Identification of Critical Risk Factors Leading to Short-Term Readmission of Diabetic Patients

Aishwarya Subash Chandra Bose[1], Choo Ming Hui Raymond[1],
Jiang Zhiyuan[1], Siew Yaw Hoong[1],
Mao Bowen[1], Wang Jingli[1]

Student Affiliation:
[1] M. Tech students, Institute of Systems Science, National University of Singapore, Singapore

ABSTRACT

The topic of short-term readmission in public hospital has recently become more important due to the high cost involved and limited public healthcare resources. Identification of inpatients with high risk of short-term readmission can effectively reduce such cost, and alternative medication can be provided in advance to reduce frequent readmission. In our research, 10 years (1999-2008) of clinical care data from 130 hospitals and 70,000 diabetic inpatients in the US was used for data mining and analytics for two objectives: (1) To predict diabetes patient readmission using machine learning classification methodologies such as logistics regression, support vector machine, XGBoost and neural network. Our results showed that XGBoost achieved the best prediction with over 82% accuracy and 83% average F1 score. The performance of logistics regression and neural network also exceeded current developments. (2) To extract critical risk factors that correlates with readmission of diabetic patients. Our results showed that *number of inpatient* and *number of diagnoses* are the two most critical factors in readmission prediction. These results provided valuable suggestions to inpatients monitoring policy that may reduce short-term readmission and public healthcare cost in the future.

1. INTRODUCTION

Diabetes is a chronic disease associated with abnormally high levels of sugar (glucose) in the blood. The two types of diabetes are referred to as Type I & Type II wherein the body does not produce or properly use insulin, a hormone that helps the glucose to get into your cells to give them energy. Over time, having too much glucose in blood may cause serious problems such as heart disease and stroke. It is noted that the management of hyperglycemia and other risk factors control at the early stage in hospitalized diabetic patients has a significant effect on the outcome of diabetes treatment, in terms of both morbidity and mortality [1].

However, the burden of hospitalized diabetic patients is substantial and increasingly costly, and readmission plays a significant role in it. Patients facing a high risk of readmission need to be identified at the time of being discharged from the hospital, to facilitate improved treatment to reduce the chances of their readmission [2]. Patients with diabetes represented about 9% of the US population in 2014, but they account for approximately 25% of hospitalizations (over eight million per year) [3-5]. The hospital readmission rate for certain conditions not only affect the cost of the health care adversely, but it is also considered an indicator of hospital health care quality. Providing care in advance that is respectful of and responsive to individual patient preferences, needs, and values is a better way to ensure proper clinical decisions [6].

Enormous previous literatures of risk factors analysis are devoted to developing new tools for predicting diabetic hospital readmission risks [7]. Strack et al studied the impact of HbA1c on readmissions [1]. Jiang et al explored demographic and socioeconomic factors which may influence diabetic readmission rates [8]. Bhuvan et al evaluated different machine-learning algorithms, such as Bayesian Networks, Adaboost, and Random Forest, considering both short-term and long-term readmissions for diabetic patients using public data of patients [9].

In order to reduce the diabetic patients' hospital readmission rate and to provide a more effective risk control and interventions, an ensemble understanding of the latent causes and risk factors for readmission is crucial [10]. Besides, the previous study demonstrated that focus on patients with defined disorders may yield a higher reward in terms of improved patient care than attempts to reduce readmissions in the general population of the inpatients [11]. Therefore, reducing readmission rates of diabetic patients has the potential to greatly reduce health care costs while simultaneously improving corresponding health care quality [7].

Considering the above situation, the focus of this study is twofold: first, to explore the dataset containing patients' data and perform data cleaning, data exploration; and second, to build four different analytical methods (Logistic Regression, XGBoost, Support Vector Machine and Neural Network) to improve the prediction of readmission rate of the diabetic patients and compare them.

2. MATERIALS

2.1. Data Exploration

The dataset used for this study is extracted from Health Facts database (Cerner Corporation, Kansas City, MO), a national data warehouse with comprehensive clinical records across hospitals throughout the United States. Before further investigations, all data in the database is de-identified (in compliance with the Health Insurance Portability and Accountability Act of 1996) to make sure that the continuity of patient encounters within the same health system (EHR system) could be preserved properly. The selected dataset for this study contains 10 years' (1999-2008) of clinical care records in 130 US hospitals. The full dataset is available online at UCI Machine Learning Repository and the detailed information of attributes is listed in the list of attributes and descriptions in the initial dataset attached in Appendix A [1].

The selected dataset consists of 50 features including patient's demographic, diagnoses details, medications used, etc. With a focus on analyzing diabetic encounters (i.e. at least one of the three primary diagnosis was diabetes), only 101,766 records are applicable for this study. The dependent variable to predict is 'readmitted', the days of inpatient readmission; which has 3 valid values: "<30" if patient was readmitted within 30 days, ">30" if patient was readmitted in more than 30 days and "No" if no readmission occurred. The rest of the 49 variables are a mixture of numeric and nominal types.

2.2. Data Preparation

2.2.1 Data Cleaning

Firstly, variables with extremely high percentage of missing values (namely '*weight*' (97%), '*payer_code*' (52%) and '*medical_specialty*' (53%)) are dropped due to their poor interpretation. Meanwhile, variables with neglectable number of missing values ('*gender*' (3 records)) are assessed and the corresponding records containing such missing values are also dropped from further investigation. In addition, records with all 3 diagnosis values (Primary (*diag_1*), Secondary (*diag_2*), and Additional (*diag_3*)) missing are also removed. With a purpose of predicting readmission, records where the patients died during hospitalization are also removed as they have no chance of readmission.

Secondly, variables are classified based on their nature to reduce complexity so that data could fit properly in modelling to have meaningful interpretations. In the original dataset, each diagnosis variable ('*diag_1*', '*diag_2*' and '*diag_3*' representing primary, secondary and additional diagnoses) contains around 700-900 unique ICD (International Classification of Diseases) codes. Following ICD-9 Guideline, those codes are classified and recategorized into 9 diseases as shown in Table 1 below:

Table 1. ICD9 Diagnosis Codes Mapping

Diagnosis Code	Mapped New Diagnosis Code
[140, 240)	Neoplasms
250.x	**Diabetes**
[390, 460) and 785	Circulatory
[460, 520) and 786	Respiratory
[520, 580) and 787	Digestive
[800, 1000)	Injury
[710, 740)	Musculoskeletal
[580, 630) and 788	Genitourinary
"V", "E" and others	Others

With the focus on diabetic readmission in this study, the corresponding "*Diabetes"* records (i.e. at least one of '*diag_1*', '*diag_2*' & '*diag_3*' fell into "*Diabetes"* category) are filtered out for analysis.

Besides the 3 levels of diagnoses, recategorization is also performed on other applicable variables:

- '*admission_type_id*', recategorized from 8 into 4 categories ("*urgent*", "*elective*", "*newborn*" & "*others/unknown*")
- '*discharge_disposition_id*', recategorized from 29 into 4 categories ("*discharged to home*", "*discharged/transferred to higher or same priority*", "*discharged/transferred to lower priority*" & "*others/expired*")
- '*admission_source_id*', recategorized from 26 into 4 categories ("*referral*", "*transfer*", "*emergency*" & "*others*")

In the dataset, there are 23 kinds of medications with one variable representing each kind. As most of them only has a limited number of tested samples, the medication information seems to be over detailed and cannot contribute well in interpretation. Thus, medication variables with less than 100 numbers of tested samples are dropped and only 11 meaningful medications are left.

'*Age*', presented as categorical intervals in original dataset, is converted into a continuous variable by taking the median value of each interval.

Lastly dummy coding is performed to convert categorical variables into dichotomous variables for modelling purpose.

2.2.2 Outlier Removal

When assessing the numerical variables, all values falling outside 6 Standard Deviation (SD, +3 SD and -3 SD) interval of its variable-mean are identified as outliers of this variable. Records containing any numbers of outliers are all removed (0.3% records affected).

2.2.3 Data Scaling (Normalization)

To ensure that all variables will be scaled to the same range of (0,1) and a numerical variable will have the same range as its dummy variables, data scaling is performed by applying the formula of:

$$X_{sd} = \frac{(X_i - X_{min})}{(X_{max} - X_{min})} \quad (1)$$

2.2.4 Data Splitting

After the above processes, the cleaned-up dataset with 35,477 records of 70 variables is further randomly split into 70% training data and 30% testing data.

2.2.5 Data Balancing

In the original dataset, there is a data imbalance in 'readmitted' variable with only 10% of its records are marked as '< 30 days' which will give a baseline accuracy of 90% with poor precision. To reduce the effects of this data imbalance and improve the modelling result, synthetic minority over-sampling technique (SMOTE) is applied to obtain equal representation of the overrepresented and underrepresented classes.

3. METHODOLOGY

In order to predict more readmitted diabetic inpatients, 4 models (Logistics Regression, XGBoost, SVM and neural network) are constructed according to the supervise learning framework. Using the features from section 2.2.1, comply to the steps including baseline model training, tuning key parameters to optimize model performance, validating the model performance, testing and evaluating the model performance.

The original target "*readmitted*" has 3 categories, "*<30 days*", *">30 days*" and "*No Readmission*". Since the research target just focus on recognizing early-stage readmission (namely repeatedly frequently shortly readmitted), hence the target is defined into Class "0" ('*>30 days*' and '*No Readmission*') and Class "1" ('*<30 days*').

Our objective of modeling is further developed to improve the prediction of readmission (True Positive) while balancing the number of False Positive and False Negative.

3.1. Logistics Regression

Logistic regression is a statistical method of using multiple independent variables to predict the outcome of binary dependent variable, the results can be either 0 or 1. The method derives its name from the Logistic Function, also called Sigmoid Function, used in determining the binary outcome of the prediction. Logistics regression generates an equation with the coefficients, standard errors and significance levels to predict a logit transformation of the probability of the outcome. The coefficients can be used to estimate the odds ratios for each of the independent variables in the model.

In our case, the logistics regression model is built to predict the diabetes inpatients readmitted, listed the significant factors according to the coefficients. The steps are followed as listed below:

- Calculate the correlation across variables to avoid collinearity.
- Use RFE (Recursive Factor Elimination) to select the significant factors according to P-values hypothesis.
- Build Logistics Regression model, evaluate on the test dataset.

3.2. Support Vector Machine

Support Vector Machine is a supervised model for classification and regression analysis. In classification analysis, hyperplanes were constructed to separate the input data into predicted classes. New data are then tested against the decision function to see which side of the hyperplane they fall on to determine their classification. The support vector machines (SVM) require the solution of the following optimization problem:

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i \quad (2)$$

$$subject\ to \quad y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i,\ \xi_i \geq 0 \quad (3)$$

For parameters of the model includes Cost (*C*), *gamma* (detail level), we did a grid search through cross validation to select the optimal parameters for training the model. Parameters tested were:

- Weight for output classes (0.1, 0.3, 0.5, 0.7, 0.9)
- Cost (*C*=1, 10, 100, 1000)
- *Gamma* (0.1, 0.01, 0.001, 0.0001)
- Kernel functions (*sigmoid*, *RBF*, *polynomial*)
- 3-fold cross validation

The best results from the grid search were used to construct the SVM model for prediction with the test dataset. The prediction was then compared with the actual results and classification report, confusion matrix and ROC curves were created.

3.3. Neural Network

Neural network is a machine learning method for both supervised and unsupervised learning for classification and regression analysis. It is an interconnected network of nodes that simulate how the neurons in the brain function. During model training, the nodes take in the input variables, process them according to the activation function chosen, and output the predicted results that best match the actual data. This neural network would then become the statistical model that is used on new data to predict the outcome.

In our case, the neural networks structure configuration is listed in Table 2 after tuning the nodes number of hidden layer. Sequentially, the *epoch* is set as 50 and *batch_size* as 100, Adaptive Moment Estimation (Adam) is applied to compute adaptive learning rate for neural network learning.

Table 2. Structure and Parameters Configuration

Layer (Activation type)	Output Shape	Param
Input Layer	70	0
Hidden Layer 1 (relu)	70	4970
Hidden Layer 2 (relu)	20	1420
Output Layer (softmax)	2	42
Total params 6,432		

In addition, two optimised approaches for building the neural networks are implemented. Firstly, dropout rate is set to 0.1 to avoid overfitting, as neural networks tend to overfit the training data if the network is trained under a large epoch. Secondly, since ROC-AUC is considered more valuable to predict readmission cases instead of overall accuracy, the AUC is customised with a cost function to enhance the model's ability to track recall. Figure 1 shows the learning curves of loss variation and AUC trend during training.

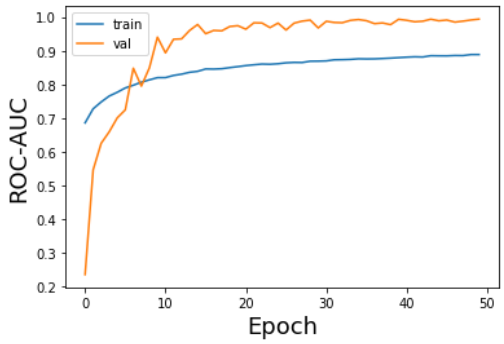

Fig 1. Neural Network Model ROC-AUC vs Epoch

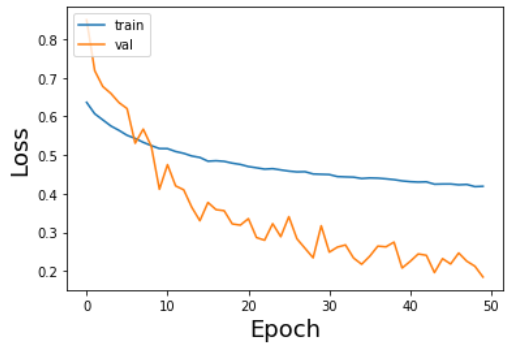

Fig 2. Neural Network Model Loss Curve vs Epoch

3.4. XGBoost

Extreme Gradient Boosting (XGBoost) [13], is an ensemble boosting method. It is based on the Gradient Boosted Decision Tree algorithm, but uses a more regularised model formalisation to control over-fitting. Boosting refers to the technique of adding new models to correct the mistakes of previous models. Models are added until no further improvements can be made. It is called gradient boosting because it uses a gradient descent algorithm to minimise the loss when adding new models.

XGBoost is known for its performance, speed and flexible parameter tuning. Starting with a baseline, training metrics are set to "AUC" and 6 sets of parameters are tuned step by step:

- Fix learning rate and number of estimators for tuning tree-based parameters.
- Maximum tree depth "*max_depth*" in range (3, 10, *step* = 1) and Minimum sum of instance child weight "*min_child_weight*" in range (1, 6, *step* = 2)
- Minimum loss reduction "*gamma*" in range (0, 0.5, *step* = 0.1)
- Tuning "*subsample*" in range (0.6, 1, *step* = 0.1*) and "colsample_bytree" in range (0.6, 1, step = 0.1).*
- Tuning Regularization Parameters "*reg_alpha*" in set (0.001, 0.01, 0.1, 1, 10).
- Reduce the Learning rate "*learning_rate*" to 0.01

Following these 6 steps, the best key parameters are identified using grid search and 5-fold cross validation, and the XGBoost model was built and applied to the test data to compare the results.

4. PREDICTION RESULTS

4.1. Evaluation criteria

Since our target class "1" is a small sample from a highly skewed data distribution, hence overall accuracy is tend to have less priority (large negative sample affect the detection result) when evaluate the test performance. Additonally, 3 harmonic indicators are utilized to measure the prediction performance, which are F1-score, Area Under Curve of Receiver Operating Characteristics Curve (AUC-ROC) and Area Under Curve of Precision Recall Curve (AUC-PR).

$$Accuracy = \frac{Tp+Tn}{Tp+Tn+Fp+Fn} \quad (4)$$

$$Precision = \frac{Tp}{Tp+Fp}, \quad Recall = \frac{Tp}{Tp+Tn} \quad (5)$$

$$F1\_score = \frac{2 \cdot Precision \cdot Recall}{Precision+Recall} \quad (6)$$

4.2. Fine tuning model parameters

In order to improve the predictions in modeling, parameter tuning is used to modify the key parameters in SVM, XGBoost and Neural Networks. The optimal parameters sets are listed in the Table 3.

Table 3. Optimal Parameters for 3 Models

Models	Optimal parameters
SVM	*class_weight*=8.9 *C*=10, *gamma*=0.01
XGBoost	*max_depth*=3, *min_child_weight*=6 *gamma*=0.1, *colsample_bytree*=0.9 *subsample*=0.8, *reg_alpha*=0.1
Neural Network	*dropout_rate*=0.1 *nodes_hiddenLayer*=20

4.3. Test prediction result

After training, 4 models are applied to the test data for predictions, and the results compared using 5 indicators: accuracy, F1-score, average F1-score, ROC-AUC and PR-AUC. Results are listed in the Table 4, and the ROC curves in Figure 3.

Table 4. Comparing Accuracies of Different Models (Class 1)

Indicators	Logistic Regression	XGBoost	SVM	Neural Network
Accuracy	0.68	**0.82**	0.64	0.65
F1-score	0.29	0.27	0.25	0.28
Avg F1-score	0.74	**0.83**	0.71	0.70
ROC-AUC	0.69	**0.66**	0.67	0.68
PR-AUC	0.249	0.191	0.219	0.246

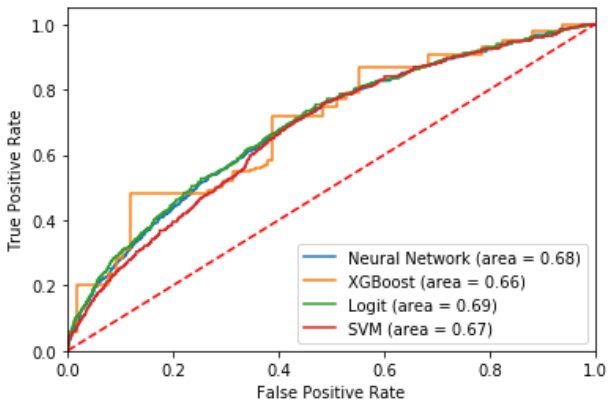

Fig 3. ROC Curves of Different Models

F1-score and ROC-AUC results are similar across all 4 models. Logistic regression model shows the best F1-score of 0.29 and ROC-AUC of 0.69. Although XGBoost does not have the best F1-score (0.27) or ROC-AUC (0.66), its overall accuracy of 0.82 outperforms the other models.

In order to compare the models' prediction capability with current developments, models from similar studies [9] are listed in Table 5 for comparison. In these studies, Bayes Network, Naïve Bayes, Neural Network and Random Forest models were used to predict readmissions. Among these models, the best performing one in terms of Precision-Recall Area Under Curve (PR-AUC) was the Random Forest model (0.242) [9]. In comparison, our Logistic Regression and Neural Network models show better PR-AUC at 0.249 and 0.246 respectively (Figure 4).

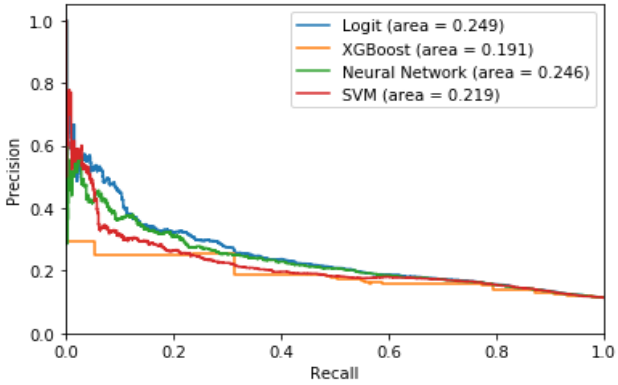

Fig 4. Precision / Recall curve of Different Models

Table 5. Comparison with Current Developments

Benchmark	PR-AUC	Models	PR-AUC
Adaboost	0.167	**Logistics Regression**	**0.249**
Bayes Network	0.208	XGBoost	0.191
Neural Network	0.233	**Neural Network**	**0.246**
Random Forest	0.242	SVM	0.219

5. RISK FACTORS ANALYTICS

To identify the critical risk factors to monitor upon admission, the Logistic Regression and XGBoost models were studied to identify variables that contribute the most to readmission in both models. The RFE algorithm was used on the Logistic Regression model to determine the top 10 variables contributing to readmission. Similarly, the relative importance of the variables in the XGBoost model was calculated based on the F score. It's noted that while it's possible to determine the polarity of the relationship in the Logistic Regression model, analysis of the XGBoost model is only limited to the importance of the variable.

From the results, the "*Number of inpatient visits in the past one year*" variable has the most contribution to readmission in both models (Fig 5 and 6). The second variable that has a significant contribution to readmission in both models is the "*Number of Diagnoses entered into the EHR System*". Based on these observations, it can be interpreted that if a patient has a high frequency of admission to the hospital in the past one year, and has a high number of medical diagnoses entered to the EHR system (e.g. a patient with a history of Diabetes, Coronary Artery Disease, Kidney Failure entered into the system compared to a patient with a history of only Diabetes), there is a very strong likelihood that the patient will be

readmitted to the hospital. Also, it's observed from both models that "*Number of Outpatient Visits in the past 1 year*" is an important variable and has potential in reducing readmission rates. As such, high risk patients can have their risk reduced by increasing outpatient follow-up.

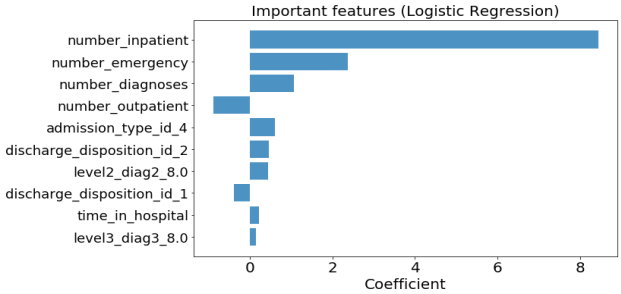

Fig 5. Ranking of Important Variables from Logistic Regression

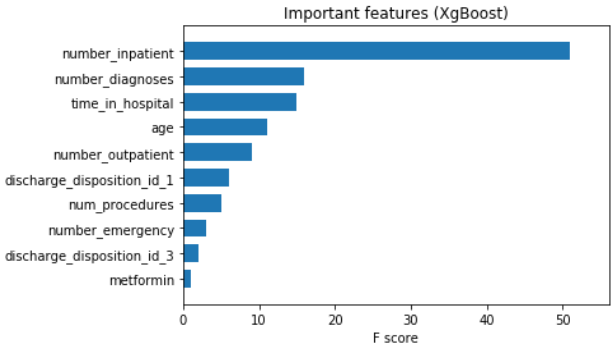

Fig 6. Ranking of Important Variables from XG Boost

With the common important predictors across both models identified, a study was done to understand the percentage of patients that were readmitted based on each scenario (e.g. 8% of patients with 1 inpatient visit in the past year were readmitted). Figure 7 shows that the percentage of patients readmitted is positively correlated to the inpatient visits in the past year. In addition, Figure 8 shows that as the number of diagnoses reaches 10, a spike in the readmission rates is observed.

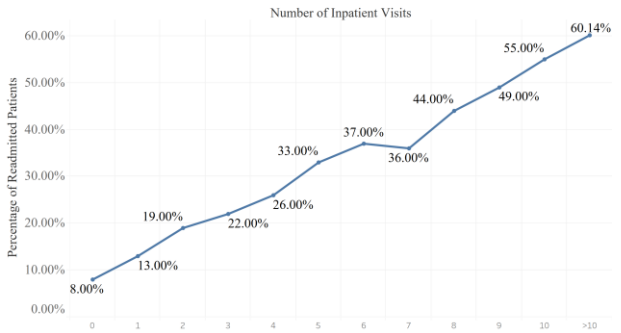

Fig 7. Percentage of Readmitted Patients by the Number of Inpatient Visits in the Past 1 Year.

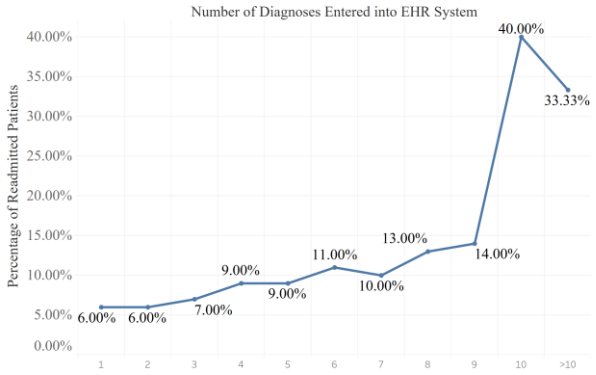

Fig 8. Percentage of Readmitted Patients by the Number of Diagnoses Entered into the EHR System

6. CONCLUSIONS

The US Medicare Payment Advisory Commission states that $12 billion is spent on preventable readmissions annually [12]. As such, patients facing high risk of readmission need to be identified at the point discharge, not only to reduce costs on the patients, but also to improve treatment to reduce the chances of their readmission [2]. While other papers predicting readmission using the dataset have confirmed that the entire dataset contains diabetic encounters, it was noted during the data cleaning stage that 70% of the records do not have diabetes (ICD9 code 250.X) indicated in any of the diagnoses (diag1, 2 and 3) [1, 2]. Due to the imbalance of classes in the *readmitted* attribute in the dataset (10% vs 90%), the SMOTE technique was used to applied to obtain equal representation of the overrepresented and underrepresented classes. Four analytical models (Logistic Regression, XG Boost, SVM and Neural Network) were built to predict the readmission rate of diabetic patients, with fine tuning performed on the SVM, XGBoost and Neural Network models using the ROC-AUC as the basis. The logistic regression and Neural Network models were found to have outperformed current benchmarks [9]. Lastly, variables were evaluated by their contribution to the logistic regression and XGBoost model, with *number of inpatient* and *number of diagnoses* identified as common critical factors.

While there is an improvement in the prediction of short-term readmission in this paper, the current limitation is that it's not known if diabetes is the cause of their readmission (e.g. a patient's Coronary Heart Disease may not be caused by their existing diabetes condition). In addition, long-term readmission as defined as readmission more than 30 days from discharge in the dataset, which is not helpful as an admission after 31 days is the same as one 4 years later.

With a better data definition, the study can be focused on readmission caused by diabetes within a year of discharge, resulting in more actionable insights. With high risk diabetic patients identified, future research could investigate the effectiveness and means of preventing readmission, resulting in better outcomes.

REFERENCES

[1] Strack, B., Deshazo, J. P., Gennings, C., Olmo, J. L., Ventura, S., Cios, K. J., & Clore, J. N. (2014). Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. *BioMed Research International*, *2014*. https://doi.org/10.1155/2014/781670

[2] Bhuvan M S∗, Ankit Kumar†, Adil Zafar‡ , Vinith Kishore∗ Identifying Diabetic Patients with High Risk of Readmission.

[3] Centers for Disease Control and Prevention. National Diabetes Statistics Report: Estimates of Diabetes and Its Burden in the United States, 2014. Atlanta, *GA: U.S. Department of Health and Human Services, 2014.*

[4] HCUP Nationwide Inpatient Sample (NIS) 2012. Agency for Healthcare Research and Quality (AHRQ). 2014. http://hcupnet.ahrq.gov/HCUPnet.jsp. Accessed June 15, 2014

[5] HCUP Nationwide Inpatient Sample (NIS). Agency for Healthcare Research and Quality (AHRQ); 2011. http://hcupnet.ahrq.gov/HCUPnet.jsp. Accessed August 11, 2013.

[6] Daniel J. Rubin, "Hospital Readmission of Patients with Diabetes," *Current Diabetes Reports, vol. 15, no. 4, 2015.*

[7] Reena Duggal, Suren Shukla, Sarika Chandra, Balvinder Shukla, Sunil Kumar Khatri. Predictive risk modelling for early hospital readmission of patients with diabetes in India. *Int J Diabetes Dev Ctries (October–December 2016)* 36(4):519–528. DOI 10.1007/s13410-016-0511-8.

[8] Jiang HJ, Stryer D, Friedman B, Andrews R. Multiple hospitalizations for patients with diabetes. *Diabetes Care*. 2003;26(5):1421–6.

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

[10] Umpierrez GE, Isaacs SD, Bazargan N, You X, Thaler LM, Kitabchi AE. Hyperglycemia: an independent marker of in hospital mortality in patients with undiagnosed diabetes. J Clin Endocrinol Metab. 2002;87(3):978–82.

[11] Benbassat J, Taragin M. Hospital readmissions as a measure of quality of health care: advantages and limitations. *Arch Intern Med. 2000*;160(8):1074–81.

[12] S. F. Jencks, M. V. Williams and E. A. Coleman, "*Rehospitalizations among Patients in the Medicare Fee-for-*Service Program," The New England Journal of Medicine, vol. 360, (14), pp. 1418-1428, 2009.

[13] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. ACM, 2016: 785-794.

Appendix A

Table A: The List of Attributes and Descriptions in the initial

Attributes Name	Feature Name	Type	Description and Values	% Missing
encounter_id	Encounter ID	Numeric	Unique identifier of an encounter	0%
patient_nbr	patient_nbr	Numeric	Unique identifier of a patient	0%
race	Race	Nominal	Values: Caucasian, Asian, African American, Hispanic, and other	2%
gender	Gender	Nominal	Values: male, female, and unknown/invalid	0%
age	Age	Nominal	Grouped in 10-year intervals: [0, 10), [10, 20), . . ., [90, 100)	0%
weight	Weight	Numeric	Weight in pounds.	0%
admission_type_id	Admission type	Nominal	Integer identifier corresponding to 9 distinct values, for example, emergency, urgent,	0%
discharge_disposition_id	Discharge disposition	Nominal	Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available	0%
admission_source_id	Admission source	Nominal	Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room and transfer from a hospital	0%
time_in_hospital	Time in hospital	Numeric	Integer number of days between admission and discharge	0%
payer_code	Payer code	Nominal	Integer identifier corresponding to 23 distinct values, for example, Blue Cross\Blue Shield, Medicare, and self-pay	52%
medical_specialty	Medical specialty	Nominal	Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family\general practice, and surgeon	53%
num_lab_procedures	Number of lab procedures	Numeric	Number of lab tests performed during the encounter	0%
num_procedures	Number of procedures	Numeric	Number of procedures (other than lab tests) performed during the encounter	0%
num_medications	Number of medications	Numeric	Number of distinct generic names administered during the encounter	0%
number_outpatient	Number of outpatient visits	Numeric	Number of outpatient visits of the patient in the year preceding the encounter	0%
number_emergency	Number of emergency visits	Numeric	Number of emergency visits of the patient in the year preceding the encounter	0%
number_inpatient	Number of inpatient visits	Numeric	Number of inpatient visits of the patient in the year preceding the encounter	0%
diag_1	Diagnosis 1	Nominal	The primary diagnosis (coded as first three digits of ICD9); 848 distinct values	0%
diag_2	Diagnosis 2	Nominal	Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values	0%
diag_3	Diagnosis 3	Nominal	Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values	1%
number_diagnoses	Number of diagnoses	Numeric	Number of diagnoses entered to the system	0%
max_glu_serum	Glucose serum test result	Nominal	Indicates the range of the result or if the test was not taken. Values: “>200,” “>300,” “normal,” and "none" if not measured	0%
A1Cresult	A1c test result	Nominal	Indicates the range of the result or if the test was not taken. Values: “>8” if the result was greater than 8%, ">7" if the result was greater than 7% but less than 8%, “normal” if the result was less than 7%, and “none” if not measured.	0%
Metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, citoglipton, insulin, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone, metformin, rosiglitazone, metformin-pioglitazone	23 features for medications	Nominal	Indicates whether the drug was prescribed or there was a change in the dosage. Values: “up” if the dosage was increased during the encounter, “down” if the dosage was decreased, “steady” if the dosage did not change, and “no” if the drug was not prescribed	0%
change	Change o fmedications	Nominal	Indicates if there was a change in diabetic medications (either dosage or generic name). Values: “change" and "no change"	0%
diabetesMed	Diabetes medications	Nominal	Indicates if there was any diabetic medication prescribed. Values: “yes” and “no”	0%
readmitted	Readmitted	Nominal	Days to inpatient readmission. Values: “<30” if the patient was readmitted in less than 30 days, “>30” if the patient was readmitted in more than 30 days, and “No” for norecord of readmission.	0%