Spring 3-27-2019

Predicting Hospital Readmissions of Diabetic patients - A Machine Learning Approach

Giridhar Reddy Bojja
Dakota State University

Omar El-Gayar
Dakota State University

Predicting Hospital Readmissions of Diabetic patients - A Machine Learning Approach

Giridhar Reddy Bojja, M.S, Omar El-Gayar, Ph.D
Giridhar.Bojja@trojans.dsu.edu , Omar.el-gayar@dsu.edu
College of Business and Information Systems - Dakota State University

Abstract

- Hospital readmission is an indicator of the quality of care and is a driver for the increasing cost of healthcare. Like other chronic diseases, Diabetes is associated with a higher risk of hospital readmission. In this research, we evaluate several machine learning approaches to predict the probability of hospital re-admissions for diabetic patients. The data set used for this study contains more than 100,000 diabetic patient data and 55 variables including length of stay, insulin, and in-patient visits from hospitals in the United States. We leverage several pre-processing techniques and investigate the performance of the various models. The significant variables contributing to the analysis are the number of in-patients, length of stay, number of medications, number of diagnoses, and age. The results demonstrate the viability of the techniques in providing a better understanding of factors influencing hospital re-admission.

Background

- A hospital readmission is an episode when a patient who had been discharged from a hospital is admitted again within a specified time interval. It is an indicator of the quality of care and could be a major driver for healthcare cost.
- According to survey by Agency for Healthcare Research and Quality (AHRQ) in 2011, it was found that more 3 million patients were readmitted within 30 days from discharge date. In 2012, there were 23,700 cases of re-admissions due to unchecked diabetes alone costing around $251 million.
- Although, identifying patients who are expected to be readmitted in 30 days of discharge is a complex task for hospitals. Techniques that can help to predict the likelihood for readmission and to identify the factors contributing to readmission can be of significant value to healthcare providers. Specifically, such techniques, allows provider to optimize their interventions for high-risk patients and to ultimately reduce readmission rates via improved processes.

Literature Review

- Rubin (2015) reports that patients with diabetes have a high risk of readmission than those without diabetes. Eby et al. (2015) use logistic regression and identify inpatient visits, race, diabetes prior to stay, and heart problem as strong predictors but they did not account the class imbalance problem. Munnangi and Chakraborty (2015) use a support vector machine and found inpatient and outpatient visits, primary diagnosis, mode of admission, patient condition were key factors in their analysis. In another study, Duggal et al., (2016) demonstrate that decision trees out-perform logistic regression and nave Bayes in predicting the readmission of diabetic patients. Contributing features include readmission department, length of stay, patient's medical history but they havent addressed the class imbalance issue. However, Hempstalk and Mordaunt (2016) report better results using logistic regression over nave Bayes and decision trees. In another study, Bhuvan et al., (2016) compares nave Bayes, random forest, adaboost and neural networks and demonstrate improved prediction results using random forest. Relevant features include a number of inpatient visits, discharge disposition ID, admission source, number of diagnosis as strong predictors but did not address class imbalance problem.
- While prior research demonstrates the viability of machine learning in predicting hospital readmission for diabetes patients, none of the studies accounts for the class imbalance problem inherent in the underlying data set. Further, there is considerable variability in the performance of the techniques and the resultant set of relevant features. In a more recent study, Dagliati et al., (2018) attempts to account for the class imbalance problem and demonstrates that logistic regression outperformed random forest and SVM. With a focus on type-2 diabetes, significant features include gender, age, time of diagnosis, BMI, HbA1c.
- From the literature, previous studies have produced mixed results and did not address the class imbalance problem. While Dagliati et al. (2018) attempted to account for such problem, the study is limited to patients with type 2 diabetes with a considerably smaller data set. This study addresses the class imbalance problem while using a significantly larger data set that includes patients with different types of diabetes. Further, the study evaluates a larger set of machine learning techniques including a number that have not been considered in prior research. The study leverages several pre-processing techniques, conduct feature analysis and investigate various model performances.

Methodology

- The dataset used for this study is obtained from UCI Machine Learning Repository and contains more than 100,000 diabetic patient data. As we are dealing with early re-admissions, we focus on less than 30 days readmission. Accordingly, we re-label the output variable 'readmitted' as '1' the number of days to readmission is less than 30 days, '0' otherwise. We retained unique records and removed duplicates. Discharge dispositions Id's which are related to deaths are deleted.

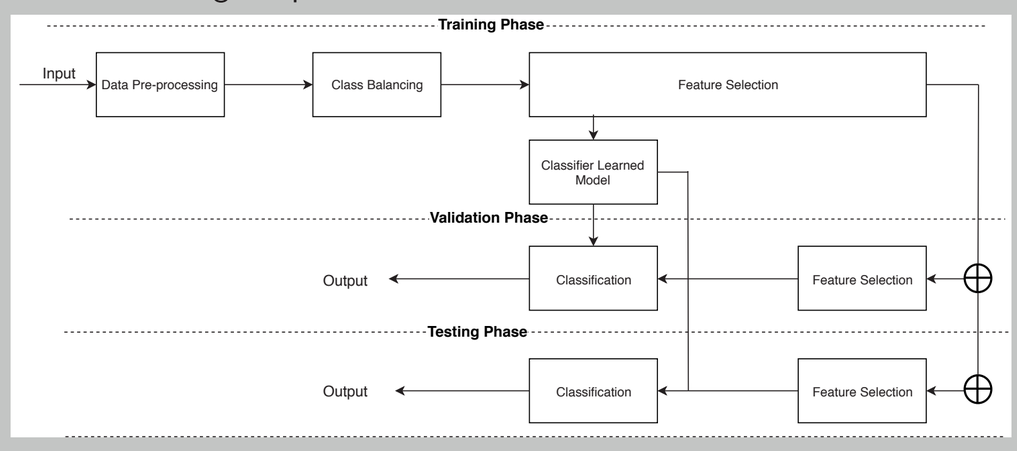

Figure 1: Methodology

- We performed feature engineering and separated data into 3 features numerical, categorical and others. Based on information from (Beata el al., 2014 we define diagnosis into nine different categories. We perform missing value analysis for each feature. We employ one-hot encoding technique to encode the categorical variables. We perform missing value imputation adding dummy variables. Overall, we end up having total of 71 features for our analysis.
- We split the data into Train, Validation and Test datasets using the percentages 70, 15, and 15, respectively. We use random sampling to address the data imbalance problem and perform feature analysis to know important features for our study.
- Using Python-Sklearn, we evaluate a number of ML techniques for this study. We hyper-tune the algorithms using grid search and randomized search, where a list of parameter values are tested using cross-validation technique to determine best fit. We calculate the Area under the Curve (AUC) and use it as a performance metric for evaluating the models. 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 in less than 30 days were only 11%.

Results and Discussion

- Our preliminary results show LightGBM performs slightly better than the other techniques (Table 1). The results of Microsoft's LightGBM for precision, recall and specificity are 0.116, 0.577, 0.575 respectively. While the results are in line with other published results by Dagliati et al. (2018) that accounted for the class imbalance problem, LightGBM, AdaBoost and Random Forest slightly out-performed Logistic Regression reported in (Dagliati et al., 2018).

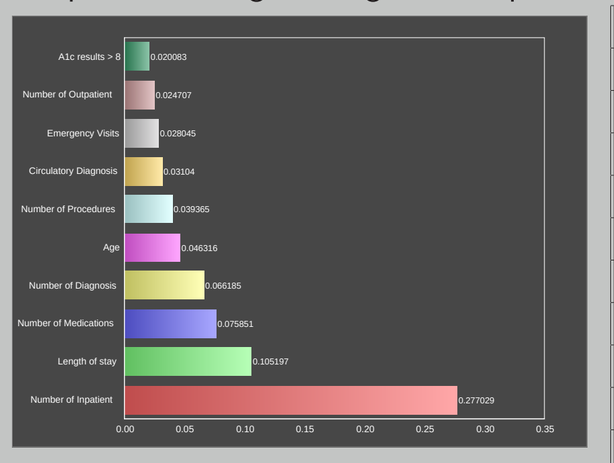

Figure 2: Feature importance by random forests

Models	AUC
Logistic Regression	0.603
Stocastic Gradient Descent	0.603
Gradient Boosting	0.599
Random Forest	0.606
SVM	0.600
Naive Bayes	0.577
Decision Trees	0.601
AdaBoost	0.617
CatBoost	0.599
LightGBM	0.620

Table 1: Performance of models

- Tree-based models calculate feature importance by keeping the best performing features as close to the root of the tree. Given an importance score to each feature the larger the score the most important the feature. Number of in-patient visits, Length of stay, Number of Medications, Number of Diagnoses, and Age were top fivein our analysis (Figure 2). Our results show age, length of stay, A1c are consistent with Dagliati et al. (2018). Whereas, number of diagnosis, number of inpatient visits are consistent with Bhuvan et al. (2016). Length of stay was consistent with (Duggal et al., (2016). The Number of medications and Circulatory diagnoses are unique findings in our study.

Conclusion

- In this study, we evaluated various machine learning models to predict readmissions of high-risk patients. Extending prior research, we performed class balancing considering the skewness of data. Our results show LightGBM slightly out-performing logistic regression and decision trees which were widely adopted in the literature. Some of the key features that drove readmissions are number of in-patients, length of stay, number of medications and number of diagnosis. Extending this research, we plan to further investigate the performance of classifiers with the goal to improve the accuracy of prediction of readmission risk. Further research is also warranted to explore the relevant feature space particularly with the variability of findings across research studies. The latter is particularly important as it can translate to proactive processes and policies aimed at addressing the factors that significantly influence hospital readmissions. Given the increasing cost of hospital readmissions and the increased emphasis on quality of care, the accuracy and validity of prediction models remain an important, yet elusive goal.