Diabetic Patients' Readmission Prediction

Multi-Class Classification on an Imbalanced Dataset

By Umer Farooq, Enid Roman, and Sangeetha Sasikumar

Data 621, Business Analytics and Data Mining

Abstract

The "Diabetes 130-US hospitals for years 1999-2008" dataset addresses the pressing issue of high 30-day readmission rates in diabetic patients, a chronic condition affecting 9.3% of the U.S. population. Focusing on factors contributing to readmissions, the study aims to predict high-risk patients for targeted intervention, enhancing care quality, patient experience, and overall population health while curbing costs. Leveraging comprehensive patient data spanning nearly a decade, the methodology employs data analysis, machine learning, and statistical modeling to uncover key predictors and develop a predictive model for high-risk patients. Additionally, the research explores the efficiency of diabetes medications, offering valuable insights into treatment strategies. Anticipated outcomes include the identification of significant risk factors, enabling targeted interventions and a predictive model that reduces readmission rates and associated costs. Insights into the medication's effectiveness will guide treatment decisions, contributing to overall patient well-being. The study aligns with healthcare industry priorities, emphasizing quality indicators, cost reduction, and patient-centered care. Ultimately, this research seeks to make a meaningful impact by improving the management and outcomes of diabetic patients, addressing a critical aspect of healthcare quality and cost containment

Keywords Clinical trials, Glucose, Metformin, Multinomial Logistic Regression, Random Forest, Readmission rate.

Introduction

The background of the problem lies in the prevalence and impact of Diabetes Mellitus (DM), a chronic condition with heightened blood sugar levels that can lead to severe health complications and diminish the quality of life. As highlighted by the World Health Organization (WHO), DM results frominadequate insulin production or inefficient insulin use by the body. A research article underscores the substantial societal burden, estimating that 9.3% of the U.S. population has DM, with 28% undiagnosed. Of particular concern is the high 30-day readmission rate for hospitalized diabetic patients, ranging between 14.4% and 22.7%, surpassing rates for all hospitalized individuals. This issue has prompted increased attention from government agencies and healthcare systems, emphasizing 30-day

readmission rates as a crucial indicator of patient complexity and care quality. The problem statement centers on identifying the factors contributing to this elevated readmission rate and predicting high-risk diabetic patients to enhance care quality, patient experience, and population health while reducing costs. The impact on business is significant, as hospital readmissions constitute a major contributor to medical expenditures, and addressing diabetes-related readmissions aligns with healthcare quality measures and cost reduction targets. By providing insights into risk factors and effective medications for diabetes, this research seeks to contribute to improved patient outcomes, reduced healthcare costs, and enhanced overall healthcare quality.

Literature Review

When it comes to predicting early hospital readmissions, several studies have made notable contributions, particularly in non-diabetic patient cohorts. Rubin (2015) highlighted the dearth of attention given to readmissions in diabetic patients, possibly due to the myriad severe side effects associated with the disease, complicating primary diagnosis during hospitalization. Futoma et al. (2015) conducted a comprehensive comparison of machine learning models (SVM, Random Forest, Logistic Regression, and Neural Network) for predicting early hospital readmissions across 280 Diagnosis Related Groups (DRGs). Building on this foundation, our investigation employed the same four baseline models and extended the exploration to include the efficacy of ensemble learning algorithms.

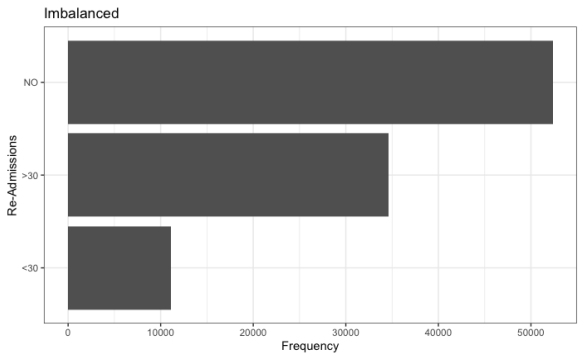
Frizzell et al. (2017) utilized a machine learning approach to predict 30-day all-cause readmissions following heart failure hospitalization, providing valuable insights despite not focusing on diabetic patients. Similarly, Golas et al. (2018) and Mortazavi et al. (2016) explored machine learning techniques for heart failure-related readmissions, offering relevant methodologies and experimental results. Our study, detailed in Section 5, surpassed existing literature by achieving an improved AUC score of 0.80, contrasting with scores ranging from 0.54 to 0.72 in related studies (Frizzell et al., 2017).

Duggal et al. (2016) specifically investigated early readmission risk prediction in diabetic patients, comparing classification models based on two-year longitudinal observations. In contrast, our approach distinguishes itself in methodology and data, focusing on clinical and demographic characteristics of individual hospitalization visits rather than two-year longitudinal observations. Our primary contributions lie in conducting an in-depth study of machine learning methods for predicting early readmission risk in diabetic patients. Moreover, we extended beyond conventional approaches by exploring ensemble learning methods and identifying top risk factors associated

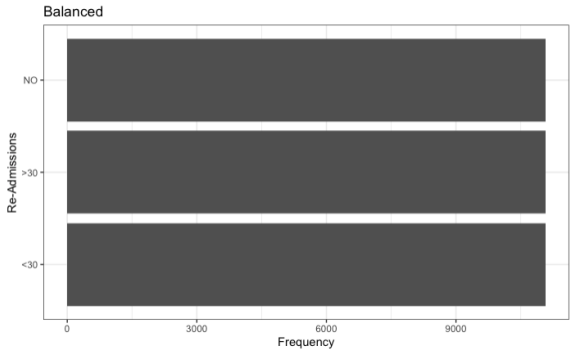
with early readmission in this specific patient data population.

Methodology:

We started our analysis with the loading and exploratory data analysis (EDA) of the dataset. Using the read_csv function, the data is imported, and missing values, represented by '?', are appropriately handled by replacing them with NA. We dropped the columns that had too many NA values, for example the “Weight” column. To address potential issues with duplicate entries, a check is conducted on the 'encounter_id' column, leading to the removal of any duplicated rows. This meticulous data cleaning process ensures a high-quality dataset for subsequent analyses. Following data cleansing, the script strategically addresses the challenge of class imbalance within the target variable ('readmitted'). The downSample function is used to down sample (rows are not duplicated but removed) the data, ensuring a balanced representation of classes that could influence model training. Figure below shows the target variable before the balance.



As we can see that there is a class Imbalance in the responses of our target variable. Here is the balanced bar graph after downsampling to the minimum responses.



Next, the dataset undergoes type conversion, with character columns transformed into factors (refer to data dictionary in appendix) to facilitate subsequent modeling steps. Multinomial logistic regression can handle both categorical and continuous independent variables. Further data exploration involves splitting the dataset into training and validation sets, leveraging functions such as split_columns to differentiate between continuous and discrete variables. Normalization of continuous variables and one-hot encoding of categorical variables are crucial steps in preparing the data for modeling. The script aptly scales continuous variables using the scale function and employs model.matrix for one-hot encoding, enhancing the compatibility of the dataset with machine learning algorithms.The focal point of the script revolves around the construction and evaluation of predictive models.

Once the dataset has been meticulously prepared, the model training phase begins. The dataset is split into training and testing sets to facilitate model evaluation. Multinomial logistic regression, k-Nearest Neighbors (KNN), and Random Forest classifiers are trained on the training set. Each model undergoes a tuning process for hyperparameters to optimize its performance. For instance, the number of neighbors in KNN or the depth of trees in Random Forest are adjusted. Before feeding the training data set to different algorithms a challenge was to determine which particular set of variables from the data set would be more statistically significant to impact the outcome of our models. A technique known as Forward Stepwise was carried out using Akaike Information Criterion (AIC) to select the predictors that lowered AIC. The downsampled dataset is also utilized for training models. The aim is to build models that capture the underlying patterns in the data, providing accurate predictions on unseen instances.

Model evaluation is a crucial step to assess the performance and generalization capabilities of trained models. Performance metrics such as accuracy, precision, recall, F1-score, and balanced accuracies are computed. The confusion matrix provides insights into the model's behavior, revealing how well it correctly classified instances for each class. Ensemble methods may be considered to combine predictions from multiple models, enhancing overall performance. The evaluation phase aims to not only measure the effectiveness of the models but also to identify areas for improvement. It serves as a critical feedback loop, guiding further iterations of model development and fine-tuning for enhanced predictive capabilities.

Experimentation and Results:

During the course of this exploration, we utilized the capabilities of Random Forest, k-Nearest Neighbors (KNN), and Multinomial Logistic Regression models to handle the issues presented by an imbalanced dataset. In the first stage, the data were carefully inspected and preprocessed. Missing values were strategically imputed, class imbalance was reduced by downsampling, and numerical features were uniformly scaled. The efficacy of Multinomial Logistic Regression was then assessed using conventional classification criteria, requiring meticulous hyperparameter adjustment and a thorough evaluation procedure. Subsequently, the KNN algorithm was utilized, focusing on determining the ideal number of neighbors and distance measures to achieve higher predicted accuracy. The robust Random Forest model was put through a rigorous training and assessment process.

In order to evaluate the effect on model performance, a downsampled version of the dataset was also added. This allowed for a comparison between

models trained on the original and downsampled datasets. The experimentation's outcomes not only shed light on the advantages and disadvantages of each model but also lay a solid basis for upcoming improvements that might include more sophisticated model structures or group techniques. All things considered, the goal of this all-encompassing strategy is to improve our comprehension of predictive modeling when it comes to imbalanced datasets and to direct future efforts toward developing a reliable and efficient solution. Below table summarize all the results from experimentation

Models	Sensitivity	Specificity	Pos.Pred.Value	Neg.Pred.Value	Precision	Recall	F1	Prevalence	Balanced.Accuracy
Multinomial (<30)	0.02	1.00	0.37	0.89	0.37	0.02	0.04	0.11	0.51
Multinomial (>30)	0.30	0.84	0.51	0.69	0.51	0.30	0.38	0.35	0.57
Multinomial (NO)	0.87	0.32	0.60	0.69	0.60	0.87	0.71	0.53	0.60
Reduced Multinomial (<30)	0.01	1.00	0.42	0.89	0.42	0.01	0.01	0.11	0.50
Reduced Multinomial (>30)	0.24	0.86	0.49	0.68	0.49	0.24	0.32	0.35	0.55
Reduced Multinomial (NO)	0.90	0.26	0.58	0.68	0.58	0.90	0.70	0.53	0.58
Down Sampled Multinomial (<30)	0.42	0.79	0.50	0.73	0.50	0.42	0.46	0.33	0.60
Down Sampled Multinomial (>30)	0.39	0.74	0.43	0.71	0.43	0.39	0.41	0.33	0.56
Down Sampled Multinomial (NO)	0.60	0.67	0.48	0.77	0.48	0.60	0.53	0.33	0.64
Down Sampled Reduced Multinomial (<30)	0.37	0.78	0.45	0.71	0.45	0.37	0.41	0.33	0.57
Down Sampled Reduced Multinomial (>30)	0.26	0.80	0.39	0.68	0.39	0.26	0.31	0.33	0.53
Down Sampled Reduced Multinomial (NO)	0.65	0.57	0.43	0.77	0.43	0.65	0.52	0.33	0.61
Multinomial Cross Validated (<30)	0.42	0.78	0.49	0.73	0.49	0.42	0.46	0.33	0.60
Multinomial Cross Validated (>30)	0.38	0.74	0.43	0.71	0.43	0.38	0.40	0.33	0.56
Multinomial Cross Validated (NO)	0.60	0.67	0.48	0.77	0.48	0.60	0.53	0.33	0.64
Multinomial with Coded Matrix (<30)	0.42	0.79	0.50	0.73	0.50	0.42	0.45	0.33	0.60
Multinomial with Coded Matrix (>30)	0.38	0.75	0.43	0.71	0.43	0.38	0.40	0.33	0.56
Multinomial with Coded Matrix (NO)	0.61	0.67	0.48	0.77	0.48	0.61	0.54	0.33	0.64
Random Forest (<30)	0.41	0.72	0.42	0.71	0.42	0.41	0.42	0.33	0.56
Random Forest (>30)	0.31	0.75	0.38	0.68	0.38	0.31	0.34	0.33	0.53
Random Forest (NO)	0.55	0.66	0.45	0.75	0.45	0.55	0.49	0.33	0.61
KNN (<30)	0.37	0.77	0.44	0.71	0.44	0.37	0.40	0.33	0.57
KNN (>30)	0.32	0.75	0.39	0.69	0.39	0.32	0.35	0.33	0.53
KNN (NO)	0.59	0.63	0.44	0.75	0.44	0.59	0.51	0.33	0.61
Support Vector Machines (<30)	0.35	0.84	0.52	0.72	0.52	0.35	0.41	0.33	0.59
Support Vector Machines (>30)	0.36	0.77	0.44	0.71	0.44	0.36	0.39	0.33	0.57
Support Vector Machines (NO)	0.70	0.59	0.46	0.80	0.46	0.70	0.55	0.33	0.64

Table 1: Classification Model Metrics

Discussion and Conclusions:

The model performance table that is displayed offers important information about the advantages and disadvantages of different classifiers that have been used on the dataset in various scenarios. It was consistently difficult for Multinomial Logistic Regression to correctly identify examples in the '<30' and '>30' categories when it comes to imbalance dataset, indicating possible limits in capturing intricate relationships within these groups. For the 'NO' category, on the other hand, the model demonstrated noteworthy sensitivity, Recall and balanced accuracy, highlighting its ability to accurately anticipate negative cases.

Both the Reduced Multinomial Logistic Regression and the Down Sampled Reduced Multinomial Logistic Regression demonstrated how feature reduction and downsampling affected model performance. Downsampling improved the '<30' category's sensitivity, albeit at the expense of specificity. Downsampling and feature reduction increased sensitivity even further, but at the sacrifice of specificity, highlighting the fine balance needed to handle unbalanced datasets.

In every category, Random Forest demonstrated balanced sensitivity and specificity, making it a strong performer. It is an attractive option for this classification task because of its capacity to manage skewed datasets. KNN performed competitively, highlighting its applicability in situations where sensitivity and specificity must be balanced.

When it came to the 'NO' category, Support Vector Machines (SVM) demonstrated a high sensitivity. However, they had trouble attaining a high sensitivity for the '<30' and '>30' categories. The SVM performance characteristics highlight how crucial it is to select models that are in line with the particular objectives and priorities of the classification task.

In conclusion, the experimentation provided a nuanced understanding of how different models respond to imbalanced datasets and various preprocessing techniques. The findings underscore the need for a thoughtful approach in selecting and configuring models based on the desired trade-offs between sensitivity and specificity. Future model development might involve fine-tuning hyperparameters, exploring ensemble methods, neural networks or considering additional feature engineering to further enhance predictive capabilities. The insights gained from this discussion contribute to a comprehensive strategy for building effective and reliable classification models in the context of imbalanced datasets.

References:

Frizzell, J. D., Liang, L., Schulte, P. J., Yancy, C. W., Heidenreich, P. A., Hernandez, A. F., Bhatt, D. L., Fonarow, G. C., and Laskey, W. K. (2017). Prediction of 30-day all-cause readmissions in patients hospital- ized for heart failure: comparison of machine learn- ing and other statistical approaches. *JAMA cardiology*, 2(2):204–209.

Golas, S. B., Shibahara, T., Agboola, S., Otaki, H., Sato, J., Nakae, T., Hisamitsu, T., Kojima, G., Felsted, J., Kakarmath, S., et al. (2018). A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of electronic medical records data. *BMC medical infor- matics and decision making*, 18(1):44.

Duggal, R., Shukla, S., Chandra, S., Shukla, B., and Kha-tri, S. K. (2016). Predictive risk modelling for early hospital readmission of patients with diabetes in in- dia. *International Journal of Diabetes in Developing Countries*, 36(4):519–528.

Futoma, J., Morris, J., and Lucas, J. (2015). A comparison of models for predicting early hospital readmissions. *Journal of biomedical informatics*, 56:229–238.

Mortazavi, B. J., Downing, N. S., Bucholz, E. M., Dhar- marajan, K., Manhapra, A., Li, S.-X., Negahban, S. N., and Krumholz, H. M. (2016). Analysis of ma- chine learning techniques for heart failure readmis- sions. *Circulation: Cardiovascular Quality and Out- comes*, pages CIRCOUTCOMES–116.

Rubin, D. J. (2015). Hospital readmission of patients with diabetes. *Current diabetes reports*, 15(4):17.

https://letsgethealthy.ca.gov/goals/redesigning-the-health-system/reducing-hospital-readmissions/#:~:text=The%20rate%20of%20unplanned%20hospital,are%20associated%20with%20high%20costs

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8418292/

https://my.clevelandclinic.org/health/diseases/7104-diabetes#symptoms-and-causes

Appendices:

- **Supplemental tables and/or figures.**

Data Dictionary
Data ID Mapping

- **R statistical programming code.**

The Code