Data698 - Project Proposal

Hospital Diabetic Cases Readmission

Monu Chacko, Md Forhad Akbar, Abdelmalek Hajjam, Shovan Biswas

10/21/2020

- Team Members
- Overview and Motivation
- Goal and Modeling
- Data Source
- References

Team Members

- Monu Chacko
- Md Forhad Akbar
- Abdelmalek Hajjam
- Shovan Biswas

Introduction and Overview

Hospitalizations account for almost one-third of the total health care spending in the United States. A substantial portion of those hospitalizations are readmissions, which is why the Hospital Readmissions Reduction Program (HRRP) was created by CMS with the goal of improving the quality of care for patients and reducing healthcare spending.

Hospital readmissions within 30 days of discharge (30-d readmissions) have become a high-priority healthcare quality measure and target for cost reduction.

The Centers for Medicare and Medicaid Services (CMS) now penalizes hospitals for excess readmissions rates for certain conditions, with up to a three percent payment reduction. For many health systems, this can translate into millions of dollars. In fact, just one or two excess readmissions in populations such as diabetes, Heart Failure and Knee Replacements can tip a hospital over its allowable readmission rate.

The most common reasons for readmission according to primary discharge diagnoses were diabetes, heart failure, procedural complications, chest pain, shortness of breath, acute kidney failure, and urinary tract infection.

Diabetes is the seventh leading cause of death in the United States. Diabetes is the condition with the 3rd most all-cause, 30-day readmissions for Medicaid patients, and in 2011, American hospitals spent over $41 billion caring for diabetic patients who were readmitted within 30 days of discharge [7]. Readmission of diabetes patients can usually be avoided if additional attention is paid to these patients with high readmission risk and appropriate actions are taken. This makes early prediction of the hospital readmission risk an important problem. Being able to predict which patients will be readmitted could help save hospitals billions of dollars while also improving quality of care [7].

Goal and Modeling

The goal of this project is to see how well we can predict 30-day hospital readmission of diabetes patients, i.e. classify and identify patients that are at risk of being readmitted after being discharged. We should be able to predict whether the patient will be readmitted to the hospital or

not. For predicting it most accurately we will be using various prediction models for this purpose.

Because this is a classification problem, to predict whether a patient will be readmitted or not, we will be using six different classifiers, namely, Support Vector Machines, Generalized logistic regression, Artificial Neural Networks, Random Forest Classifier, Naïve Bayes Classifier and Decision Trees. The models will also be used to help determine what factors are the most important in predicting hospital readmission for diabetic patients.

After building, training, and testing the model, our models should be able to classify patients (yes/no) being readmitted within 30 days.

Data Source

The dataset represents 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks [7]. It consists of 10,000 records and 52 features representing patient and hospital outcomes. This dataset is made publicly available by the UCI Machine Learning Repository.

The data contains such attributes as patient number, race, gender, age, admission type, time in hospital, medical specialty of admitting physician, number of lab test performed, HbA1c test result [7], diagnosis, number of medications, diabetic medications, number of outpatients, inpatient, and emergency visits in the year before the hospitalization, etc.

This dataset will undergo intense cleaning procedures to be ready for modeling. Feature selection for these models will be done by conducting Correlation Analysis, and eliminating features with class imbalance. Data will be converted and pre-processed by removing null values and changing categorical variables.

Literature Review

Before embarking upon our research, we reviewed the literature and got a sense of the history of what has been done in the past. Reducing hospital readmissions continues to be a high-priority task across the healthcare systems, because unnecessary hospitalizations not only expose the patients to potential infections (especially in times of COVID19 like this), but also increase costs.

Since we are doing this research for a Data Science project, we like to explore the applicability of machine learning and other data science techniques. So, we checked the literature and found significant work have been done in that regard, especially in the area of readmissions for diabetes. Extensive study of HbA1C tests, as predictor for readmissions within 30-day period has been performed. Almost 70,000 encounters were considered with about 50 features. Algorithms like Logistic Regression, Decision Tree, Random Forest, Adaboost and XGBoost were deployed in such studies [7]. While readmission rates remained the highest for circulatory diagnoses, readmission rates for patients with diabetes appeared to be associated with the decision to test for HbA1c, rather than values of its result.

More than $25 billion are spent each year on diabetic readmissions alone. Sarthak et al evaluated existing models and proposed new embedding based on deep and neural network (DNN), which can predict whether a hospitalized diabetic patient would be readmitted within 30-days or not, with an accuracy of 95.2% and AUROC of 97.4%, based on data collected from 130 US hospitals between 1999-2008 [4]. The results have been encouraging with patients having changes in medications. Identifying prospective patients for readmission could reduce readmission costs. Every time a patient is admitted to a hospital, creatinine level is tested. Ben-

Assuli et al used multivariate logistic regression model on this indicator (creatinine) from Electronic Health Records of 5,103 patients [1]. Their findings suggest that three significant components impacting readmission are age, gender and creatinine levels.

Different researchers approached with different combinations of machine learning. Joshi and Alehegn (2017) reviewed various data mining techniques, which are generally used to predict chronic diseases. Classification techniques of machine learning, e.g. Support Vector Machine, Naïve Bayes and Decision Tree work well in generating better diagnoses predictions for diabetes and other diseases [11]. In order to classify the risk of diabetes, Nongyao et al [12] applied four machine learning classification methods namely Decision Tree, Artificial Neural Networks, Logistic Regression and Naïve Bayes. They improved the model with Bagging and Boosting techniques. Although Random Forest algorithm gives optimum results among all the machine learning algorithms they used, Hammoudeh et al [5] used a balanced combination of Convolutional Neural Networks and data engineering against real life data, to predict readmissions in early stages and thereby reduce costs and preserve reputation of the hospitals.

Patient wellness score integrates many lifestyle components and a holistic patient prospective. Agarwal et al used a large comprehensive survey of over 5000 people, conducted by the CDC, to build machine learning models. 8 out of the 9 models were shown to have a statistically significant ($p = 0.05$) increase in area under the receiver operating characteristic when using the hybrid approach when compared to expert-only models [2]. We also reviewed some literature on use of big data to evaluate risk of readmissions for diabetes patients. A paper [8] by Salian recommends a prescient model that can discover the hazard indicators on patients with perpetual diabetes. The data was analyzed on Hadoop cluster. Weight, plasma, glucose, age,

pregnancy, family work was found to be some of the top indicators of readmission for diabetic patients.

As of 2012, according to American Diabetic Association, more than 29 million Americans were diabetic. S. Behara's paper compared multilayer perception (MPL) and Bayesian networks (BNs) in diabetic patient classification on a cohort of of 5000 individuals, surveyed by the CDC. This paper [3] showed that MLP models predict with larger AUC, higher accuracy and lower RMSE.

Among the few other papers, we consulted were by Goudjerkan [6], where they used Random Forest for feature selection and SMOTE algorithm for data balancing. They also proposed MLP on data collected from 130 American hospitals. The proposed combination of data engineering and MLP was found to outperform existing practices. Two other papers by Kumar et al [9] used predictive analysis on in Hadoop and Baechle et al [10] used Naïve Bayes.

Methodology

Data

The dataset we will be using was obtained from the UCI Machine Learning Repository. It represents 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks [7]. It consists of 100000 observations and more than 50 features. The response variable is a binary outcome named "readmitted" (0 or 1), referring to whether or not the patient was readmitted within 30 days after his/her visit Fig 1. We have a classification problem in hand.

Many steps were taken to clean this dataset such as removing features that were irrelevant, redundant or containing too many missing values. We combined other variables under one variable as

seen necessary. After our final features were settled, correlation matrix was built to verify variables dependency. After that a dimension reduction algorithm (PCA) was run and showed that our 2 readmission classes were not well separated fig 2. The **final dataset** consists of **70420 observations** and **42 variables** (41 independent, 1 response).

Our Data seems to be highly imbalanced; 64129 cases (91%) were **NOT** readmitted and 6291 cases (9%) were readmitted. Algorithms in general tend to be bias toward the majority class (i.e. when one class dominates the other) when dealing with classification problems. To solve this problem, we will be using SMOTE (Synthetic Minority Over-sampling Technique) to smartly create new instances of the minority class, so the number of minority instance will be almost similar of the number of instances in the majority class (these are not duplicate). Our Data will be spitted in 2 proportions, one set for training (80%) and another set for testing (20%). SMOTE will only be applied to the training set.

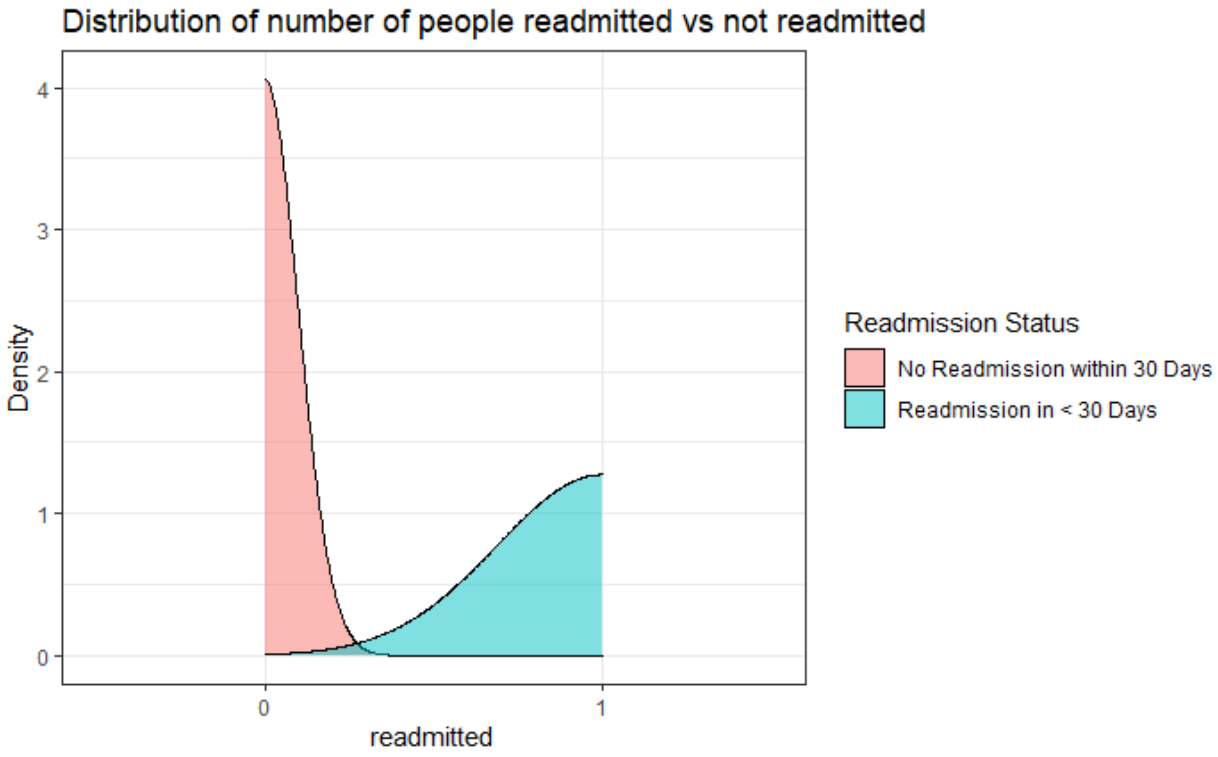


Fig 1

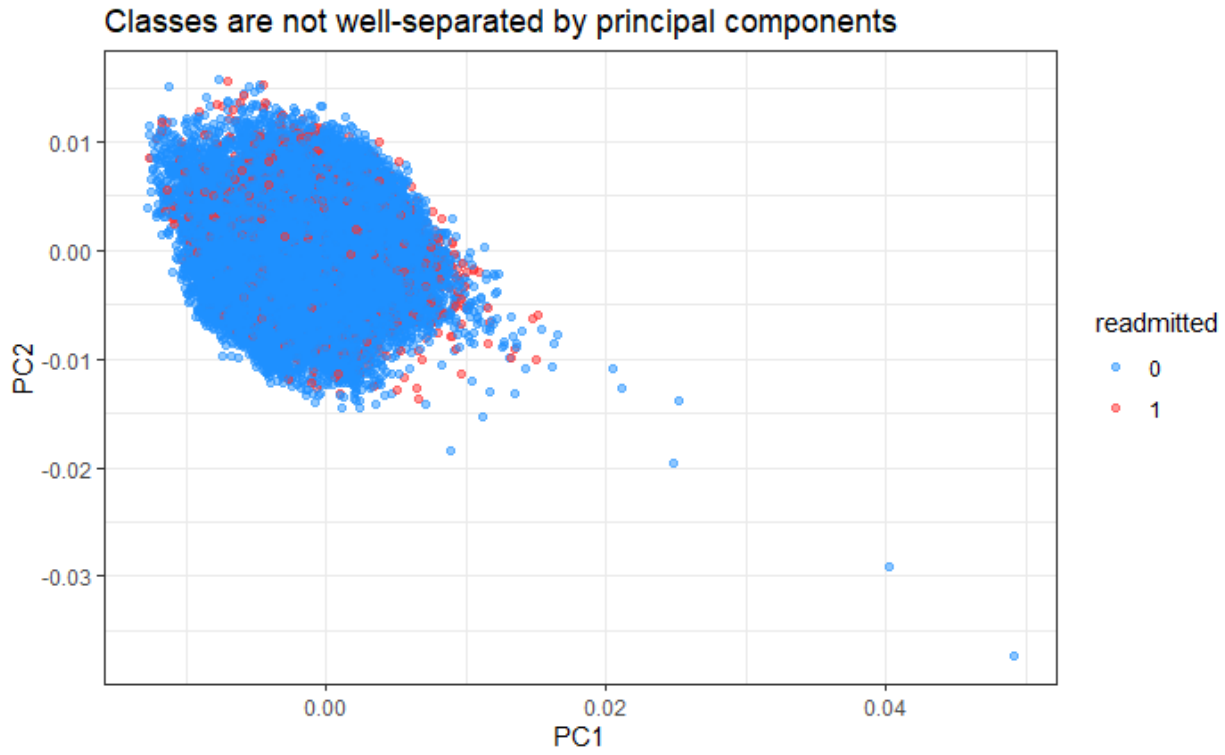
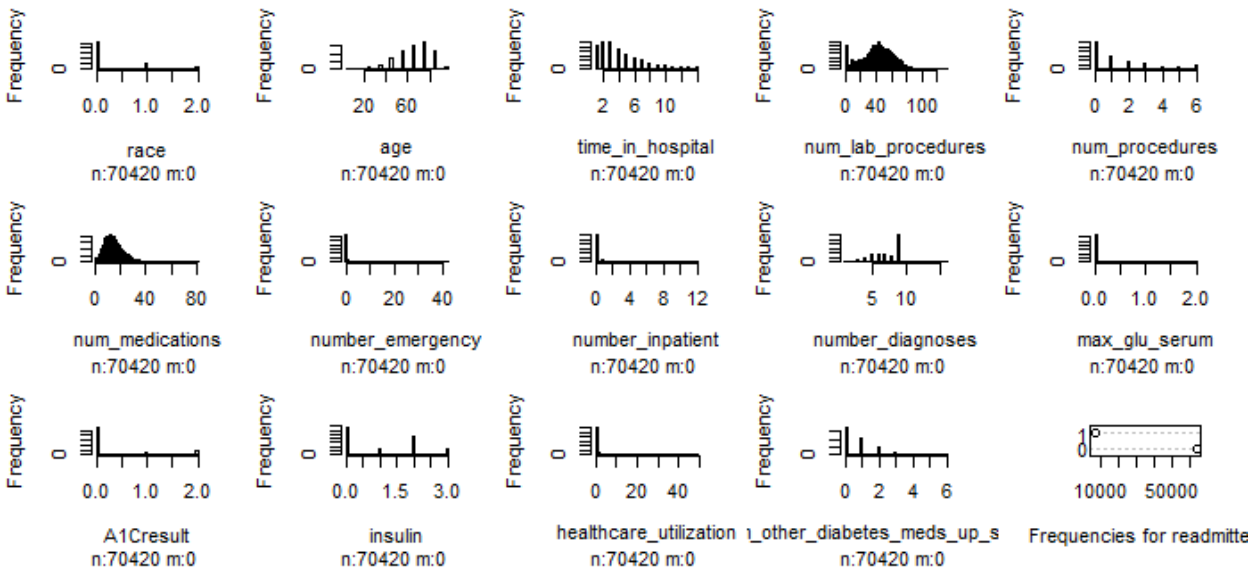


Fig 2

Distribution of few variables



Algorithms

This is a classification problem (1 or 0), to predict whether a patient will be readmitted or not, we will be using six different classifiers, namely, Support Vector Machines, Generalized logistic regression, Decision Trees, Random Forest Classifier, Naïve Bayes Classifier and Artificial Neural Networks.

Finally, we will use the latest algorithms in machine learning called **autoML**, an algorithm that runs many models in parallel using many trials while searching for the best hyper parameters for every algorithm that returns the best accuracy. This will be accomplished through **h2o** framework (locally), as well Azure ML (on the cloud).

We will compare all models to decide for the best that gives the best accuracy.

References

[1] O. Ben-Assuli, R. Padman, M. Leshno, and I. Shabtai, “Analyzing Hospital Readmissions Using Creatinine Results for Patients with Many Visits,” *Procedia Computer Science*, 21-Sep-2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050916321998?via=ihub. [Accessed: 21-Oct-2020].

[2] A. Agarwal, C. Baechle, R. S. Behara, and V. Rao, “Multi-method approach to wellness predictive modeling,” *Journal of Big Data*, 01-Jan-1970. [Online]. Available: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-016-0049-0. [Accessed: 21-Oct-2020].

[3] R. S. Behara, A. Agarwal, V. Rao, and C. Baechle, “Predicting the occurrence of diabetes using analytics.” [Online]. Available: http://faculty.eng.fau.edu/ankur/files/2017/08/Predicting-the-Occurence-of-Diabetes-Using-Analytics.pdf. [Accessed: 21-Oct-2020].

[4] Sarthak, S. Shukla, and S. P. Tripathi, “EmbPred30: Assessing 30-Days Readmission for Diabetic Patients Using Categorical Embeddings,” *Smart Innovations in Communication and Computational Sciences Advances in Intelligent Systems and Computing*, pp. 81–90, 2020. [Online]. Available: https://arxiv.org/pdf/2002.11215.pdf. [Accessed: 21-Oct-2020].

[5] A. Hammoudeh, G. Al-Naymat, I. Ghannam, and N. Obied, “Predicting Hospital Readmission among Diabetics using Deep Learning,” *Procedia Computer Science*, vol. 141, pp. 484–489, 2018. [Online]. Available: https://switchpointventures.com/wp-content/uploads/2018.11_predicting-hospital-readmission-among-diabetics-using-deep-learning.pdf. [Accessed: 21-Oct-2020].

[6] T. Goudjerkan and M. Jayabalan, “Predicting 30-day hospital readmission for diabetes patients using multilayer perceptron. ” [Online]. Available: https://thesai.org/Downloads/Volume10No2/Paper_36-Predicting_30_Day_Hospital_Readmission_for_Diabetes_Patients.pdf. [Accessed: 21-Oct-2020].

[7] B. Strack, J. P. DeShazo, C. Gennings, J. L. Olmo, S. Ventura, K. J. Cios, and J. N. Clore, “Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records,” *BioMed Research International*, 03-Apr-2014. [Online]. Available: https://www.hindawi.com/journals/bmri/2014/781670/. [Accessed: 21-Oct-2020].

[8] S. Salian and D. G. Harisekaran, “Big Data Analytics Predicting Risk of Readmissions of Diabetic Patients,” *International Journal of Science and Research (IJSR)*, Apr. 2015. [Online]. Available: https://www.ijsr.net/archive/v4i4/SUB152923.pdf. [Accessed: 21-Oct-2020].

[9] N. M. S. kumar, T. Eswari, P. Sampath, and S. Lavanya, “Predictive Methodology for Diabetic Data Analysis in Big Data,” *Procedia Computer Science*, 08-May-2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050915005700?via=ihub. [Accessed: 21-Oct-2020].

[10] C. Baechle and A. Agarwal, “A framework for the estimation and reduction of hospital readmission penalties using predictive analytics,” *Researchgate*, Dec-2017. [Online]. Available: https://www.researchgate.net/publication/320818442_A_framework_for_the_estimation_and_reduction_of_hospital_readmission_penalties_using_predictive_analytics. [Accessed: 21-Oct-2020].

[11] R. Joshi and M. Alehegn, “Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach,” *International Research Journal of Engineering and Technology (IRJET)*, Oct. 2017. [Online]. Available: https://d1wqtxts1xzle7.cloudfront.net/54976577/IRJET-V4I1077.pdf?1510396014=&response-content-disposition=inline%3B+filename%3DAnalysis_and_prediction_of_diabetes_dise.pdf&Expires=1603304802&Signature=eA5hHU8cZgbcv1Lg43~IXVLeIkNTfTXtfbDrGeRM6x7pQ9PH3viAw5AE7sSaV2hepMzkJtt3uIGLnWyr8aqKEv9VQFu0oNExGmRCIelVVFOp36UYB~1Kp-czNjY1k25X6eses846ByctxiHrgwOdqBL4ZGDFPE6xjOWWqBc3lmR69vNBRNUQGeaHy3Jm~12yMzZnsebERsLqWRdIs-6Uc87w-iOUapLKNRcI0OuA65xgXKvyUDuHAV8yhCGRGhkNRB6iYgjRo-JJwZP0YQXize2Yt7XR7~fYxOtI6LjrIy-Zth4ZdYsxpeBasIbybcgosi32~WSQMbyZBjKdLNXLKQ__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA].

[12] N. Nai-Arun and R. Moungmai, “Comparison of Classifiers for the Risk of Diabetes Prediction,” *Procedia Computer Science*, 14-Nov-2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050915031786. [Accessed: 21-Oct-2020].