PREDICTING EARLY READMISSION OF DIABETIC PATIENTS WITHIN 30 DAYS OF DISCHARGE

DA 620 CAPSTONE PROJECT

Daniel Hernandez

DECEMBER 20TH 2023

TABLE OF CONTENTS

CONTENT	PAGES
BACKGROUND	2
PROBLEM SCENARIO/BUSINESS ISSUE	2-3
OBJECTIVE/GOALS OF THE PROJECT	3
DATA EXPLOARTION/DATA VISUALIZATION	3-8
DATA MANIPULATION	9
METHODOLOGY/MODEL BUILDING	9-10
MODEL SELECTION	11-12
CONCLUSIONS/RECOMMENDATIONS	12-13
BIBLIOGRAPHY	14

BACKGROUND

Controlling blood sugar levels is essential for preventing complications like heart disease, damage to the kidneys, and nerve issues. Effective diabetes care, including medication, lifestyle modifications, and monitoring, improves the general health of those who have the disease. The management of diabetes during hospital stays presents several issues, including managing customized dietary demands, managing drug regimens, and dealing with blood sugar fluctuations just to name a few. Management can be made more difficult by poor communication between healthcare teams and a lack of diabetes education for staff members as well as patients. The implementation of patient-specific care plans, collaboration among specialists, and timely monitoring are crucial in addressing these obstacles and guaranteeing optimal management of diabetes in a hospital environment.

Early readmission poses challenges for both patients and healthcare facilities. If you look at the patients' side first there are instances where individuals may experience prolonged recovery, increased health care costs and are at a higher risk of complication. On the other side of that coin healthcare facilities can be strained with minimal resources, impact quality of care and in turn these facilities can lead to financial penalties under certain reimbursement models. In this project it will be essentials to address the root causes of readmissions for the improvement of patient outcomes and reducing that burden that healthcare systems may face.

PROBLEM SCENARIO/BUSINESS ISSUE

Higher readmission rates are a result of many difficulties in controlling diabetes when a patient is in the hospital. These include inconsistent glucose monitoring, medication errors, inadequate diabetes management training for personnel, and insufficient post-discharge care education for patients. These problems are made worse by a lack of coordination between

inpatient and outpatient treatment. Hospital expenses and patient outcomes may be significantly impacted by uninformed diabetes management. Poorly regulated blood sugar levels raise the risk of consequences such infections, cardiovascular problems, and delayed recovery from wounds. This can be caused by inconsistent or inadequate management. Unstable diabetes can increase the need for interventions, lengthen hospital stays, and raise the expense of healthcare.

OBJECTIVE/GOALS OF THE PROJECT

The objective of the project is to develop a predictive model to identify factors influencing early readmission of diabetic patients within 30 days of discharge. The main goals include gathering comprehensive data on diabetic patients, identifying relevant factors that could influence early readmission, employ statistical and machine learning techniques to build a robust predictive model that can analyze and identifies factors and predict the likelihood of early readmission. Validate the model using historical data and ensure its accuracy, sensitivity, and specificity, integrate the predictive model into healthcare systems to assist clinicians in identifying high risk patients and implementing preventive measures, and to assess the impact of the predictive model on reducing early readmission rates and improving patient outcomes.

For those with diabetes, addressing these problems through comprehensive treatment plans, employee education, and enhanced communication can help lower readmission rates. Putting into practice individualized, evidence-based diabetes management programs is essential to enhancing patient outcomes and reducing medical costs.

DATA EXPLORATION/DATA VISUALIZATION

Started off the data exploration chapter by doing a summary of the numerical features in the diabetes dataset, it included the mean, standard deviation, minimum and maximum values, and quartiles (25^{th}, 50^{th}, 75^{th} percentiles). Within this dataset time in hospital stood out, 2-3 days

was the average amount of days those patients stayed. I also did a correlation analysis to show the correlation between some of the numerical features. Generally, longer stays might indicate more sever conditions, complex treatments, or complications if you look at the correlation analysis num of medications and time in the hospital yielded a correlation of 0.466135 one of the highest in the group, if one or more features are added it might tell a different story within that mix of variables and may cause a further look. In figure 1 a histogram showed that the highest number of days recorded in the hospital was 3 days followed by 2 days.

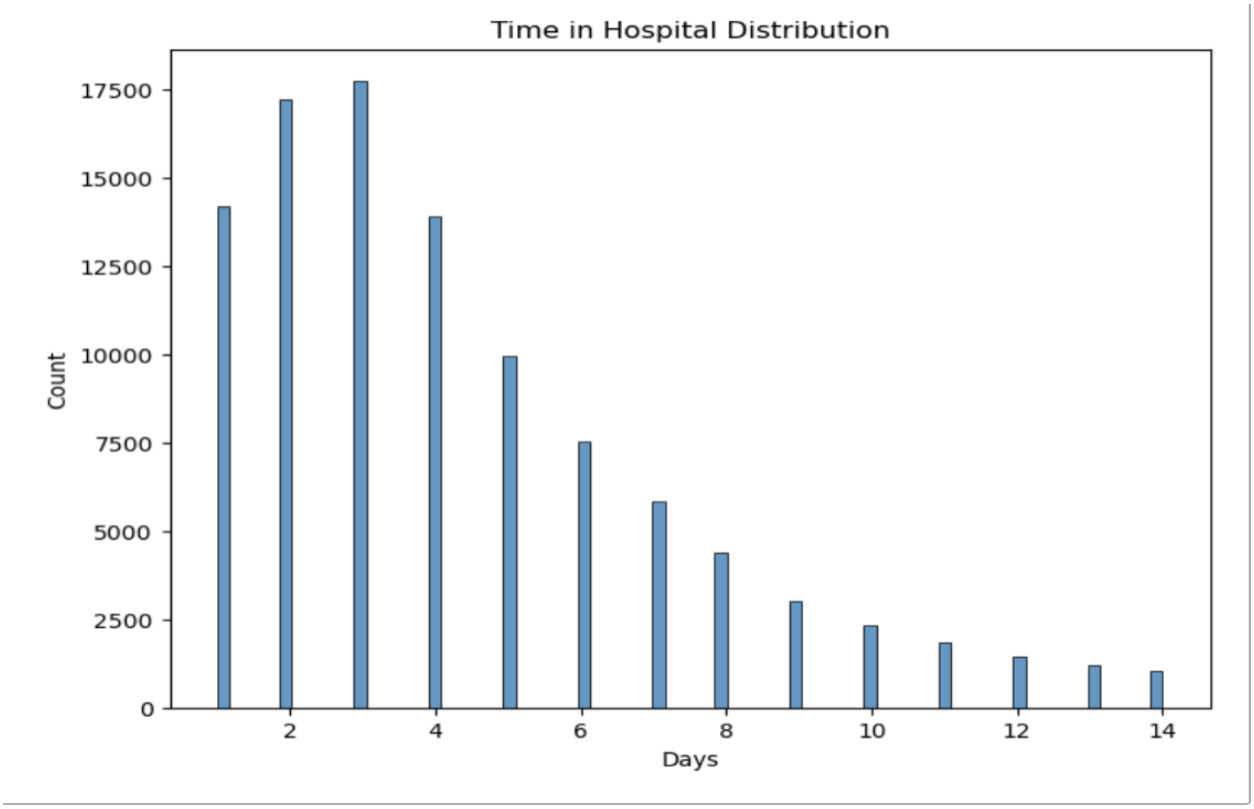


Figure 1 Number of days recorded in the hospital.

Readmission distribution in this dataset provides valuable insights into serval aspects of patient care and healthcare management. A skewed distribution with a higher number of readmissions might indicate potential gaps in care quality, discharge planning or a post discharge follow up. In

this instance we see that over 50000 people where not readmitted over 30000 people were readmitted over 30 days and only 10000 was readmitted in less than 30 days.

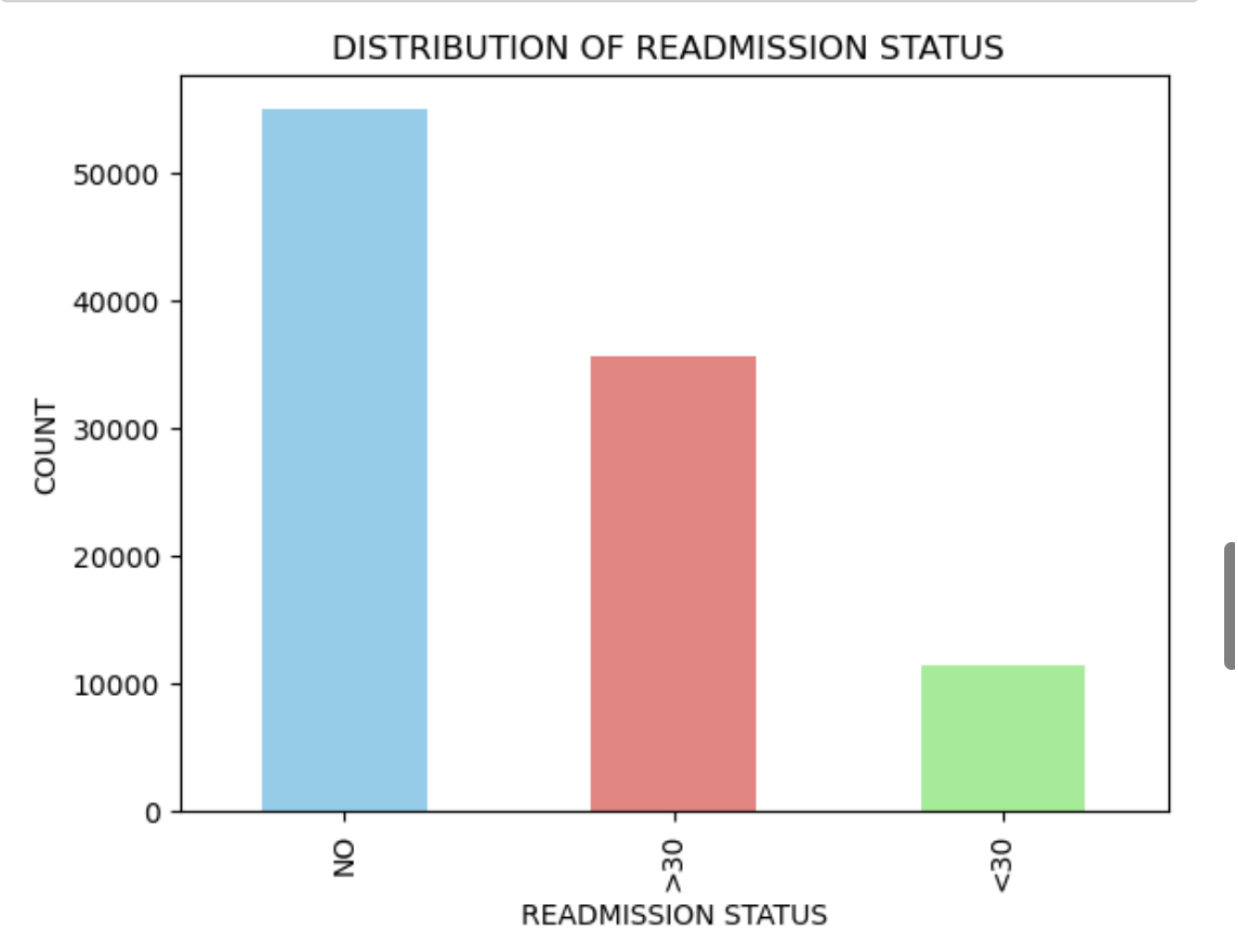


Figure 2 Bar chart of readmission status

Within this dataset Caucasian is 2.5 times the amount of the next race (African American). Examining the distribution of races helps in identifying potential disparities in healthcare access, utilization, and outcomes among different racial groups. Disproportionate representation might indicate underlying health inequities that need attention. Different racial groups might have varying prevalence's of certain diseases or different risk factors associated with specific health conditions. Analyzing race distributions alongside specific health issues helps in understanding these correlations.

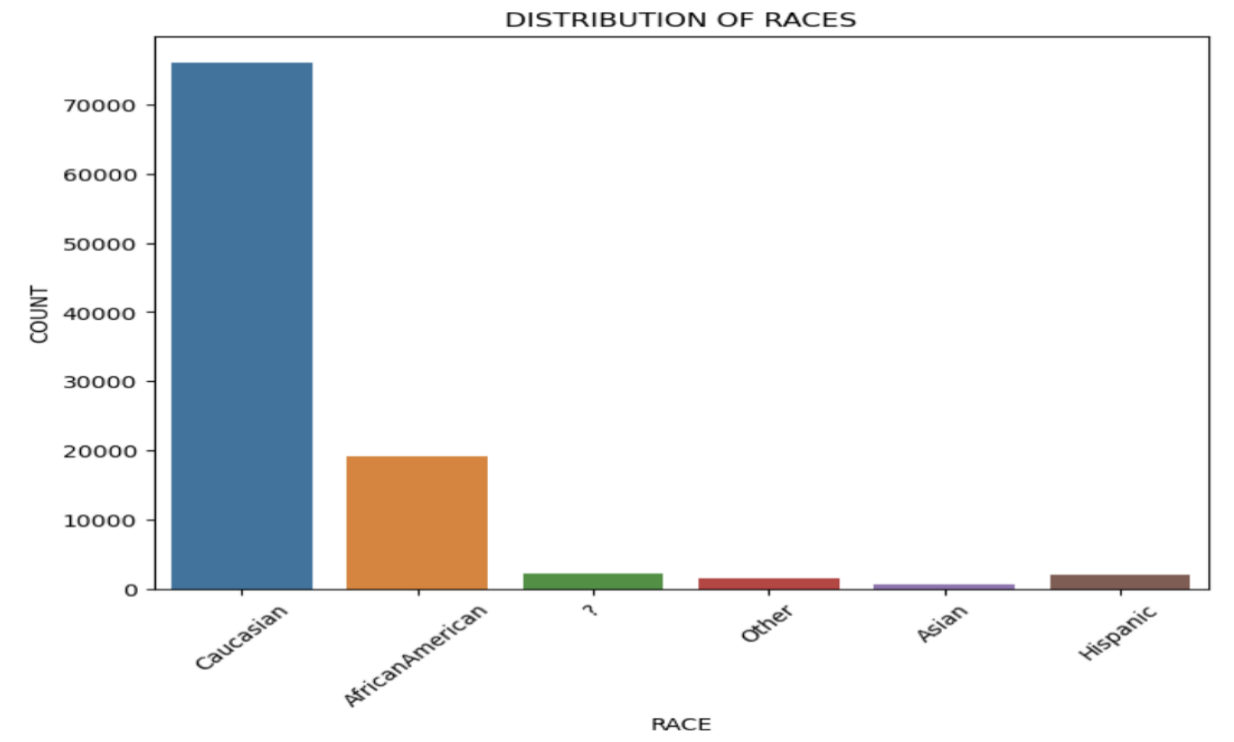


Figure 3 Distribution of Races

According to this dataset ages 70-80 has the most readmission counts under 30days with over 3000 of that age being readmitted. Understanding readmission rates across age groups helps in stratifying patients based on their risk of readmission. Certain age groups might be more prone to readmissions due to specific health conditions, complications, or comorbidities. Different age groups may exhibit varying patterns of diseases and health conditions. Analyzing readmission rates by age can highlight prevalent conditions within each age group and guide targeted disease management strategies. Higher readmission rates in certain age groups may impact resource allocation within healthcare settings. Hospitals can allocate resources and plan interventions more effectively by considering the readmission patterns across age groups.

```
readmitted   age
<30          [70-80)     3069
             [60-70)     2502
             [80-90)     2078
             [50-60)     1668
             [40-50)     1027
             [30-40)      424
             [90-100)     310
             [20-30)      236
             [10-20)       40
             [0-10)         3
>30          [70-80)     9475
             [60-70)     7897
             [80-90)     6223
             [50-60)     5917
             [40-50)     3278
             [30-40)     1187
             [90-100)     808
             [20-30)      510
             [10-20)      224
             [0-10)        26
NO           [70-80)    13524
             [60-70)    12084
             [50-60)     9671
             [80-90)     8896
             [40-50)     5380
             [30-40)     2164
             [90-100)    1675
             [20-30)      911
             [10-20)      427
             [0-10)       132
Name: age, dtype: int64
```

Figure 4 Readmitted age groups.

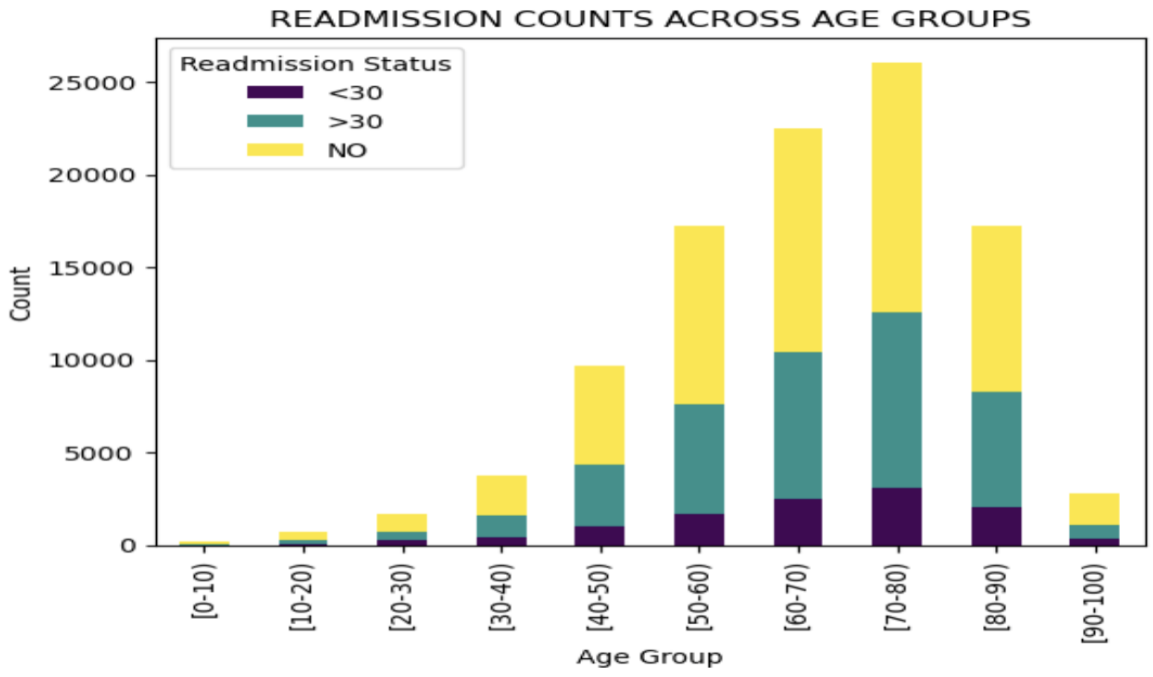


Figure 5 Readmission of ages bar chart.

Admission types include 1: emergency, 2: urgent, 3, elective, 4: newborn, 5: not available, 6: null trauma, 7: center, 8, not mapped. I did a bar chart for the time in hospital and admission type. The relationship between admission type and time spent in the hospital can reveal important insights about various aspects related to patient care, treatment, and resource allocation within a healthcare setting. Different admission types might correlate with varying degrees of illness severity. Emergency admissions might generally have shorter hospital stays due to urgent care needs, while elective admissions could imply planned procedures with longer expected stays. Analyzing the time spent in the hospital per admission type can help healthcare administrators allocate resources effectively. For instance, if certain admission types consistently result in longer stays, it might indicate the need for more resources in those areas.

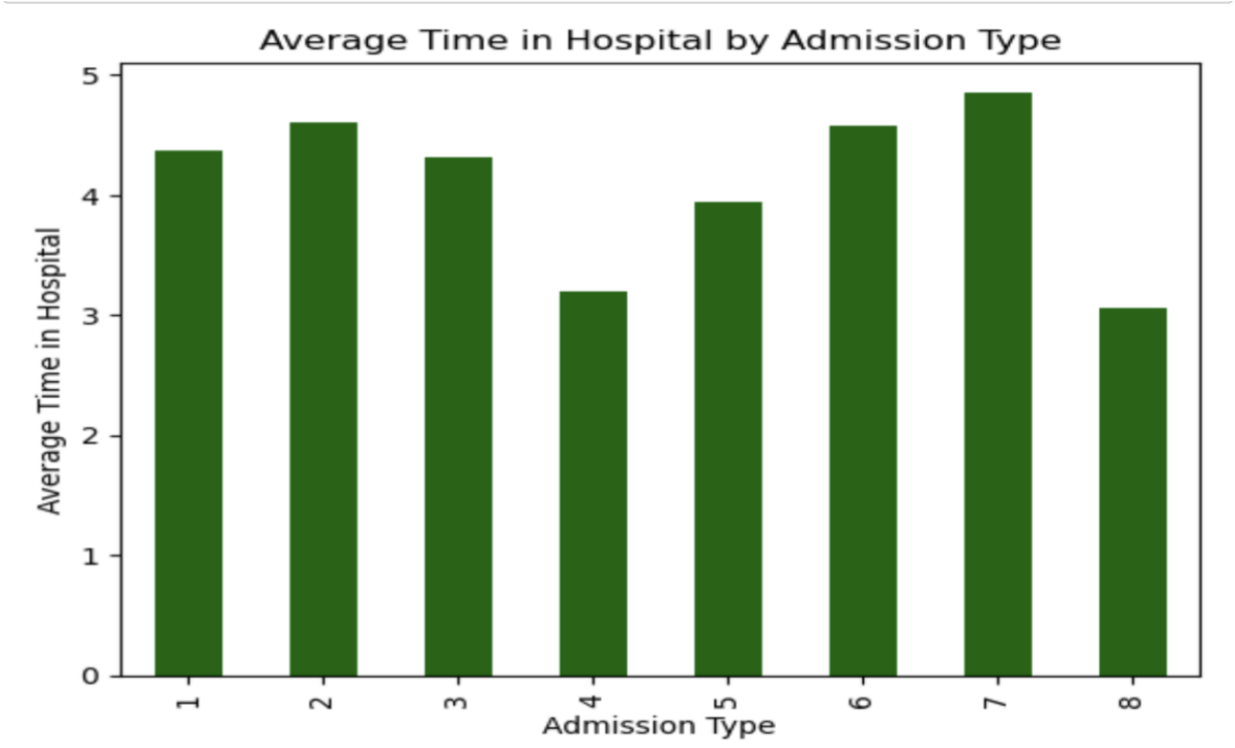


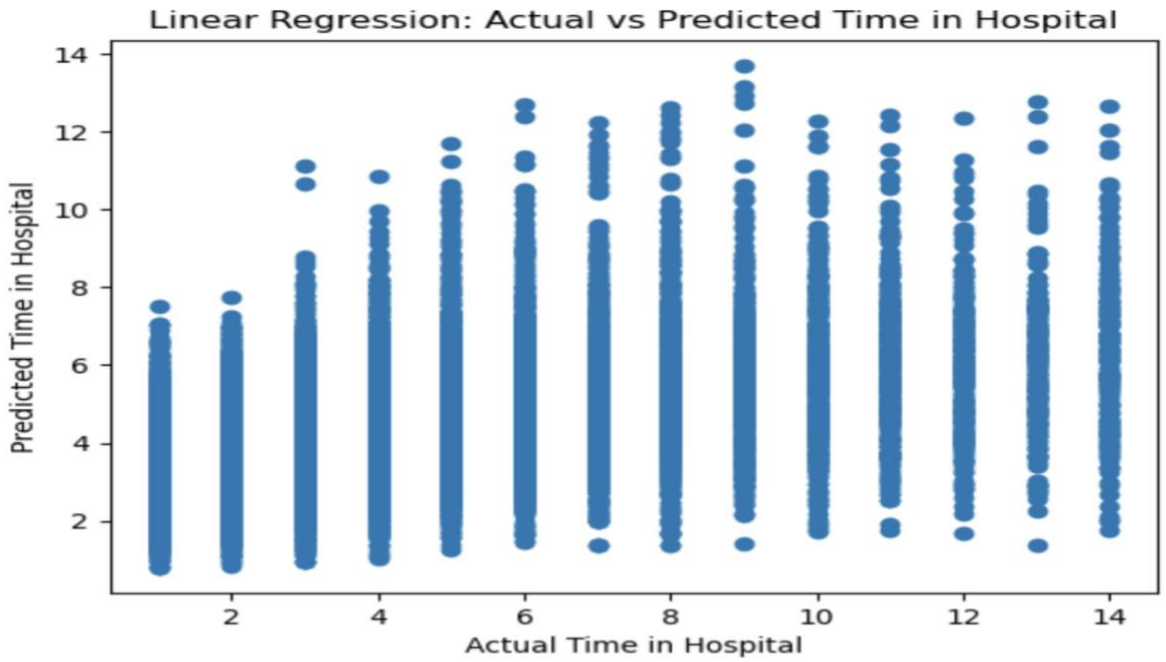
Figure 6 Average time in hospital by admission type.

DATA MANIPULATION

Within this dataset there was little to do in terms of manipulating the data in my estimation. Isnull identifies missing values in the data frame and sum calculates the sum of missing values for each column. The result is a series where each column name is associated with the count of missing values within their respective columns and in this case no columns had any null or missing values. I also performed a one hot encoding on two categorical columns (race and gender) in the data frame, I converted these categorical columns into numerical representations because it is suitable for machine and logistical models. I did a mapping dictionary which converted age groups that was represented as strings into numerical values. Lastly, I dropped a column (weight) because it provided no context because it was represented by a question mark (?) and it made no sense keeping the column because it may affect the dataset in the long run, but this weight column could have been a great indicator for readmitting patients.

METHODOLOGY/MODEL BUILDING.

I performed two predictive models one using logistic regression for reclassification and another using random forest for more complex predictions. Logistic Regression standardizes the features to ensure they have similar scales. Utilized numerical features (num_lab_procedures, num_procedures, num_medications to predict the numerical target variable time in hospital. The use of mean squared error (MSE and R-Squared for model evaluation and visualized predictions vs actual values with a scatter plot. In the second model I used the same numerical variables to predict the binary target variable readmitted. I employed an accuracy classification report along with a confusion matrix to assess the model's performance.




```
Linear Regression Model:
Mean Squared Error: 6.447867330149717
R-squared: 0.25921285752766565
```

Figure 7 Actual vs Predicted time in the hospital.

```
Random Forest Classifier Model:
Accuracy: 0.494890439225705

Classification Report:
              precision    recall  f1-score   support

         <30       0.12      0.03      0.05      2285
         >30       0.38      0.27      0.32      7117
          NO       0.55      0.74      0.63     10952

    accuracy                           0.49     20354
   macro avg       0.35      0.35      0.33     20354
weighted avg       0.44      0.49      0.46     20354


Confusion Matrix:
 [[  67  607 1611]
 [ 210 1956 4951]
 [ 303 2599 8050]]
```

Figure 8 Classification Report and confusion matrix

MODEL SELECTION

For model selection, it is crucial to consider the nature of the problem which is analyzing the readmission of patients in less than 30 days. In this case I chose a linear regression model for a numerical prediction and a random forest classifier for binary classification. For model 1 (Linear Regression for Numerical Prediction), since the task involves predicting a numerical value, specifically the amount of time in the hospital, linear regression is one of the best choices, it is well suited for regression problems where the target variable is continuous. An easy way to understand the connections between input features and the target variable is to use Linear Regression. Metrics like Mean Squared Error (MSE) and R-squared are frequently employed for numerical prediction applications. These measures shed light on how closely the model's predictions match the actual data.

Using Random Forest Classifier for Binary Classification is an example of a binary classification issue since the aim is to predict binary outcomes, which is readmitted in more or less than 30 days or not. Random Forest can capture complex relationships and non-linear patterns in the data. This flexibility is advantageous when dealing with diverse and intricate features. By combining predictions from multiple decision tress, random forest tends to be more robust and less prone to overfitting compared to individual decision trees. This is beneficial for improving generalization to unseen data.

Linear Regression is computationally efficient and scales well with large datasets. Random Forest, although more computationally intensive, is still scalable and often performs well with moderate-sized datasets. Linear Regression offers simplicity and interpretability, while Random Forest introduces more complexity. Both models were evaluated using appropriate metrics for their respective tasks. For regression (Linear Regression), Mean Squared Error and

R-squared were used. For classification (Random Forest), accuracy, classification report, and confusion matrix were employed. In summary, the model selection was driven by the characteristics of the problem (readmission of patients in less than 30days), the nature of the target variables, and the specific strengths of each model type.

CONCLUSION/RECOMMENDATIONS

Why is the data important?

Healthcare datasets, such as the one under examination, are invaluable resources for shaping informed decision-making and improving patient outcomes. In this dataset, each entry represents a patient encounter, encompassing a wealth of information ranging from demographic details to medical procedures and medications. The importance of this data lies in its potential to drive evidence-based practices, optimize resource allocation, and enhance the overall efficiency of healthcare systems.

Analysis of demographic features, including race, gender, and age, revealed the composition of the patient population. Understanding these demographics is crucial for tailoring healthcare services to diverse patient needs. Examination of medication-related features provided insights into the prevalence and changes in drug prescriptions. These insights are vital for optimizing treatment plans, ensuring patient adherence, and managing medication-related risks. Exploration of features like outpatient visits, emergency visits, and inpatient stays shed light on patient healthcare utilization. Recognizing these patterns is essential for resource planning, identifying potential areas for intervention, and enhancing overall healthcare efficiency.

Building predictive models, particularly Linear Regression and Random Forest Classifier, facilitated the identification of factors influencing patient readmission. Understanding these

predictors empowers healthcare providers to implement targeted interventions, potentially reducing readmission rates.

Recommendations

Implement patient-centered interventions based on demographic insights to address the unique needs of different patient groups. Enhance medication management protocols by closely monitoring medication usage trends, addressing patterns of dosage changes, and ensuring safe and effective drug regimens. Optimize healthcare resources based on insights into patient healthcare utilization patterns. This includes planning for outpatient services, emergency care, and inpatient Develop targeted strategies for preventing patient readmissions, leveraging insights from predictive models. This might involve interventions related to post-discharge care, medication adherence, and lifestyle management. Establish a system for continuous monitoring of healthcare data to adapt interventions dynamically. Regularly updating models and strategies ensures responsiveness to evolving patient needs and healthcare dynamics.

In conclusion, this dataset provides a comprehensive lens into the healthcare landscape, enabling evidence-based decision-making. The insights garnered can guide strategic initiatives, fostering a patient-centric, efficient, and responsive healthcare system. Continuous monitoring and adaptation based on these findings will contribute to ongoing improvements in patient care and healthcare resource management.

BIBLOGRAPHY

"Diabetes 130-US Hospitals for Years 1999-2008." *UCI Machine Learning Repository*, archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008. Accessed 21 Dec. 2023.

Dungan, Kathleen M. "The Effect of Diabetes on Hospital Readmissions." *Journal of Diabetes Science and Technology*, U.S. National Library of Medicine, 1 Sept. 2012, www.ncbi.nlm.nih.gov/pmc/articles/PMC3570838/.

Rubin, Daniel J. "Correction to: Hospital Readmission of Patients with Diabetes - Current Diabetes Reports." *SpringerLink*, Springer US, 13 Mar. 2018, link.springer.com/article/10.1007/s11892-018-0989-1.

Peggy Chou, MD. "Reducing Hospital Readmission for Diabetes: Context & Solutions." *Stability Health*, 7 June 2022, stabilityhealth.com/reducing-hospital-readmissions/.