Hospital Readmission Probability In Diabetic Patients: Predictive modeling with machine learning classifiers

1. CONTEXT

1.1. PROBLEM STATEMENT

This report outlines the development and evaluation of predictive models for hospital readmissions, with a focus on the utilization of machine learning techniques to automatically generate features from longitudinal patient data. Hospital readmission refers to the scenario where a patient, following their initial hospital discharge, is admitted back to the hospital within a specified period. It is a crucial healthcare metric, signifying both the quality of hospital services and the financial impact on healthcare expenditure. In 2011, American hospitals spent over $41 billion on diabetic patients who got readmitted within 30 days of discharge. Hospital readmissions present a significant challenge to healthcare systems due to their associated costs, and our objective was to improve the accuracy of predicting such readmissions for diabetic patients.[1] By identifying the factors influencing readmission and developing predictive models early on, hospitals can potentially save millions of dollars while simultaneously enhancing the quality of patient care.[2] To address this critical issue, we leveraged medical claims dataset to explore the following question:

- What is the probability of a diabetic patient getting readmitted within 30 days of discharge?

1.2. METHODS

We analyzed patients claims data from 1998 to 2008 in the US for patients suffering from diabetes. Our target value is diabetic patients who get readmitted within 30 days of their previous admission. Success in prediction is measured using the area under the curve (AUC) statistic, precision, accuracy, recall, and F1 score.

2. DATA ANALYSIS

2.1. DATASET DESCRIPTION

Our analysis is based on a comprehensive dataset comprising a decade's worth of hospital readmission data from 1998-2008, specifically focused on diabetes-related diagnoses. The dataset, "hospital_readmissions.csv," consists of 101,766 data entries, with 50 columns capturing various aspects of hospital readmissions. This dataset serves as the foundation for our study, encompassing the data analysis, preprocessing steps, and the application of machine learning classification models to address our objectives.

2.2. DATASET UNDERSTANDING

In our project, we are employing the data mining process, focusing on classification techniques. However, it's important to acknowledge that we must be vigilant regarding potential biases related no information of a patients' family's medical history for diabetes or any geographical information in our dataset. We understand these biases can manifest in various stages of the data mining process and can have significant implications on our analysis and model outcomes.

These potential bias directions encompass various stages of the data mining process and model development. Here are some key potential bias directions to be aware of:

Data Collection Bias:

Sampling Bias: Bias can occur if the sample of data collected is not representative of the broader population. For example, in our dataset there may be a possibility of being overrepresented or underrepresented of a certain geographic area as we don't have the data available.

2.3. DATA PREPRATION

2.3.1. DATA GRANUALITY

We are working on individual patient level dataset which entails following information:

- **Encounter ID:** Unique identifier of an encounter
- **Patient number:** Unique identifier of a patient
- **Race** Values: Caucasian, Asian, African American, Hispanic, and other
- **Gender** Values: male, female, and unknown/invalid
- **Age** Grouped in 10-year intervals: 0, 10), 10, 20), …, 90, 100)
- **Weight:** Weight in pounds
- **Admission type:** Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available
- **Discharge disposition:** Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available
- **Admission source:** Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital
- **Time in hospital:** Integer number of days between admission and discharge
- **Payer code:** Integer identifier corresponding to 23 distinct values, for example, Blue Cross/Blue Shield, Medicare, and self-pay Medical

- **Medical specialty:** Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family/general practice, and surgeon
- **Number of lab procedures:** Number of lab tests performed during the encounter
- **Number of procedures:** Numeric Number of procedures (other than lab tests) performed during the encounter
- **Number of medications:** Number of distinct generic names administered during the encounter
- **Number of outpatient visits:** Number of outpatient visits of the patient in the year preceding the encounter
- **Number of emergency visits:** Number of emergency visits of the patient in the year preceding the encounter
- **Number of inpatient visits:** Number of inpatient visits of the patient in the year preceding the encounter
- **Diagnosis 1:** The primary diagnosis (coded as first three digits of ICD9); 848 distinct values
- **Diagnosis 2:** Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values
- **Diagnosis 3:** Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values
- **Number of diagnoses:** Number of diagnoses entered to the system 0%
- **Glucose serum test result:** Indicates the range of the result or if the test was not taken. Values: ">200," ">300," "normal," and "none" if not measured
- **A1c test result:** Indicates the range of the result or if the test was not taken. Values: ">8" if the result was greater than 8%, ">7" if the result was greater than 7% but less than 8%, "normal" if the result was less than 7%, and "none" if not measured.
- **Change of medications:** Indicates if there was a change in diabetic medications (either dosage or generic name). Values: "change" and "no change"
- **Diabetes medications:** Indicates if there was any diabetic medication prescribed. Values: "yes" and "no"
- 24 features for medications for the generic names: **metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, sitagliptin, insulin, glyburide-metformin, glipizide-metformin, glimepiride- pioglitazone, metformin-**

rosiglitazone, and metformin- pioglitazone, the feature indicates whether the drug was prescribed or there was a change in the dosage. Values: “up” if the dosage was increased during the encounter, “down” if the dosage was decreased, “steady” if the dosage did not change, and “no” if the drug was not prescribed

- **Readmitted:** Days to inpatient readmission. Values: “<30” if the patient was readmitted in less than 30 days, “>30” if the patient was readmitted in more than 30 days, and “No” for no record of readmission.

2.3.2. DEALING WITH MISSING VALUES

Variable weight contains approximately 98% of the missing values so there is no significance in filling those missing values, so we decided to drop these variables. Variable Payer code and medical specialty contains approximate 40% missing values, so we also dropped these variables. Variables race, diag_1, diag_2, diag_3 and gender contain very few missing values as compared to other attributes which we dropped so for these attributes we also decided to drop those where missing values contains. [3]

Also, variables (drugs named citoglipton and examide), all records have the same value. So essentially these cannot provide any interpretive or discriminatory information for predicting readmission, so we decided to drop these two variables.

2.3.3. ENCODING VARIABLES

The original data used string values for gender, race, medication change, & each of the 24 drugs used. We encoded the “medication change” feature from “No” (no change) and “Ch” (changed) into 0 and 1. We also reduced both A1C test result and Glucose serum test result into categories of Normal, Abnormal and Not tested.[3]

2.3.4. ENCODING THE OUTCOME VARIABLE

We are looking whether the patient gets readmitted to the hospital within 30 days or not. The variable has < 30, > 30 and No Readmission categories. To reduce our problem to a binary classification, we combined the readmission after 30 days and no readmission into a single category making readmission within 30 days as 1 (11,357 patients) and no readmission within 30 days as 0 (90,409 patients).

2.3.5. FEATURE ENGINEERING

2.3.5.1. SERVICE UTILIZATION

The data contains variables for number of inpatient (admissions), emergency room visits and outpatient visits for a given patient in the previous year. These are (crude) measures of how much hospital/clinic services a person has used in the past year. We added these three to create a new variable called service utilization. The idea is to see which version gives us better results.

2.3.5.2. NUMBER OF MEDICATION CHANGES

The dataset contains 24 features for 24 drugs which indicate for each of these, whether a change in that medication was made or not during the current hospital stay of patient. We decided to count how many changes were made in total for each patient and declared that a new feature. The reasoning here was to both simplify the model and possibly discover a relationship with a number of changes regardless of which drug was changed.

The dataset contained up to three diagnoses for a given patient (primary, secondary, and additional). However, each of these had 700–900 unique ICD codes and it is extremely difficult to include them in the model and interpret meaningfully. Therefore, we collapsed these diagnosis codes into 9 disease categories in an almost similar fashion to that done in the original publication using this dataset. These 9 categories include Circulatory, Respiratory, Digestive, Diabetes, Injury, Musculoskeletal, Genitourinary, Neoplasms, and Others. Although we did this for primary, secondary, and additional diagnoses, we eventually decided to use only the primary diagnosis in our model.

2.3.7 FEATURE SELECTION

We used Lasso, or Least Absolute Shrinkage and Selection Operator for feature selection in our model. Lasso adds an L1 regularization term to the cost function.

The cost function for linear regression is typically represented as:

$Cost = SSE + \lambda * (\alpha * |\theta_1| + \alpha * |\theta_2| + ... + \alpha * |\theta n|)$

This regularization term encourages some feature coefficients to become zero, resulting in feature selection. The strength of the regularization term determines the sparsity of the selected features. Lasso optimizes the cost function to find a subset of important features while setting others to zero, simplifying the model. This process helps identify the most relevant features for the predictive model.[2]

3. MODELING

After performing Lasso regression on the data as a feature selection process, we were resulted with 74 columns to perform modelling. Upon running a logistic regression model, we realized the model was skewedly distributed in the target variable making the model exhibit bias towards the majority class, which is not readmitted within 30 days. The model couldn't efficiently predict patient's readmission within 30 days making the F1 score 0.

3.1. SMOTE (synthetic minority oversampling technique) ￼

SMOTE is one of the most commonly used oversampling methods to solve the imbalance problem. It aims to balance class distribution by randomly increasing minority class examples by replicating them. We were careful not to perform SMOTE on the entire dataset, but to only the train set after test-train split as it should reflect the real-world distribution and we wanted to avoid our model being overly optimistic due to testing it with a balanced test dataset.[4]

3.2. MODEL SELECTION

Given the evident imbalance in our dataset, we confidently opted for boosting models. Boosting algorithms are known for their remarkable efficacy in handling imbalanced datasets. They adaptively concentrate on minority class instances by employing weighted sampling and loss functions. Through their ensemble approach, these models amalgamate multiple weak learners, thereby enhancing overall performance and generalization.

Boosting methods offer flexibility, enabling us to fine-tune hyperparameters to achieve optimal results. Additionally, they often exhibit robustness to noisy data, making them a reliable choice in challenging scenarios. When coupled with resampling techniques, boosting becomes a potent tool for tackling class imbalance.

In our approach, we selected four robust models to address this issue:

1. Random Forest
2. Catboost Classifier
3. XGBoost Classifier
4. LightGBM Classifier

3.2.1. METHOD

Each model underwent rigorous evaluation through a 5-fold cross-validation process, where the training and validation data sets were systematically partitioned into distinct subsets in every iteration. This approach ensured a comprehensive assessment of model performance under varying conditions.

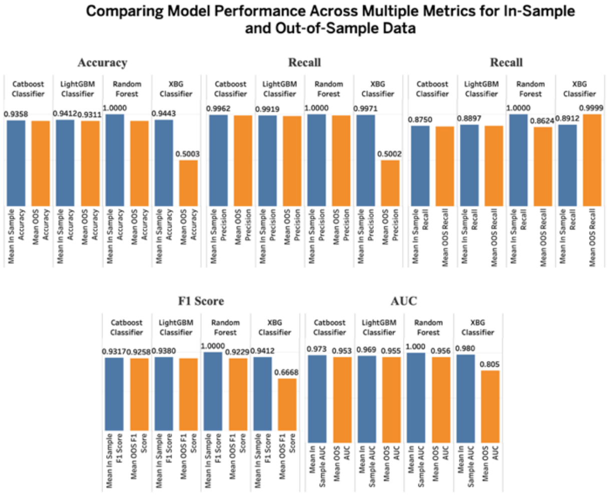
To address the class imbalance issue, Synthetic Minority Over-sampling Technique (SMOTE) was systematically applied to the training set following each partitioning, equitably augmenting the representation of minority class instances.

Multiple performance metrics, including Area Under the Curve (AUC), F1 Score, Recall, Precision, and Accuracy, were meticulously recorded during the evaluation process. These metrics serve as a crucial basis for comparative analysis.

In summary, we documented these performance metrics for both the out-of-sample (OOS) evaluation, representing the model's performance on unseen data, and the in-sample evaluation, reflecting the model's performance on the training data. Each metric serves as a quantitative measure of the model's effectiveness in classification tasks. A brief definition of these metrics is as follows:

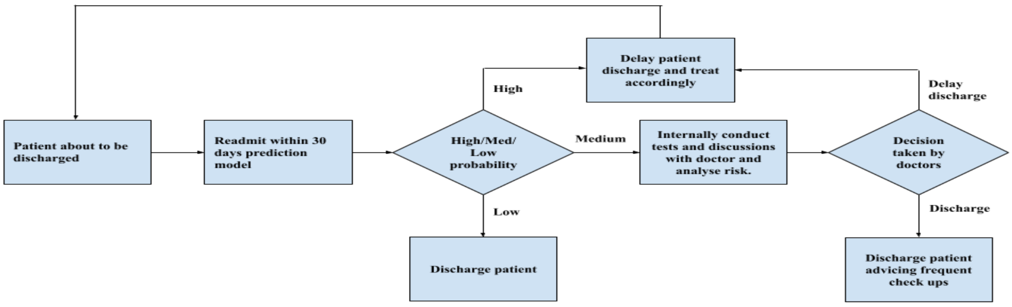
- **AUC (Area Under the Curve):** AUC quantifies the ability of a model to distinguish between positive and negative classes. A higher AUC indicates superior classification performance.
- **F1 Score**: F1 Score balances precision and recall, providing a harmonic mean. It is particularly useful when dealing with imbalanced datasets.
- **Recall**: Recall measures the model's capability to correctly identify positive instances. It is also known as a true positive rate or sensitivity.
- **Precision**: Precision quantifies the accuracy of positive predictions made by the model. It is also known as a positive predictive value.
- **Accuracy**: Accuracy represents the ratio of correctly predicted instances to the total number of instances. It's a fundamental measure of overall model correctness.

4. EVALUATION



Based on our model evaluation, we can see LightGBM Classifier, Catboost Classifier, Random Forest are three classifiers which are consistently performing well on both out of sample and in sample across all metrics. However, LightGBM Classifier performs slightly better than the remaining and hence we'll be using LightGBM during deployment. The accuracy(93.11%) and efficiency of the LightGBM model can greatly benefit hospital finances in several ways. Firstly, by accurately predicting readmissions, hospitals can proactively implement targeted interventions, reducing the number of unnecessary readmissions. This directly saves on the costs associated with readmission, including additional treatments, hospital stays, and other resources. Secondly, the predictive model allows hospitals to allocate their resources more efficiently. By identifying patients at a higher risk of readmission, healthcare facilities can provide tailored care and support, focusing on those who need it the most. This can lead to a reduction in resource wastage and better utilization of staff and infrastructure. Moreover, avoiding penalties imposed by regulatory bodies for high readmission rates is another financial benefit. Hospitals with excessive readmissions often face financial penalties, which can be mitigated by early interventions guided by the model's predictions[3].

5. DEPLOYMENT


Delay patient discharge and treat accordingly
High
Delay discharge
Patient about to be discharged
Readmit within 30 days prediction model
High/Med/ Low probability
Medium
Internally conduct tests and discussions with doctor and analyse risk.
Decision taken by doctors
Low
Discharge
Discharge patient
Discharge patient advicing frequent check ups

The proposed model offers significant potential for deployment within the healthcare industry, with a particular focus on hospital settings. Hospitals can leverage this model to estimate the likelihood of a diabetic patient being readmitted within 30 days after their initial discharge. This predictive capability empowers hospitals to implement tailored strategies based on the predicted probabilities.

In scenarios where the model indicates a high probability of readmission, hospitals can take proactive measures such as conducting additional diagnostic tests, providing enhanced treatment, and potentially delaying the patient's discharge. These actions not only contribute to the overall well-being of the patient but also enhance the hospital's reputation for providing comprehensive and patient-centric care. For cases where the predicted probability falls within a medium range, hospitals have the opportunity to engage in consultation with medical professionals. This collaborative approach allows hospitals to make informed decisions, including the continuation of patient monitoring or, in consultation with the attending physician, the option to discharge the patient with guidance on the importance of frequent check-ups. By implementing this model, hospitals can optimize patient care, minimize readmissions, and improve patient outcomes while maintaining operational efficiency.

5.1 ISSUES HOSPITAL SHOULD BE AWARE OF

It's essential for healthcare providers to recognize that the dataset underpinning this model is derived from the years 1998-2008. Given the dynamic nature of healthcare, trends and patient behaviors may have evolved since that time. Consequently, hospitals should be mindful of potential shifts in healthcare patterns and patient demographics that may not be fully captured by the historical dataset.

5.2 RISKS ASSOCIATED

This predictive model primarily relies on patients' physical attributes, including sex, race, medical diagnoses, and medical history, to estimate the likelihood of readmission within 30 days. However, it's crucial to acknowledge the inherent limitations of this approach. Notably, the model cannot discern the specific reasons why patients might be readmitted, as there are various factors at play. Patients could return within 30 days due to factors such as medication non-compliance or other unanticipated circumstances that the model may not be equipped to predict. Therefore, the model serves as a valuable tool for assessing risk but should be complemented with clinical judgment and individualized patient care. Models typically don't have insight into patients' daily behaviors, lifestyle choices, or their adherence to treatment plans. These factors can significantly impact readmission risk but are challenging to predict accurately. Unexpected life events, such as accidents or other health issues, can lead to readmissions, and these events are often difficult to predict in advance. Socioeconomic conditions, access to healthcare, and support systems can play a vital role in a patient's readmission risk. Models may not incorporate these factors effectively.

5.3 MITIGATE RISKS

Regular Patient Check-ups: Hospitals should establish a standard protocol for regular follow-up appointments with discharged patients, particularly those identified as high/medium-risk by predictive models. These follow-up visits serve as opportunities to monitor the patient's condition, address any potential health issues, and assess adherence to prescribed treatments. Engaging patients in their post-discharge care through regular check-ups fosters better communication, trust, and ensures early intervention if issues arise.

Collect High-Quality Data: Hospitals must prioritize the collection of comprehensive and accurate patient data, extending beyond medical records to include social determinants of health, lifestyle factors, and patient-reported outcomes. High-quality data are vital for improving the accuracy and reliability of predictive models, allowing healthcare providers to make more informed decisions. Data-driven insights derived from quality data can facilitate better risk stratification, leading to more effective interventions.

Telemedicine and Remote Monitoring: Invest in telemedicine and remote patient monitoring solutions to extend care beyond the hospital's walls. Remote monitoring technology allows hospitals to track patient vital signs, medication adherence, and symptom progression, thus enabling early intervention and potentially reducing the need for readmission.

APPENDIX

TEAM CONTRIBUTION

Team Members	Task(s) Undertaken
Caio Oliveira Neto Teixeira	Data Preparation, EDA & Report creation
Sathish Prasad	Modeling & Report creation
Nikita Singh	Data Preparation, EDA & Report creation
Jenny Zhang	Feature Engineering & Report creation
Lance Qiu	Modeling & Report creation

REFERENCES

1. https://www.ischool.berkeley.edu/projects/2017/what-are-predictors-medication-change-and-hospital-readmission-diabetic-patients
2. https://chat.openai.com/share/a1aa041b-60cf-4b54-87e1-0c60265cef77
3. https://www.modernhealthcare.com/safety-quality/hospital-readmissions-penalties-increasing-2024-cms#:~:text=For%20the%20upcoming%20year%2C%2070.1,virtually%20unchanged%20from%20last%20year.
4. https://chat.openai.com/share/d97e87c4-46eb-4ce7-beb1-824659d77461