

**Predictive model for**

**hospital readmissions**

by Jane Tran - Capstone project

Domain background

Hospital readmission for diabetic patients is a significant concern in the United States due to the increasing prevalence of diabetes and its associated complications. According to the American Diabetes Association, more than 30 million people in the U.S. have diabetes, and the estimated annual cost of diabetes in the country is $327 billion. Hospital readmission rates for diabetic patients are alarmingly high, with studies indicating that up to 22.7% of diabetic patients are readmitted to the hospital within 30 days of discharge, much higher than the rate for all hospitalised patients (8.5 - 13.5%) (Ostling et al., 2017). Furthermore, the Centers for Medicare & Medicaid Services (CMS) reported that in 2019, the national average readmission rate for diabetes-related complications was 18.9%, making it one of the leading causes of hospital readmissions. These statistics highlight the urgent need for effective interventions and strategies to reduce hospital readmissions and improve the quality of care for diabetic patients in the U.S. healthcare system. Additionally, reducing readmission rates for diabetic patients has a great potential to reduce medical cost significantly. Therefore, the objective of this study is to predict the likelihood of a diabetic patient being readmitted.

Business perspective – Business context, stakeholders and value The high rates of hospital readmission for diabetic patients in the U.S. represent both challenges and opportunities for healthcare providers, payers, and related stakeholders.

On one hand, the financial burden of hospital readmissions, estimated to be in the billions of dollars annually, poses significant cost challenges for hospitals and healthcare systems. In addition, penalties imposed by CMS for excessive readmissions under the Hospital Readmissions Reduction Program (HRRP) can impact hospital reimbursements, further impacting their bottom line.

On the other hand, addressing the issue of hospital readmissions for diabetic patients presents an opportunity for healthcare providers to implement innovative solutions that improve patient outcomes and reduce healthcare costs. This may include investing in care coordination, patient



education, remote monitoring, and other evidence-based interventions to prevent hospital readmissions and enhance patient care. Furthermore, there may be a demand for specialized services, technologies, and products that support diabetic patients' management of their condition, creating potential business opportunities for companies that can offer effective solutions in this space. Overall, addressing the issue of hospital readmissions for diabetic patients in the U.S. from a business perspective requires a proactive approach to implement cost-effective strategies that prioritize patient care and outcomes while managing financial implications.

This project aims to help a group of hospitals to better understand diabetic patient readmissions. The hospitals gave access to ten years of information on diabetic patients readmitted to the hospital after being discharged. The doctors want to assess if initial diagnoses, number of procedures, or other variables could help them better understand the probability of readmission.

Business questions:

How to identify patients with high probability of readmission based on initial diagnoses, number of procedures, and other variables?

Stakeholders

Stakeholders involved in predicting hospital readmission risk may include:

● **Healthcare providers:** Physicians, nurses, and other healthcare professionals who are responsible for patient care and management may be stakeholders in predicting hospital readmission risk. They may use predictive models and risk assessment tools to identify patients at higher risk of readmission, and develop interventions or care plans accordingly.

● **Hospital administrators:** Hospital administrators, including hospital managers, executives, and administrators, may have a stake in predicting hospital readmission risk. They may use predictive models and risk assessment tools to allocate resources, plan for patient discharge, and manage hospital capacity and utilization.

● **Payers and insurers:** Payers and insurers, such as health insurance companies, Medicare, and Medicaid, may be stakeholders in predicting hospital readmission risk. They may use predictive models and risk assessment tools to identify patients at higher risk of readmission and implement targeted interventions or reimbursement strategies to reduce readmission rates and manage costs.

● **Patients and families:** Patients and their families may have a stake in predicting hospital readmission risk as it affects their health outcomes and healthcare costs. They may be involved in shared decision-making and care planning based on the predicted risk of readmission.

● **Researchers and academics:** Researchers and academics in the field of healthcare and hospital readmission may be stakeholders in developing and validating predictive



models and risk assessment tools. Their research may inform the development of evidence-based interventions and policies related to hospital readmission risk prediction. ● **Policy makers and regulators:** Policy makers and regulators at the local, state, or national level may be stakeholders in predicting hospital readmission risk as they develop policies, guidelines, and regulations related to healthcare quality, patient safety, and reimbursement. Predictive models and risk assessment tools may inform their decision-making process.

● **Technology vendors and developers:** Vendors and developers of healthcare information technology solutions, including electronic health record (EHR) systems, predictive analytics software, and other technological tools, may have a stake in predicting hospital readmission risk. They may develop and provide tools, technologies or solutions that enable healthcare providers, administrators, payers, and other stakeholders to assess and manage hospital readmission risk more effectively.

Business value:

**Assumptions**

● Total number of diabetic patients in the U.S.: 30 millions (American Diabetes Association, 2022)

● The hospital readmission rate for U.S. diabetic patients: 18% (Ostling et al., 2017) ● The average cost of hospital readmission for a diabetic patient: US $10,000 (American Diabetes Association, 2018)

● This study aims to reduce the hospital readmission rate for U.S. diabetic patients by 9%

***Cost of hospital readmissions without the predictive model:***

*30,000,000 x 18% x $10,000 = $54 billions*

***Cost of hospital readmissions with the predictive model:***

*30,000,000 x 9% x $10,000 = $27 billions*

***Cost savings:***

*$54 billions - $27 billions = $27 billions*

***The projected cost savings for the healthcare sector are estimated to be $27 billion in the U.S.***

******

Data description, sources, quality

Data questions

● Does diabetes play a central role in readmission?

● On what groups of patients should the hospital focus their follow-up efforts to better monitor patients with a high probability of readmission?

Data information

The dataset represents 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. It includes 101,766 observations and 50 features representing patient and hospital outcomes. Information was extracted from the database for encounters that satisfied 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.

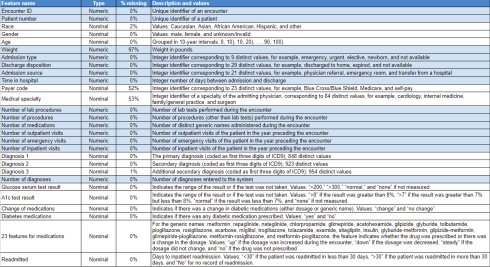
The data contains such attributes as patient number, race, gender, age, admission type, time in hospital, medical specialty of admitting physician, number of lab test performed, HbA1c test result, diagnosis, number of medication, diabetic medications, number of outpatient, inpatient, and emergency visits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

Source:

https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008

Information in the file





Data limitations

Here’s some limitations of the dataset:

● **Limited Scope:** The data set may only cover a specific time period (1999-2008) and may include data from a limited number of hospitals in the United States. This could limit the generalizability and applicability of the findings to other populations or time periods.

● **Data Quality:** The accuracy, completeness, and reliability of the data in the data set may vary, as it is dependent on the quality of data collection, documentation, and coding practices at the participating hospitals. There may be missing, inconsistent, or erroneous data that could introduce bias or impact the validity of the results.

● **Bias and Generalizability:** The data set may not be representative of the broader population or may have selection bias, as it may only include data from hospitals that voluntarily participated in the study. This could limit the generalizability of the findings to other settings or populations, and may not accurately reflect the true prevalence or characteristics of diabetes in the entire US population.

● **Lack of Longitudinal Data:** The data set may not have long-term follow-up data on individual patients, which could limit the ability to study the natural history of diabetes or track outcomes over time. Longitudinal data would be necessary to understand disease progression, treatment patterns, and long-term outcomes in a more comprehensive manner.



● **Ethical Considerations:** The data set may raise ethical concerns related to patient privacy, data security, and informed consent. It is important to ensure that the data used in the analysis are de-identified and comply with relevant privacy regulations to protect the rights and confidentiality of patients.

● **Confounding Variables:** The data set may not account for all potential confounding variables that could influence the relationship between variables of interest. Unmeasured or unknown confounding variables could impact the validity of the findings and limit the ability to draw causal inferences.

● **Limited Variables:** The data set may have limitations in terms of the variables available for analysis. It may not capture all relevant clinical, social, or environmental factors that could impact diabetes outcomes. This could limit the ability to conduct comprehensive analyses or explore specific research questions.

How the data can be sources in the future

Electronic health records (EHRs) are increasingly being adopted by healthcare systems, and they can provide a rich source of data for diabetes research. EHRs contain detailed clinical information on patients, including demographics, medical history, diagnoses, treatments, and outcomes. Researchers can collaborate with healthcare systems or organizations to access and analyze EHR data, while ensuring appropriate data privacy and security measures are in place. EHRs also allow for the collection of more up-to-date and comprehensive data, and can provide longitudinal data to track disease progression, treatment patterns, and outcomes over time.

Proposed solution, evaluation metrics and current benchmark (if any)

There are several reasons for choosing logistic regression, decision tree, and random forest models for predicting hospital readmission risk for diabetic patients using the Diabetes 130-US hospitals for years 1999-2008 Data Set:

**1. Logistic Regression:**

● Logistic regression is a widely used method for binary classification tasks, making it suitable for predicting hospital readmission risk (binary outcome of readmission or no readmission).

● It is simple and interpretable, which allows for easy understanding of the model's predictions and insights into the factors that influence hospital readmission risk. ● Logistic regression is efficient and can handle large datasets, making it suitable for the Diabetes 130-US hospitals dataset, which is a large dataset with multiple features. ● It provides probability estimates for the predicted outcomes, which can be useful in understanding the likelihood of hospital readmission for individual patients.

**2. Decision Tree:**

****

● Decision trees are non-linear models that can capture complex relationships and interactions between features, which may be present in the Diabetes 130-US hospitals dataset.

● Decision trees are interpretable and can provide insights into the decision-making process of the model, making it easier to explain and understand the results. ● Decision trees can handle both categorical and numerical features, making them suitable for datasets with a mix of data types, which is often the case in healthcare datasets.

**3. Random Forest:**

● Random Forest is an ensemble learning method that combines multiple decision trees, which can improve the model's performance and reduce overfitting.

● Random Forest can handle high-dimensional datasets with a large number of features, making it suitable for the Diabetes 130-US hospitals dataset, which has multiple features.

● Random Forest is known for its robustness and ability to handle noisy data, which can be prevalent in healthcare datasets.

● It can capture complex interactions and nonlinear relationships between features, which may be present in the data.

● Random Forest provides feature importance measures, which can help identify the most influential features for hospital readmission risk prediction.

Evaluation Metrics:

● **Accuracy:** The proportion of correct predictions out of the total predictions, which gives an overall measure of the model's accuracy.

● **Precision:** The proportion of true positive predictions (readmission) out of the total predicted positive cases (predicted readmissions), which measures the model's ability to correctly identify readmission cases.

● **Recall:** The proportion of true positive predictions (readmission) out of the total actual positive cases (actual readmissions), which measures the model's ability to capture all the readmission cases.

● **AUC-ROC:** Area under the Receiver Operating Characteristic (ROC) curve, which measures the model's ability to discriminate between readmission and non-readmission cases.

Benchmark:

The benchmark for these models can vary depending on the specific problem context and the dataset. However, a commonly used benchmark for hospital readmission risk prediction is



achieving an accuracy of around 70-80%, precision and recall of around 60-70%, and an AUC-ROC score of 0.64 or higher (Lu & Uddin, 2022). These benchmarks can serve as a reference for evaluating the performance of the logistic regression, decision tree, and random forest models on the Diabetes 130-US hospitals dataset.

Project design: What Data preprocessing methods are you thinking of using?

For the Diabetes 130-US hospitals for years 1999-2008 dataset, several data preprocessing methods can be employed to ensure the quality and reliability of the data for building logistic regression, decision tree, and random forest models. Some potential data preprocessing methods include:

Data Cleaning:

● **Handle Missing Values:** Identify and handle missing values in the dataset, which can involve techniques such as imputation (e.g., mean, median, mode imputation), deletion of rows or columns with missing values, or using advanced imputation methods (e.g., k-nearest neighbors imputation, regression imputation) depending on the nature and extent of missing data.

● **Remove Duplicates:** Check for and remove any duplicate records in the dataset, as duplicate data can introduce biases and affect model performance.

Feature Engineering:

● **Feature Selection:** Identify and select relevant features (i.e., columns) that are most informative for predicting hospital readmission risk. This can involve techniques such as univariate and multivariate feature selection methods (e.g., chi-squared test, feature importance from decision tree/random forest), domain knowledge, and feature correlation analysis.

● **Feature Encoding:** Encode categorical features (i.e., variables) into numerical representations suitable for modeling. Common methods include label encoding (e.g., assigning numeric codes to categories) and one-hot encoding (e.g., creating binary variables for each category).

● **Feature Scaling:** Scale numerical features to a common scale to prevent the domination of certain features due to their larger magnitudes. Common methods include min-max scaling (scaling features to a specific range) or standardization (scaling features to have zero mean and unit variance).

Data Transformation:



● **Handling Outliers:** Identify and handle outliers, which can involve techniques such as winsorizing (replacing extreme values with values at a specified percentile), log transformation, or robust scaling (scaling features based on their interquartile range) to mitigate the impact of outliers on model performance.

● **Handling Skewed Data:** If the dataset contains highly skewed features, such as features with a long tail distribution, applying data transformation techniques such as log transformation or power transformation (e.g., Box-Cox transformation) can help normalize the data and improve model performance.

Data Splitting:

**Splitting Data:** Split the dataset into training, and test sets to evaluate model performance. Common techniques include random sampling or time-based splitting (e.g., using a specific time period for test data) depending on the nature of the dataset.

Resampling Techniques:

SMOTE (Synthetic Minority Over-sampling Technique) and ENN (Edited Nearest Neighbors) are two popular techniques used for addressing class imbalance in imbalanced datasets.

● SMOTE is a synthetic oversampling technique that generates synthetic examples of the minority class by interpolating between existing minority class instances. This technique helps to increase the representation of the minority class and create a more balanced dataset. SMOTE has been widely used in various machine learning algorithms and has shown promising results in improving the performance of models on imbalanced datasets.

● ENN, on the other hand, is an undersampling technique that removes instances from the majority class that are misclassified by their nearest neighbors in the minority class. ENN helps to reduce the number of majority class instances and can be effective in eliminating noisy instances from the majority class that may negatively impact model performance.

SMOTE can be used to oversample the minority class to balance the class distribution, and ENN can be used to remove noisy instances from the majority class. The combination of SMOTE and ENN can help to create a more balanced dataset, which can improve the performance of logistic regression, decision tree, and random forest models in predicting hospital readmission risk for diabetic patients.

Reference (APA7 Style)



American Diabetes Association (Ed.). (2018, May). The cost of diabetes. The Cost of Diabetes | ADA. Retrieved April 23, 2023, from https://diabetes.org/about-us/statistics/cost-diabetes

American Diabetes Association. (2022, July 28). Statistics about diabetes. Statistics About Diabetes | ADA. Retrieved April 23, 2023, from

https://diabetes.org/about-us/statistics/about-diabetes

Davis, S., Zhang, J., Lee, I., Rezaei, M., Greiner, R., McAlister, F. A., & Padwal, R. (2022). Effective hospital readmission prediction models using machine-learned features. BMC Health Services Research, 22(1). https://doi.org/10.1186/s12913-022-08748-y

Fernández Alberto, Krawczyk, B., García Salvador, Galar, M., Herrera, F., & Prati, R. C. (2018). Learning from imbalanced data sets. Springer.

Hosseinzadeh, A., Izadi, M., Verma, A., Precup, D., & Buckeridge, D. (2013). Assessing the predictability of hospital readmission using Machine Learning. Proceedings of the AAAI Conference on Artificial Intelligence, 27(2), 1532–1538. https://doi.org/10.1609/aaai.v27i2.18995

Kampan, P. (2006). Effects of Counseling and Implementation of Clinical Pathway on Diabetic Patients Hospitalized with Hypoglycemia.

Lin, C.-Y., Singh, H. S., Kar, R., & Raza, U. (2017). What Are Predictors of Medication Change and Hospital Readmission in Diabetic Patients?

Lu, H., & Uddin, S. (2022). Explainable stacking-based model for predicting hospital readmission for Diabetic patients. Information, 13(9), 436. https://doi.org/10.3390/info13090436

Michailidis, P., Dimitriadou, A., Papadimitriou, T., & Gogas, P. (2022). Forecasting hospital readmissions with machine learning. Healthcare, 10(6), 981.

https://doi.org/10.3390/healthcare10060981

Ostling, S., Wyckoff, J., Ciarkowski, S. L., Pai, C.-W., Choe, H. M., Bahl, V., & Gianchandani, R. (2017). The relationship between diabetes mellitus and 30-day readmission rates. Clinical Diabetes and Endocrinology, 3(1). https://doi.org/10.1186/s40842-016-0040-x

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, 1–11.

https://doi.org/10.1155/2014/781670