Predicting Readmissions of Diabetic Patients Utilizing Machine Learning Methods

Yuwei Xia[†]
School of Economic Information Engineering
Southwestern University of Finance and Economics
Chengdu Sichuan China
xiayuwei0226@yahoo.com

ABSTRACT

Health care has become more and more important in U.S. nowadays and the service provided by health systems are now evaluated by a program called Hospital Readmissions Reduction Program (HRRP). This program aims to improve Americans' health care via penalising health systems that have higher than expected readmission rates. Thus, health systems are supposed to try hard decreasing their readmission rates. Nevertheless, patterns of readmissions are hard to identify only according to clinical expertise, owing to the diversity and complexity of demographic, social, diseases and diseases-related characteristics. Therefore, this research targets diabetic patients and identifies readmission patterns of them. Seven machine learning models are applied to do the classification, containing Logistic Regression, Decision Tree, Support Vector Machine, Naive Bayes, K-Nearest Neighbour, Random Forest and Gradient Boosting Decision Tree. Nevertheless, all these models behave badly due to the extremely imbalanced dataset. Thus, a method called Borderline-SMOTE is utilized to rebalance dataset. Afterwards, models are trained again and performance of these models is evaluated and compared. The results show that performance of all models improves, while Random Forest, whose f1-score is 93.4% and recall 88.7%, outperforms all the other methods, including those used in previous researches. Finally, top 20 features contribute most to readmissions of diabetic patients are identified by Random Forest.

1 INTRODUCTION

Health care has become one of the most popular industries nowadays. In U.S., a program called Hospital Readmissions Reduction Program (HRRP) launched by the Centres for Medicare and Medicaid services (CMS) was established to improve Americans' health care via penalising health systems with higher than expected readmission rates. Situations that count as readmission include all-cause unplanned readmissions happening within 30 days of discharge from the index (i.e., initial) admission and readmissions to the same hospital or another applicable acute care hospital for any reason. In order not to be penalized, it is prominent for health systems to try their best decreasing their readmission rates by analysing and discovering the patterns of readmissions. However, reality is that the causes of readmissions are difficult to evaluate only based on clinical expertise, on account of the diversity and complexity of diseases, demographic, social and disease-related characteristics [11]. Therefore, the objective of this research is to model and analyse the readmission patterns to predict patients with high-risky admission accurately. Moreover, due to the huge differences among diseases and patients, models based upon a specific patients' characteristics relevant to certain diseases are more helpful and useful than models based upon general cohort aiming to handle with multiple diseases [23].

This research targets diabetes for several reasons. For one thing, diabetes affects a large number of people in U.S. For another, diabetes could lead to perilous complications including stroke, heart disease, kidney damage and nerve damage if without ongoing and careful management [30][32]. Moreover, the costs on diabetes is numerous. As a result, decreasing 30-day readmissions of diabetic patients could not only be conducive to a great number of people via improving health care, but also reduce health care costs.

While most researches regarding diabetes intended to predict diabetes [1][2][22], a few studies have focused on readmissions of diabetes. S.Salian and G.Harisekaran aimed to determine the risk predictors that could cause readmission among diabetic patients and performed detailed analysis to predict risk of readmission of diabetic patients based on Decision Tree model[8]. S.Tutun etc. proposed a hybrid classification framework called Evolutionary Simulating Annealing LASSO Logistic Regression (ESALOR) to improve the classification of readmissions of diabetic patients[5]. R.Duggal etc. applied models including Naive Bayes, Bayes Network, Random Forest, Adaboost Tress and Neural Networks to identifying diabetic patients with high risk of readmission[23]. Although all of these studies have made some progress in figure out patterns of readmissions of diabetic patients, they have some limitations. First, the size of dataset is relatively small. Second, the imbalance of dataset classes is ignored. To be more specific, the number of patients who readmitted within 30 days is much more less than the number of patients who readmitted after 30 days or not readmitted. Third, most of the best-behaved models are not interpretable, which means we still do not know what are important factors affecting readmissions of diabetic patients.

In order to solve these problems, a dataset is obtained regarding readmission of diabetic patients from UCI Repository of Machine Learning Databases, which contains 101,765 encounters and 50 features [35]. Then data pre-processing is performed and several machine learning classification models are utilised to predict diabetic patients who are very likely to readmitted within 30 days, including Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), Naive Bayes (NB), K-Nearest Neighbour (KNN), Random Forest (RF) and Gradient Boosting Decision Tree (GBDT). The performance of models is evaluated by accuracy, precision, recall and f1-score. Nevertheless, all these models behave badly, with high accuracy but low precision, recall and f1-score. Obviously, poor performance of models results from the imbalanced dataset. Therefore, a method called Borderline-SMOTE is utilized to rebalance the dataset [37]. Afterwards, models are trained again and performance of models is compared. The results show that performance of all models improves, and the best classifier is Random Forest (RF), with the accuracy of 93.7%, precision 98.6%, recall 88.7% and f1-score 93.4%, which outperform all methods including those in previous researches [5][8][23]. Finally, top 20 most important patterns affecting readmissions of diabetic patients are decided by Random Forest. The research contributions of this paper are: 1) a resampling method is used to rebalance the dataset; 2) performance of machine learning classification models is obviously improved comparing with previous researches; 3) top 20 most important readmission patterns are determined via Random Forest model.

The remainder of this paper is organized as follows: Section 2 introduces the background and related works of our research; Section 3 describe basic information of the dataset and perform data pre-processing; Section 4 demonstrates machine learning models used in this research; Section 5 compares the performance of models and selects the best-behaved model to identify readmission patterns of diabetic patients; Section 6 concludes our work.

2 BACKGROUND AND RELATED WORKS

Health care is drawing more and more attention now. In U.S., Hospital Readmissions Reduction Program (HRRP), a program launched by launched by the Centres for Medicare and Medicaid services (CMS), aims to improve Americans' health care by penalizing health systems with higher readmission rates than expected. This program provides hospitals with a strong financial incentive to make their communication and care coordination efforts better and meanwhile, work better with patients and caregivers on post-discharge planning and applies excess readmission ratio (ERR) to measure hospital performance. The US Medicare Payment Advisory Commission has stated that $12 billion is spent on preventable readmissions annually, and a study estimated readmission cost of Medicare patients to be $17.4 billion per annum [28].

However, due to the complexity and diversity of demographic, social, diseases and disease-related characteristics, it is difficult to decrease readmission rates merely based on clinical expertise. Therefore, this research aims to predict diabetic patients with high-risk readmissions via machine learning methods and help health systems identify readmission patterns. Diabetes affects a huge number of people in U.S. In 2015, it was estimated that the number of people over 18 years of age with diagnosed and undiagnosed diabetes was 30.2 million, representing around 9.4% of the U.S. population [31]. Despite of that, the 30-day readmission rate of diabetic patients is 14.4%-22.7%, higher than the overall 30-day readmission rate, which is 8.5-13.5% [34]. For another, by impairing the body's ability to process blood glucose, it can lead to a build-up of sugars in the blood and increase the risk of dangerous complications including stroke, heart disease, kidney damage and nerve damage if without ongoing and careful management [30][32]. Moreover, it cost a large sum of money. In 2017, total costs of diagnosed diabetes in U.S. is 327 billion dollars, consist of 237 billion dollars for direct medical costs and 90 billion dollars in reduced productivity [33].

2.1 Discover important readmission features of diabetic patients with ESALOR

S.Tutun etc. proposed a hybrid classification framework called Evolutionary Simulating Annealing LASSO Logistic Regression (ESALOR) to improve the classification of readmissions of diabetic patients[5]. By applying the evolutionary strategy (ES) and simulated annealing (SA) algorithms and preventing over-training using regularization (Lasso), the coefficients of Logistic Regression (LR) is optimised. This model could help health care providers discover the most important risk features that cause readmissions of diabetic patients and allow physicians to develop new strategies to decrease readmission rates.

2.2 Determine predictors causing readmissions of diabetic patients based on Decision Tree

S.Salian and G.Harisekaran aimed to determine the risk predictors that could cause readmission among diabetic patients and performed detailed analysis to predict risk of readmission of diabetic patients based on Decision Tree[8]. Big Data analytics have been applied to evaluate the risk of readmission for diabetes patients and predictive modelling has been employed by applying Decision Tree classification method. Although the chance of readmission is successfully predicted using the above analysis to some extent, the prediction results are not that satisfying and more works need to do.

2.3 Identify patterns relevant to readmissions of diabetic patients by Random Forest

R.Duggal etc. applied models including Naive Bayes, Bayes Network, Random Forest, Adaboost Tress and Neural Networks to identifying diabetic patients with high risk of readmission and selected the best classifier, which is Random Forest[23]. Mining latent patterns in the diagnosis, medications, lab test results and basic features of patients, this model finds a strong set of statistically rules relevant to readmissions of diabetic patients. Nevertheless, the dataset used in this paper is relatively small and since it is imbalanced, all classifiers are biased towards majority class yielding high accuracy scores and low recall scores and f1-scores.

3 DATASET

3.1 Data Collection

The dataset applied for all analysis of this research is from the UCI Repository of Machine Learning Databases [35]. This dataset contains medical records of 101765 patients diagnosed with diabetes collected from 130 U.S. hospitals for 10 years (1999-2008) and includes 50 features describing the diabetic encounters, comprising demographics, diagnoses, diabetic medications, number of visits in the year preceding the encounter and payer information [36]. Features and descriptions are demonstrated in Table 1. All data satisfies the following criteria: 1) It is an inpatient encounter (a hospital admission); 2) It is a diabetic encounter, that is, one during which any kind of diabetes was entered to the system as a diagnosis; 3) The length of stay was at least 1 day and at most 14 days; 4) Laboratory tests were performed during the encounter; 5) Medications were administered during the encounter [36].

Feature name	**Description and values**
encounter_id	Unique identifier of an encounter.
patient_nbr	Unique identifier of a patient.
race	Caucasian, Asian, African American, Hispanic, and other.
gender	male, female, and unknown/invalid.
age	Grouped in 10-year intervals.
weight	Weight in pounds.
discharge_disposition_id	Integer identifier corresponding to 29 distinct values.
admission_type_id	Integer identifier corresponding to 9 distinct values.
admission_source_id	Integer identifier corresponding to 21 distinct values.
payer_code	Integer identifier corresponding to 23 distinct values.
medical_specialty	Integer identifier of a specialty of the admitting physician.
num_procedures	Number of procedures (other than lab tests) performed during the encounter.
num_medications	Number of distinct generic names administered during the encounter.
time_in_hospital	Integer number of days between admission and discharge.
num_lab_procedures	Number of lab tests performed during the encounter.
number_outpatient	Number of outpatient visits of the patient in the year preceding the encounter.
number_emergency	Number of emergency visits of the patient in the year preceding the encounter.
number_inpatient	Number of inpatient visits of the patient in the year preceding the encounter.
diag_1	The primary diagnosis (coded as first three digits of ICD9).
diag_2	Secondary diagnosis (coded as first three digits of ICD9).
diag_3	Additional secondary diagnosis (coded as first three digits of ICD9).
number_diagnoses	Number of diagnoses entered to the system.
max_glu_serum	Indicates the range of the result or if the test was not taken.
A1Cresult	Indicates the range of the result or if the test was not taken.
23 medications	The feature indicates whether the drug was prescribed or there was a change in the dosage.
change	Indicates if there was a change in diabetic medications (either dosage or generic name).
diabetesMed	Indicates if there was any diabetic medication prescribed.
readmitted	Days to inpatient readmission.

Table 1:Features and descriptions.

3.2 Data Pre-processing

Prior to feeding data into machine learning models, data pre-processing is a necessary procedure to improve the performance of models. First, data cleaning is performed. Meaningless features that are irrelevant to our target is dropped, which are "***encounter_id***" and "***patient_nbr***". Features with more than 50% missing values are dropped, including "***weight***", which has 97% missing values, "***payer_code***", which has 52% missing values, and "***medical specialty***", which has 53% missing values. Samples with missing values are simply deleted, since they only account for 3% of all samples. Second, since the dataset does not contain outliers, Min-Max scaling is used to shrink the range of the continuous data such that the range is fixed between 0 and 1. Moreover, One-Hot encoding is applied to convert categorically features into a form that can be provided to machine learning algorithms. Third, due to the fact that this research targets diabetic patients readmitted with 30 days, the '*Readmitted*' attribute is labelled as having two values: 1, if the patient was readmitted within 30 days of discharge, or 0, which covers both readmission after 30 days and no readmission at all [36]. The proportion of two classes is displayed in Figure 1. It clear that this dataset is extremely imbalanced.

$$\text{Readmitted} = \begin{cases} 1, \text{if the patient was readmitted within 30 days} \\ 0, \qquad\qquad\qquad\qquad\qquad\qquad\quad \text{otherwise} \end{cases} \quad (1)$$

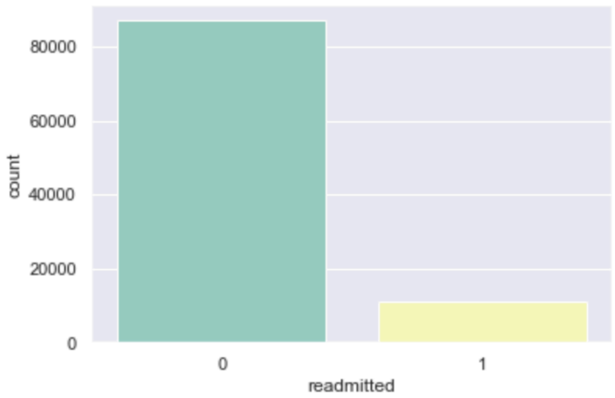

Figure 1:Proportion of two classes.

4 METHODS

After data pre-processing, 70% of the dataset is randomly divided as training set, which is used to train the machine learning models as well as parameter tuning, and 30% testing dataset to evaluate the performance of models.

Aiming to classify readmissions of diabetic patients, several widely-used classification methods are applied in this research, including Logistic Regression (LR), Decision Tree (DT), Support Vector Classification (SVC), Naive Bayes (NB), K-Nearest Neighbour (KNN), Random Forest (RF), and Gradient Boosting Decision Tree (GBDT). These models are first used to identify the patterns of readmissions of diabetic patients on training set based on features. The performance of all models is evaluated by predicting readmissions on testing set.

4.1 Logistic Regression

LR is a reasonably interpretable tool in multiple fields and is easy to implement with all kinds of programming languages. As a regression method to predict a dichotomous dependent variable, the maximum-likelihood ratio is used to determine the statistical significance of the variables in producing the LR equation [45]. LR is usually formulated mathematically as:

$$\mathrm{P(Y|X)} = \frac{1}{1+e^{-f(x)}}\ , \tag{2}$$

where X is an input matrix containing all observations and features, Y is an matrix that contains discrete and binary variables, such that y ∈ $\{0,1\}(y \in Y)$, and f(x) is a function consisting features and their corresponding weights that equals:

$$\mathrm{f(x)} = x_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \varepsilon, \tag{3}$$

where x_0, $x_1, \ldots, x_n$ represent all features, β_i represent corresponding weights for each features , and ε represents the random error process inevitably happening in the data generating process. Generally speaking, LR is well-suited for describing and testing hypotheses about relationships between a categorical outcome variable and one or more categorical or continuous variables [16].

4.2 Decision Tree

DT is a non-parametric supervised decision support tool that uses a tree-like graph or model to predicts the value of a target variable by learning simple decision rules of features [17]. Given training vectors $x_i \in R^n, (i = 1, \ldots, I)$ and a label vector y ∈ R^l, a decision tree recursively partitions the space such that the samples with the same labels are grouped together. Let the data at node m be represented by Q. For each candidate split $\theta = (\mathrm{j}, t_m)$ consisting of a feature j and threshold t_m, partition the data into $Q_{left}(\theta)$ and $Q_{right}(\theta)$ subsets:

$$Q_{left}(\theta) = (x, y)|x_j \leq\ t_m, \tag{4}$$

$$Q_{right}(\theta) = Q \backslash Q_{left}(\theta). \tag{5}$$

The impurity at m is computed using an impurity function H(), the choice of which depends on the task being solved (classification or regression).

$$\mathrm{G(Q, \theta)} = \frac{n_{left}}{N_m} H\left(Q_{left}(\theta)\right) + \frac{n_{right}}{N_m} H\left(Q_{right}(\theta)\right) \tag{6}$$

Then select parameters that minimizes the impurity:

$$\theta^* = argmax_{\theta} G(Q, \theta), \tag{7}$$

and recurse for subsets $Q_{left}(\theta^*)$ and $Q_{right}(\theta^*)$ until the maximum allowable depth is reached. DT is a proper method for predicting readmissions of diabetic patients because it uses a white box model, which demonstrates that if a given situation is observable, the explanation for the condition is easily explained by Boolean logic [17].

4.3 Support Vector Classification

SVC is one of the SVM methods that could be use to perform classification tasks. The objective of SVM is to find a hyper-plane that has the largest distance to the nearest training data points of any class in an N-dimensional space that distinctly classifies the data points, where N is the number of features. Because SVC is effective in high dimensional spaces and it can capture much more complex relationships between features by using different kernel functions, they have received considerable research interest in classification over the past years [41].

4.4 Naïve Bayes

NB is a supervised learning model based on applying Bayes's theorem with the "naive" assumption that every pair of features is independent. Bayes's theorem states the following relationship:

$$\mathrm{P}(\mathrm{y}|x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots, x_n|y)}{P(x_1, \ldots, x_n)}\ , \tag{8}$$

where y is class variable and $x_1, \ldots, x_n$ are dependent feature vectors. After adding the "naive" assumption, this relationship equals:

$$\mathrm{P}(\mathrm{y}|x_1, \ldots, x_n) = \frac{P(y)\Pi_{i=1}^{n} P(x_i|y)}{P(x_1, \ldots, x_n)}. \tag{9}$$

Since $P(x_1, \ldots, x_n)$ is constant given the input, the following classification rule is used:

$$\mathrm{P}(\mathrm{y}|x_1, \ldots, x_n) \propto \mathrm{P(y)} \prod_{i=1}^{n} P(x_i|y), \tag{10}$$

$$\mathrm{y}' = argmax_y P(y) \prod_{i=1}^{n} P(x_i|y), \tag{11}$$

and Maximum A Posteriori (MAP) estimation can be applied to estimate P(y) and $\mathrm{P}(x_i|y)$. Despite the apparently over-simplified assumptions, NB has worked quite well in many real-world applications and is well known for its computational efficient and ability to handle missing data naturally and efficiently [10].

4.5 K-Nearest Neighbor

KNN is an instance-based classifier method. The advantage that k-nearest neighbours have over other algorithms is the fact that the neighbours can provide an explanation for the classification result; this case-based explanation can provide an advantage in areas where black-box models are inadequate [18].

4.6 Random Forest

RF consist of a set of decision trees, each of which acts as a weak classifier and is trained independently. The fundamental concept of RF is that a large number of relatively uncorrelated decision trees operating as a committee will outperform any of the individual constituent models. By pooling the responses from multiple decision trees, a strong classifier with higher predictive accuracy is formed and the class of an input is determined.

4.7 Gradient Boosting Decision Tree

GBDT is a predictive model in the form of an ensemble of decision trees and affords strong predictive power with a differentiable loss function [13]. It is a powerful classification technique applied in a wide range of fields due to its high accuracy, fast training and prediction time and small memory footprint.

4.8 Performance Evaluation Methods

To evaluate the performance of the different models, multiple performance criteria are applied, including accuracy, precision, recall, and f1-score. Accuracy, precision represents the ability of a classifier not to label as positive a sample that is negative, recall represents the ability of a classifier to find all the positive samples, and f1-score can be interpreted as a weighted harmonic mean of the precision and recall. These measures are calculated as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}, \tag{12}$$

$$\text{Precision} = \frac{TP}{TP+FP}, \tag{13}$$

$$\text{Recall} = \frac{TP}{TP+FN}, \tag{14}$$

$$\text{F1} - \text{score} = \frac{2TP}{2TP+FN+FP}, \tag{15}$$

where TP, FP, TN and FN are the number of true positives, false positives, true negatives and false negatives respectively.

5 EXPERIMENTS

5.1 Evaluation Results

Training set is used to train models and testing set is used to evaluate the performance of models. The results in Table 2 show that all models behave badly.

Model	Accuracy	Precision	Recall	F1
LR	0.887	0.488	0.024	0.046
DT	0.82	0.176	0.163	0.169
SVC	0.886	0.0	0.0	0.0
NB	0.146	0.113	0.954	0.201
KNN	0.877	0.230	0.036	0.036
RF	0.886	0.139	0.002	0.003
GBDT	0.887	0.243	0.345	0.248

Table 2:Accuracy, precision, recall and f1-score of seven machine learning models.

5.2 Rebalance Dataset

Obviously, the poor performance of models results from the extremely imbalanced dataset, with only 11,082 (11.3%) encounters are labelled as 1. Thus, a resampling method, Borderline-SMOTE, is used to balance data and improve the performance of models [24]. Based on SMOTE, Borderline-SMOTE is a minority over-sampling method, which only over-sample the minority examples near the borderline since the borderline examples of the minority class are more easily misclassified than ones far from the borderline [37]. After resampling the data, the number of encounters labelled as 1 is 87,061, which is the same as number of encounters labelled as 0. Hence the data is balanced. The final dataset has 174,122 encounters and 2,399 attributes in total. Then 70% of the dataset is randomly divided as training set, and 30% testing set.

Then models are trained again and the results are shown in Table 3: Evaluation Results of Seven Classification Models Clearly, after rebalancing data, performance of all models improves, which proves the Borderline-SMOTE method to be effective.

Model	Accuracy	Precision	Recall	F1
LR	0.697	0.698	0.697	0.697
CART4	0.900	0.905	0.894	0.900
ID3	0.897	0.903	0.896	0.890
SVC	0.560	0.534	0.834	0.665
NB-Gaussian	0.542	0.522	0.993	0.684
NB-Multinomial	0.664	0.654	0.699	0.675
NB-Complement	0.665	0.652	0.698	0.676
NB-Bernoulli	0.801	0.794	0.811	0.802
KNN	0.683	0.615	0.979	0.755
RF	0.937	0.986	0.887	0.934
GBDT	0.910	0.970	0.846	0.904

Table 3: Evaluation Results of Seven Classification Models.

Comparing the performance of models, we could find that RF outperforms all the other models, with accuracy of 93.7%, precision of 98.6%, recall of 88.7% and f1-score 93.4%. Thus we choose RF as the optimal model to identify readmission patterns of diabetic patients.

In order to improve the prediction ability of RF, the method of GridSearchCV is used for parameter tuning, which is a module in

ScikitLearn which uses the method of cross validation and provides an efficiency way to search for optimal parameters for an estimator. In this research, the cross-validation splitting strategy is 3-fold cross validation. After parameter tuning, the optimal values of parameters determined are shown in Table 4 and the evaluation results of RF are displayed in Table 5 and the heat map of confusion matrix is displayed in Figure 2. It is obvious that RF performs better after parameter tuning. Moreover, comparing results of RF with best results[5] of previous works[5][8][23], which are also shown in Table 5, RF has higher accuracy, precision, recall and f1-score, which means this research is really meaningful.

Parameters	Values
Criterion	Gini
Max_features	100
Min_samples_leaf	1
Min_samples_split	2
N_estimators	170

Table 4: Optimal Values of Random Forest.

Evaluation	RF(this passage)	ESALOR[5]
Accuracy	0.946	0.762
Precision	0.997	0.77
Recall	0.894	0.77
F1-score	0.943	0.86

Table 5: Evaluation Results of RF with Optimal Parameters.

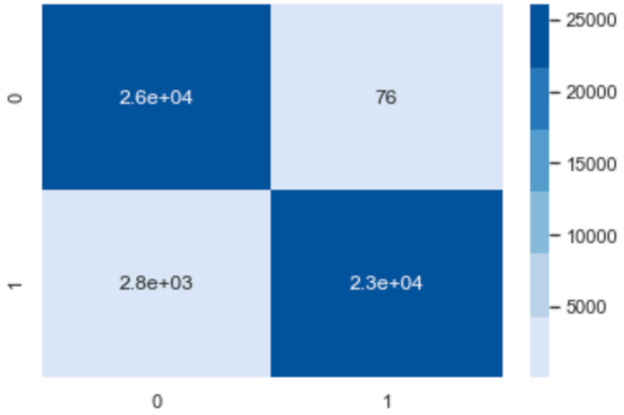

Figure 2: Heat map of confusion matrix of RF.

After obtaining the optimal parameters of RF, importance of each features are identified and top-20 most important features and how important they are are shown in Figure 3.

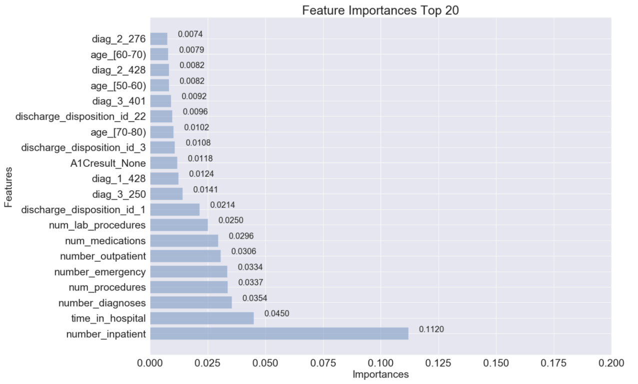

Figure 3: Top 20 most important features determined by RF.

6 CONCLUSION

Health care has drawn more and more attention all over the world these days. In U.S., a program called Hospital Readmissions Reduction Program (HRRP) are trying to improve Americans' health care by penalizing health systems with higher readmission rates than expected. Thus health systems are supposed to decrease their readmission rates as much as possible. However, due to the complexity and diversity of demographic, social, diseases and disease-related characteristics, it is difficult to decrease readmission rates merely based on clinical expertise. Therefore, this research aims to predict diabetic patients with high-risk readmissions via machine learning methods and help health systems identify readmission patterns. After data pre-processing, several models are trained, including Logistic Regression, Decision Tree, Support Vector Machine, Naive Bayes, K-Nearest Neighbour, Random Forest and Gradient Boosting Decision Tree. However, all these models perform badly. Thus a method called Borderline-SMOTE is used to rebalance the data and the models are trained again. The results show that this method really improves the ability of models. Then the performance of several models are compared, and Random Forest (RF) outperforms all the other models and is selected for further work. After parameter tuning, the accuracy of RF is 94.6%, precision is 99.7%, recall is 89.4%, and f1-score 94.3%, higher than all models including those in previous works. Finally, top 20 most important features are identified by RF.

Limitations of this research are: 1) values of possibly important features such as "***weight***" are missing; 2) how these features influence the readmissions of diabetic patients has not decided yet. Future work could be: 1) collecting more data and discovering whether these possibly important but missing features have influence on readmissions of diabetic patients; 2) determining whether these important features have positive influence or negative influence on readmissions of diabetic patients.

REFERENCES

[1] Yu, W., Liu, T., Valdez, R., Gwinn, M., & Khoury, M. J. (2010). Application of support vector machine modeling for prediction of common diseases: the case

of diabetes and pre-diabetes. BMC Medical Informatics and Decision Making, 10(1).doi:10.1186/1472-6947-10-16.

[2] Kemal Polat, Salih Gunes, Ahmet Arslan, A cascade learning system for classification of diabetes disease: Generalized Discriminant Analysis and Least Square Support Vector Machine, Expert Systems with Applications, Volume 34, Issue 1, 2008, Pages 482-487, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2006.09.012.

[3] Kevin Plis, Razvan Bunescu, Cindy Marling, Jay Shubrook, Frank Schwartz. A Machine Learning Approach to Predicting Blood Glucose Levels for Diabetes Management. Modern Artificial Intelligence for Health Analytics: Papers from the AAAI-14.2014.

[4] Zheng, T., Xie, W., Xu, L., He, X., Zhang, Y., You, M., … Chen, Y. (2017). A machine learning-based framework to identify type 2 diabetes through electronic health records. International Journal of Medical Informatics, 97, 120–127. doi: 10.1016/j.ijmedinf. 2016.09.014.

[5] Salih Tutun, Sina Khanmohanmmadi, Lu He and Chun-An Chou. (2016). A Meta-heuristic LASSO Model for Diabetic Readmission Prediction. Proceedings of the 2016 Industrial and Systems Engineering Research Conference.

[6] S. B. Kotsiantis. Supervised Machine Learning: A Review of Classification Techniques. 2007.Informatica 31 (2007). Pages 249-268.

[7] Abdullah A. Aljumah, Mohammed Gulam Ahamad, Mohammad Khubeb Siddiqui, Application of data mining: Diabetes health care in young and old patients, Journal of King Saud University - Computer and Information Sciences, Volume 25, Issue 2, 2013, Pages 127-136, ISSN 1319-1578, https://doi.org/10.1016/j.jksuci.2012.10.003.

[8] Saumya Salian, Dr. G. Harisekaran. 2013. Big Data Analytics Predicting Risk of Readmissions of Diabetic Patients. International Journal of Science and Research (IJSR), Volume 4, Issue 4, ISSN 2319-7064.

[9] V. Anuja Kumari, R. Chitra. 2013. Classification Of Diabetes Disease Using Support Vector Machine. International Journal of Engineering Research and Applications (IJERA), Vol. 3, Issue 2, March-April 2013, Pages 1797-1801, ISSN 2248-9622.

[10] Neesha Joshi, Nur'Aini Abdul Rashid, Wahidah Husain. Data Mining in Healthcare – A Review. 2015. Procedia Computer Science, Volume 72, Pages 306-313. https://doi.org/10.1016/j.procs.2015.12.145.

[11] Benbassat J, Taragin M. Hospital Readmissions as a Measure of Quality of Health Care: Advantages and Limitations. *Arch Intern Med.* 2000;160(8):1074–1081. doi:10.1001/archinte.160.8.1074.

[12] B.M. Patil, R.C. Joshi, Durga Toshniwal, Hybrid prediction model for Type-2 diabetic patients, Expert Systems with Applications, Volume 37, Issue 12, 2010, Pages 8102-8108, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2010.05.078.

[13] Ji Zhu, Trevor Hastie. Classification of gene microarrays by penalized logistic regression. 2004. Biostatistics, Volume 5, Issue 3, July 2004, Pages 427–443, https://doi.org/10.1093/biostatistics/kxg046.

[14] Malladihalli S Bhuvan, Ankit Kumar, Adil Zafar, Vinith Kishore. Identifying Diabetic Patients with High Risk of Readmission. 2016. Computer Science, Artificial Intelligence, arXiv:1602.04257.

[15] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, Noemie Elhadad. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. 2015. KDD '15 Proceddings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Pages 1721-2730, ISBN 978-1-4503-3664-2. doi>10.1145/2783258.2788613.

[16] Chao-Ying Joanne Peng , Kuk Lida Lee & Gary M. Ingersoll (2002) An Introduction to Logistic Regression Analysis and Reporting, The Journal of Educational Research, 96:1, 3-14.

[17] S.R. Safavian, D. Landgrebe. A Survey of Decision Wee Classifier Methodology. 1991. IEEE Transactions on Systems. Man, and Cybernetics, Volume 21, Issue3, Pages 660-674. DOI: 10.1109/21.97458.

[18] Stephan Dreiseitl, Lucila Ohno-Machado, Logistic regression and artificial neural network classification models: a methodology review, Journal of Biomedical Informatics, Volume 35, Issues 5–6, 2002, Pages 352-359, ISSN 1532-0464, https://doi.org/10.1016/S1532-0464(03)00034-0.

[19] Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, Dimitrios I. Fotiadis, Machine learning applications in cancer prognosis and prediction, Computational and Structural Biotechnology Journal, Volume 13, 2015, Pages 8-17, ISSN 2001-0370, https://doi.org/10.1016/j.csbj.2014.11.005.

[20] Frizzell JD, Liang L, Schulte PJ, et al. Prediction of 30-Day All-Cause Readmissions in Patients Hospitalized for Heart Failure: Comparison of Machine Learning and Other Statistical Approaches. *JAMA Cardiol.*2017;2(2):204–209. doi:10.1001/jamacardio.2016.3956.

[21] Bichen Zheng, Jinghe Zhang, Sang Won Yoon, Sarah S. Lam, Mohammad Khasawneh, Srikanth Poranki, Predictive modeling of hospital readmissions using metaheuristics and data mining, Expert Systems with Applications, Volume 42, Issue 20, 2015, Pages 7110-7120, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2015.04.066.

[22] Harleen Kaur, Vinita Kumari, Predictive modelling and analytics for diabetes using a machine learning approach, Applied Computing and Informatics, 2018, ISSN 2210-8327, https://doi.org/10.1016/j.aci.2018.12.004.

[23] Duggal, R., Shukla, S., Chandra, S. et al. Int J Diabetes Dev Ctries (2016) 36: 519. https://doi.org/10.1007/s13410-016-0511-8.

[24] Guiillaume Lematre and Fernando Nogueira and Christos K. Aridas. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. 2017. Journal of Machine Learning Research, Volume 18, Number 17, Pages 1-5.

[25] Mani, S., Chen, Y., Elasy, T., Clayton, W., & Denny, J. (). Type 2 diabetes risk forecasting from EMR data using machine learning. AMIA ... Annual Symposium proceedings. AMIA Symposium, 2012, 606–615.

[26] Han Wu, Shengqi Yang, Zhangqin Huang, Jian He, Xiaoyi Wang, Type 2 diabetes mellitus prediction model based on data mining, Informatics in Medicine Unlocked, Volume 10, 2018, Pages 100-107, ISSN 2352-9148, https://doi.org/10.1016/j.imu.2017.12.006.

[27] Sholom M. Weiss and Ioannis Kapouleas. An Empirical Comparison of Pattern Recognition, Neural Nets, and Machine Learning Classification Methods. An Empirical Comparison of Pattern Recognition. Pages 781-787.

[28] Jencks SF, Williams MV, Coleman EA. Rehospitalizations among patients in the Medicare fee-for-service program. N Engl J Med. 2009;360(14):1418–28.

[29] CentersforDiseaseControlandPrevention(CDC), 2011, "National Diabetes Fact Sheet: National Estimates and General Information on Diabetes and Prediabetes in the United States, " retrieved from http://www.familydocs.org/f/CDC

[30] Rachel Nall RN MSN. An overview of diabetes tpyes and treatments. https://www.medicalnewstoday.com/articles/323627.php.

[31] https://www.medicalnewstoday.com/articles/318472.php.

[32] https://www.healthline.com/health/diabetes/effects-on-body#2.

[33] http://www.diabetes.org/diabetes-basics/statistics/.

[34] Rubin, D.J. Hospital Readmission of Patients with Diabetes (2015) 15 - 17. https://doi.org/10.1007/s11892-015-0584-7.

[35] M. Lichman. UCI machine learning repository, 2013.

[36] 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," BioMed Research International, vol. 2014, Article ID 781670, 11 pages, 2014.

[37] H. Han, W.-Y. Wang, B.-H. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," In Proceedings of the 1st International Conference on Intelligent Computing, pp. 878-887, 2005.

[38] Pedregosa, F. and Varoquaux, G. etc. Scikit-learn: Machine Learning in Python. 2011. Journal of Machine Learning Research, Volume 12, Pages 2825-2830.