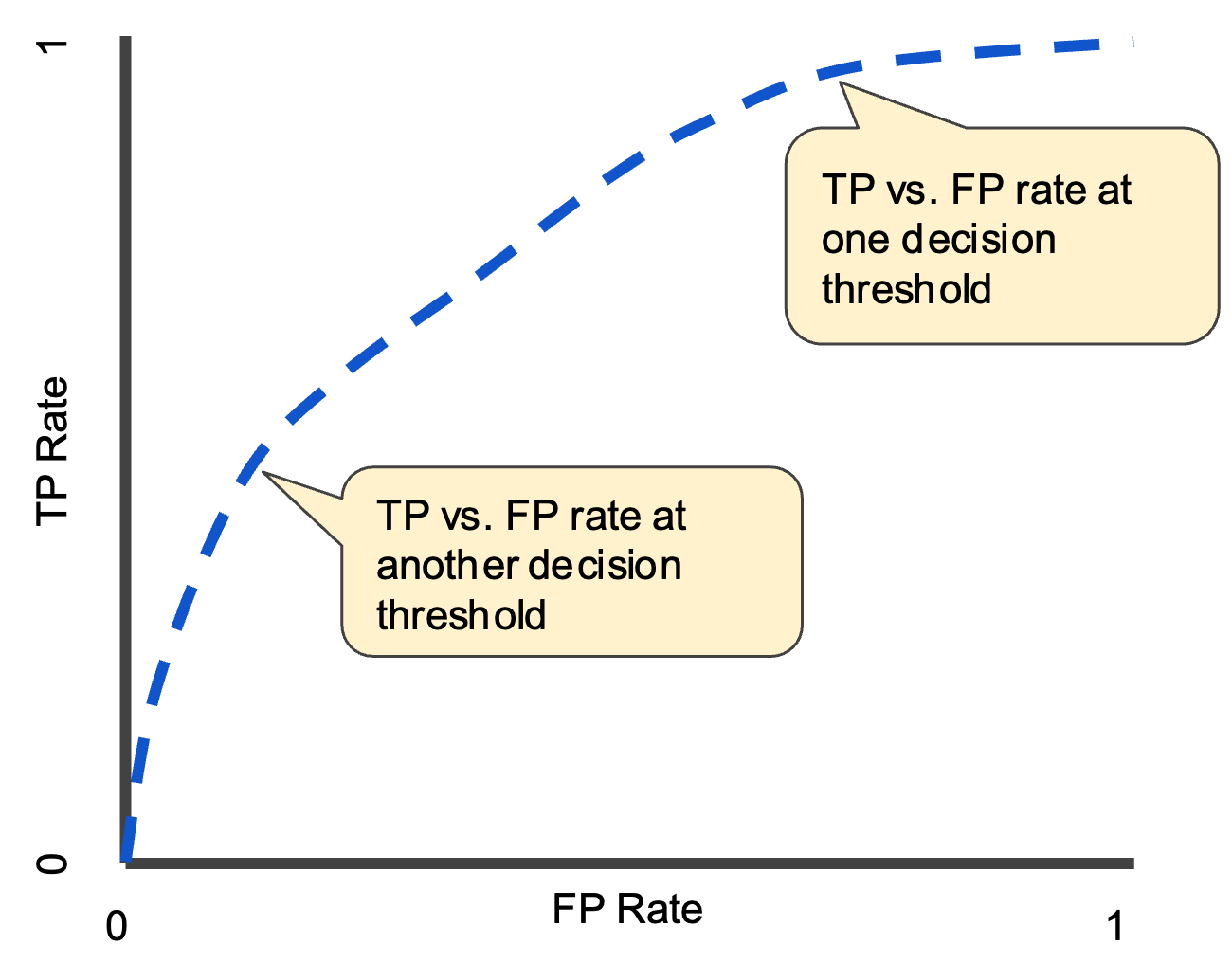




All algorithms also were evaluated using area-under-the-curve (AUC). AUC is the probability that a randomly chosen positive instance (in this study, “<30” represented as “1”) ranks higher than a randomly chosen negative one (in this study, “0”). AUC was conducted to determine strength of the predictive models in this study (show as Fig.8.). An AUC with a value equal or less than 0.5 indicates that the model is not better than a random guess, while an AUC of 1.0 indicates perfect classification. AUC is a commonly used metric for binary decision problems with highly skewed dataset. Our dataset is highly skewed because the number of patients who were readmitted within 30 days was only 11% while number of patients who didn't readmit or readmitted after 30 days was 89%. In this study. AUC of the different models were similar (0.61-

0.65), while the AUC value for random forest is 0.49.

Fig. 7. Area under the curve



TP Rate = TP/ (TP+FN)

FP Rate = FP/ (FP+TN)

The results (show in table 7) are compared by looking at all performance indicators (Accuracy, specifity, sensitivity, precision, AUC) for diabetic patients 30 days readmission. More specifically, for results of accuracy, LR performed the best, following by RF and MLP, then Adaboost.

In this use case, the goal is to determine if a patient will be readmitted to the hospital within 30 days or not. So, the correct metric to measure the efficacy of the model is to look at the precision and sensitivity. Precision gives the measure of how good we are able to predict if the patient will be readmitted, sensitivity tells us the patients which we missed to predict correctly. Although machine learning methods are becoming more regularized in medical field, but it is still challenging to predict the hospital readmission, due to its complexity and uncertainty.

Table 7 Results from predictive models

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Accuracy | Specifity | Sensitivity | Precision | AUC |
| LR | 0.647 | 0.521 | 0.663 | 0.163 | 0.617 |
| RF | 0.646 | 0.661 | 0.522 | 0.163 | 0.495 |
| MLP | 0.626 | 0.637 | 0.542 | 0.158 | 0.623 |
| Adaboost | 0.614 | 0.621 | 0.559 | 0.157 | 0.615 |

In machine learning cases, it is equally important not only have an accurate, but also an interpretable model. Apart from wanting to know which patient has high-risk of readmission, we also wondering what cause it. Tree-based models (random forest in this study) calculate feature importance by keeping the best performing features close to the root of the tree, then give an importance score to each feature. The larger the score the more important the feature. Our results show length of stay, admission type, number of inpatients, metformin, medication\_ change were the top five factors to predict diabetic patient readmission (show in fig.7.).

Through this project, we created a machine learning model that is able to predict the patients with diabetes with highest risk of being readmitted within 30 days. The best model was logistic regression, with an overall accuracy of 65%. But none of the models are strong since the True Positives (TP) are very low and False Negatives (FN) are quite high, which means we are unable to detect patients who are in high risk to readmit.

Fig. 7 Variable importance

A screenshot of a social media post

Description automatically generated

5. **Conclusion**

This study didn't get a promising performance to solve ‘Readmission Puzzle’, but it successfully identified some of the key features that drove readmissions, they're time\_in\_hospital, number\_inpatients, metformin, admission\_type, and change of medications, list by feather importance ranking.

Although even the best performance model (logistics regression) just had an overall accuracy of 64.7% with a specifity of 52.1%, sensitivity 0f 66.3%, and AUC of 61.7%, it is not good enough for health-care providers to make reliable decision based on this model, but it can help health-care providers identify the key risk factors that cause hospital readmission for diabetic patients. When the model predicts a particular patient who won't get readmitted, medical practitioner can use this prediction to get more insights and would assist him in further diagnosis. Physicians also can develop new strategies to reduce readmission rates by using the identified risk factors.

6. **Limitation and future study**

This study targets diabetic patients only, however, we believe this early work sets the stage for further research to other chronic disease like cancer, asthma, etc. In the future studies scheduled and unscheduled readmissions needs to be considered separately. Several critical features in the medical records, like medical specialty was found missing, hence a superior data collection drive needs to be done for future research in this regard. Some features that needs to be collected would be Date of Admission (to find if the season of the year has impact on patient readmission), Body Mass Index (to find impact of BMI on readmission), and Family History (to find hereditary information).

Reference

[1] [Centers for Medicare & Medicaid Service, Hospital Readmissions Reduction Program (HRRP)](https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/HRRP/Hospital-Readmission-Reduction-Program)

[2] [Centers for Disease Control and Prevention, Diabetes](https://www.cdc.gov/diabetes/basics/diabetes.html#:~:text=Diabetes%20is%20the%20seventh%20leading,diabetes%20has%20more%20than%20doubled.)

[3] [National Diabetes Statistics Report 2020](https://www.diabetesresearch.org/file/national-diabetes-statistics-report-2020.pdf)

[4] [CMS, Guide to Reducing Disparities in Readmissions](https://www.cms.gov/about-cms/agency-information/omh/downloads/omh_readmissions_guide.pdf)

[5] [Bojja, Giridhar Reddy and El-Gayar, Omar, "Predicting Hospital Readmissions of Diabetic patients - A Machine Learning Approach" (2019). Annual Research Symposium. 24](https://scholar.dsu.edu/cgi/viewcontent.cgi?article=1016&context=research-symposium)

[6] [Theodore Long, Inginia Genao, Leora I Horwitz, "Reasons for readmission in an underserved high-risk population: a qualitative analysis of a series of inpatient interviews" (2013)](https://bmjopen.bmj.com/content/3/9/e003212)

[7] [Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore, “Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records,” (2014)](https://www.hindawi.com/journals/bmri/2014/781670/)

[8] [Utomo Pujianto, Asa Luki Setiawan, Harits Ar Rosyid, Ali M. Mohammad Salah," Comparison of Naïve Bayes Algorithm and Decision Tree C4.5 for Hospital Readmission Diabetes Patients using HbA1c Measurement" (2019)](https://www.researchgate.net/publication/338128718_Comparison_of_Naive_Bayes_Algorithm_and_Decision_Tree_C45_for_Hospital_Readmission_Diabetes_Patients_using_HbA1c_Measurement)

[9] [Hephzibah Munnangi, MS, Dr. Goutam Chakraborty Oklahoma State University, Stillwater, Ok," Predicting Readmission of Diabetic Patients using the high performance Support Vector Machine algorithm of SAS® Enterprise Miner™" (2015)](https://www.researchgate.net/publication/279530614_Predicting_Readmission_of_Diabetic_Patients_using_the_high_performance_Support_Vector_Machine_algorithm_of_SASR_Enterprise_Miner)

[10] [Allen M. Glasgow, Jill Weissberg-Benchell, W. Douglas Tynan, Sandra F. Epstein, Chris Driscoll, Jane Turek and Ev Beliveau Pediatrics " Readmissions of Children With Diabetes Mellitus to a Children's Hospital" July 1991, 88 (1) 98-104;](https://pediatrics.aappublications.org/content/88/1/98)

[11] [Ahmad Hammoudeha,d, Ghazi Al-Naymata, Ibrahim Ghannamb, Nadim Obieda,c,](https://www.researchgate.net/publication/328887677_Predicting_Hospital_Readmission_among_Diabetics_using_Deep_Learning)

[" Predicting Hospital Readmission among Diabetics using DeepLearning" (2018)](https://www.researchgate.net/publication/328887677_Predicting_Hospital_Readmission_among_Diabetics_using_Deep_Learning)

[12] [Ti’jay Goudjerkan1, Manoj Jayabalan, School of Computing, Asia Pacific University of Technology & Innovation, Kuala Lumpur, Malaysia, " Predicting 30-Day Hospital Readmission for Diabetes Patients using Multilayer Perceptron" (2019)](https://thesai.org/Downloads/Volume10No2/Paper_36-Predicting_30_Day_Hospital_Readmission_for_Diabetes_Patients.pdf)

[13] [Arian Hosseinzadeh, Masoumeh Izadi, Aman Verma, Doina Precup, and David Buckeridge](https://www.academia.edu/4232658/Assessing_the_Predictability_of_Hospital_Readmission_Using_Machine_Learning)

[McGill University1140 Pine Ave West, Montreal, QC, " Assessing the Predictability of Hospital Readmission Using Machine Learning" (2013)](https://www.academia.edu/4232658/Assessing_the_Predictability_of_Hospital_Readmission_Using_Machine_Learning)

[14] [WIKIPEDIA Cross-industry standard process for data mining](https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining#:~:text=Cross%2Dindustry%20standard%20process%20for%20data%20mining%2C%20known%20as%20CRISP,most%20widely%2Dused%20analytics%20model.)

[15] [KDnuggets "CRISP-DM, still the top methodology for analytics, data mining, or data science projects"](https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html)

[16] [Suliman Hawamdeh, Hsia-Ching Chang,《Analytics and Knowledge Management》（2018）](https://books.google.com/books?id=u3NgDwAAQBAJ&pg=PT44&lpg=PT44&dq=.+Previous+research+showed+that+10+seem+to+be+the+optimal+number+for+folds+(the+optimizes+the+time+it+takes+to+complete+while+minimizing+the+bias+and+variance+associated+with+the+validation+process&source=bl&ots=qW5H60_DTr&sig=ACfU3U2bVWwvHmiyJI7quPH-aDkavydMrw&hl=en&sa=X&ved=2ahUKEwiT6834lu7pAhVSGs0KHZdUBGwQ6AEwAHoECAkQAQ#v=onepage&q&f=false)

[17] [UCI Machine Learning Repository, Diabetes 130-US hospitals for years 1999-2008 Data Set 2014-05-03](https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008)

[18] [Colleen K. McIlvennan, DNP, ANP, Zubin J. Eapen, MD, MHS, and Larry A. Allen, MD, MHS, " Hospital Readmissions Reduction Program" 2015-May-19](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4439931/)

[19] [ResDAC, Claim Inpatient Admission Type Code (FFS)](https://www.resdac.org/cms-data/variables/claim-inpatient-admission-type-code-ffs)

[20] [CMS Clarification of Patient Discharge Status Codes and Hospital Transfer PoliciesJA0801](https://www.cms.gov/Medicare/Medicare-Contracting/ContractorLearningResources/Downloads/JA0801.pdf)

[21] [ResDAC, Claim Source Inpatient Admission Code (FFS)](https://www.resdac.org/cms-data/variables/claim-source-inpatient-admission-code-ffs)

[22] [WIKIPEDIA List of ICD-9 codes](https://en.wikipedia.org/wiki/List_of_ICD-9_codes)

[23] [Dursun Delen, Glenn Walker, Amit Kadam, " Predicting Breast Cancer Survivability: A Comparison of Three Data Mining Methods" (2004)](https://pubmed.ncbi.nlm.nih.gov/15894176/)

[24] [Melody Y. Kiang, "A comparative assessment of classification methods" 1 May 2002](https://personal.utdallas.edu/~ryoung/phdseminar/DataMiningComparison-Melody.pdf)

[25] [Matt Bogard, Tuesdi Helbig, Gina Huff , Chris James, " A Comparison of Empirical Models for Predicting Student Retention"](https://www.wku.edu/instres/documents/comparison_of_empirical_models.pdf)

[26] [DursunDelen, " A comparative analysis of machine learning techniques for student retention management", November 2010](https://www.sciencedirect.com/science/article/pii/S0167923610001041)