12.2 Course Project: Milestone 5 Final Project Paper

Allyson Totpal

DSC630

6/1/25

Introduction

Background

Hospital readmissions pose both clinical and financial challenges, with chronic illness patients especially vulnerable to complications that lead to re-hospitalization. This project will explore readmission risk among patients with three highly comorbid chronic conditions - diabetes, hypertension, and chronic kidney disease (CKD) - as they often co-occur, magnify risk, and are prevalent in aging and underserved populations. Identifying these high-risk patients can assist in targeted interventions, optimization of care coordination, and reduction of costs to improve conditions for both the hospital and patient side.

Objective

To build and evaluate predictive models that estimate the likelihood of 30-day hospital readmission among patients with the comorbidity trio: diabetes, hypertension, and CKD. The goal is to:

- Identify key predictors of readmission amongst chronically ill patients
- Understand how comorbidities may impact readmission risk in patients
- Provide relevant insights into patient management strategies

Research Questions

- What comorbidities are most predictive of hospital readmission?
- Which models provide both predictive accuracy and interpretability for the data being handled?

Learning Goals

This project is aimed to:

- Apply predictive modeling to a real-word, structured healthcare dataset

- Explore how chronic conditions influence clinical outcomes
- Understand model performance trade-offs between interpretability and accuracy
- Gain hands-on experience with healthcare feature engineering and ethics

Overview of Data Used

Dataset

This project will be utilizing the UCI Diabetes dataset, which includes around 100,000 patient encounters across 130 US hospitals between the years 1999-2008 and lengths of stay up to 14 days (Cios, Clore, DeShazo, Gennings, Olmo, Strack, & Ventura, 2014). The dataset includes features such as demographics, diagnostic codes (ICD-9), medications, lab tests, admission type, length of stay, and readmission status (Cios et al., 2014).

Risks & Ethical Considerations

Some risks include:

- Outdated Dataset: The UCI Diabetes dataset contains records obtained from 1999-2008. While valuable for academic modeling, it may not accurately reflect current clinical practices, medication standards, or demographic trends and thus will limit generalizability of findings to today's environment. However, it can provide a foundation to scale to today's environment with more current data.
- Imbalanced Data: The target variable of 30-day readmission is imbalanced with fewer positive cases and could bias model predictions towards the majority class, which can be addressed with stratified sampling, class weighting, or resampling techniques.
- ICD-9 Code Limitations: The dataset uses ICD-9 diagnostic codes, codes that are less detailed than the current ICD-10 codes used in modern practice, and could cause issues with mapping the codes accurately into meaningful groupings.

Ethical considerations include:

- Bias in Historical Data: The dataset may reflect biases in healthcare delivery from 20+ years ago, including disparities in treatment across race, gender, or insurance status and models will be assessed for contrasting impact and acknowledge these biases.
- Fairness and Interpretability: High-performing models may not be easily interpretable for clinical decision-making, but the project will balance accuracy with explainability and highlight any trade-offs.
- Patient Privacy: Although the dataset is de-identified, protecting patient information is important. Therefore, the project will adhere to best practices for ethical use of healthcare data and avoid any attempt at re-identification.

Exploratory Data Analysis

As exploration of the data starts, bar plots are especially useful for explaining the data regarding the distribution of readmission status and comorbidity counts. By exploring the distribution of readmission status, we can observe that it was highly imbalanced, showing that readmissions within 30 days did not even account for half of the cases against readmissions over 30 days and no readmissions (Figure 1). In Figure 2, readmission rate by age group was explored, and observations were made showcasing that the age group 20-30 years old had the highest percentage of readmission within 30 days versus other age groups.

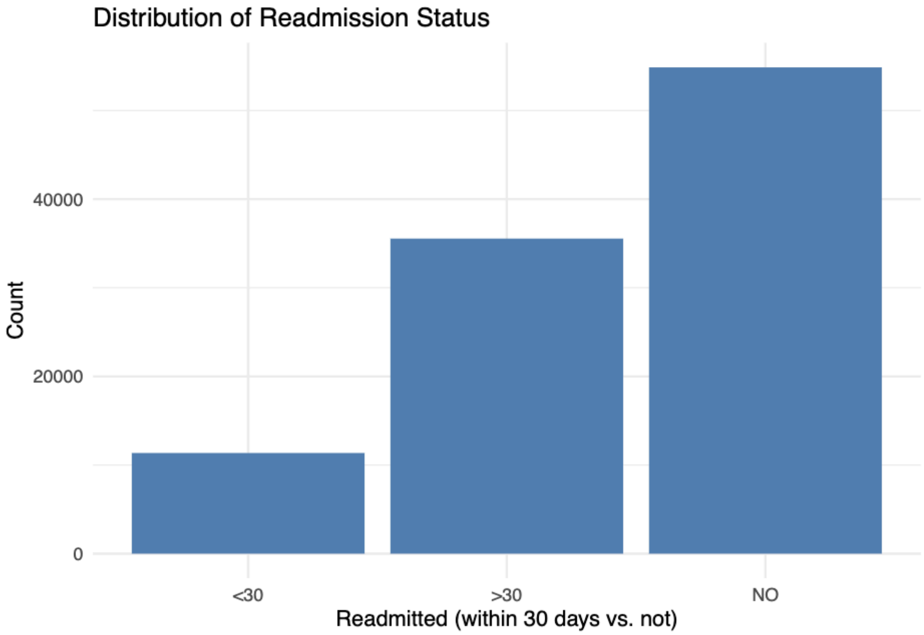

Figure 1: Distribution of Readmission Status (<30 days, >30 days, no readmission)

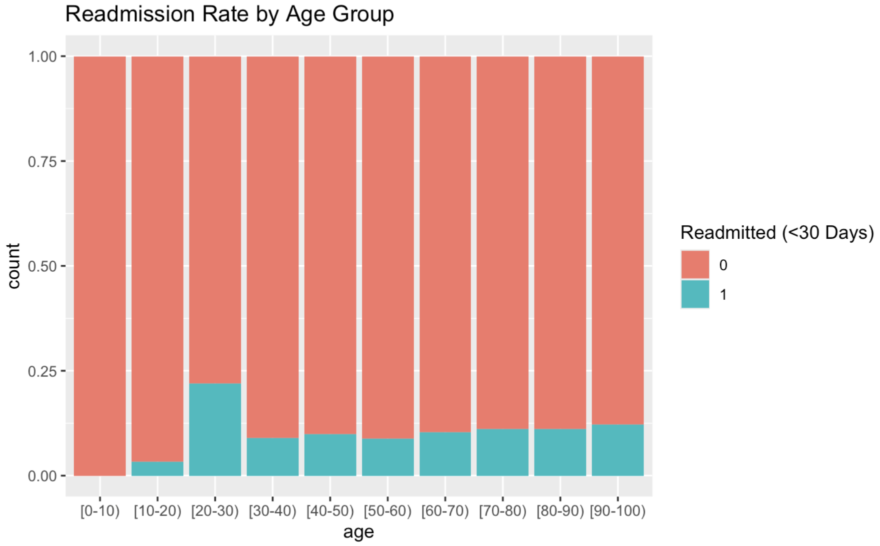

Figure 2: Readmission Rate by Age Group

Methods of Analysis

Data Preparation

Because the dataset is 100,000+ entries, processing that amount of data would be timely to run multiple models on. To help simulate pulling the last 3 years of the dataset (2005-2008) and because it did not include year of admission/entry, the dataset was filtered to only include the most recent 30% of patient records using encounter_id. To further improve efficiency while maintaining class diversity, a random sample of 10,000 records was drawn from the filtered set. Keeping with the original comorbidities while still expanding the scope, I added binary flags for patients with hypertension, CKD, and both using their respective ICD-9 codes. Variables with 30% or more missing values were removed, and categorical variables were converted to factors and one-hot encoded using model.matrix(). I also removed irrelevant and high-missing columns like weight, payer code, medical specialty, and discharge disposition ID, and removed the readmitted column as it was redundant after creating a target variable of readmit_30.

As mentioned previously, it was highlighted that the target variable (readmitted_30) was highly imbalanced, with few readmissions within 30 day encounters compared to over 30 day readmissions and non-readmission. To combat this, SMOTE (Synthetic Minority Over-Sampling Technique) was applied to duplicate positive cases and interpolate between neighbors and also conducted a post-balancing data split of 70/30. The balanced dataset will help to improve model sensitivity to enable better detection of patients likely to be readmitted.

Models & Metrics

The project will experiment with both interpretable and high-performing models including:

- Logistic Regression for baseline performance and feature interpretability
- Random Forest to capture nonlinear interactions and variable importance
- XGBoost for optimized performance, especially for imbalance data

The model performances will be evaluated using:

- ROC AUC for overall model discrimination
- Precision, recall, and F1-score for evaluating the imbalance of the dataset

- Confusion matrix to visualize trade-offs in false positives/negatives

Logistic Regression

To establish a baseline for predicting 30-day hospital readmissions, I implemented a Logistic Regression model using the balanced dataset generated through SMOTE. A Logistic Regression model was trained using all available features after cleaning to predict the binary target variable readmit_30. The model predicted probabilities for each patient in the test set and these probabilities were then thresholded at 0.5 to classify patients as either “Yes” (readmitted) or “No” (not readmitted).

Random Forest

To improve upon the performance of the Logistic Regression baseline, I implemented a Random Forest classifier as a robust ensemble learning method that constructs multiple decision trees and aggregates their results for more accurate and stable predictions. The Random Forest model was trained using the default parameters with 100 trees (ntree = 100) to reduce variance and avoid overfitting. Each tree in the forest is trained on a bootstrap sample and considers a random subset of features at each split. This randomness helps the ensemble learn diverse patterns and reduces the risk of overfitting to noisy or redundant features.

Two types of predictions were generated on the test set, class predictions (Yes or No) for accuracy and confusion matrix evaluation, and probability predictions specifically on the probability of readmission for ROC/AUC analysis,

XGBoost

Further improving on predictive performance, I implemented an XGBoost model, a gradient-boosted tree algorithm known for its high accuracy, efficiency, and ability to handle complex relationships in data. Since XGBoost requires numeric matrix input, the training and testing data were converted from data frames into matrices. The readmitted_30 outcome variable was also transformed from “Yes”/”No” to 1/0 to comply with XGBoost’s expected label format. Then we used XGBoost’s optimized data structure, DMatrix, which reduces memory usage and provides efficient training. I trained the model with the following hyperparameters for a balance of performance and generalizability:

- objective = “binary:logistic” - for binary classification
- eval_metric = "auc" - to track AUC score during training
- nrounds = 100 - number of boosting rounds
- max_depth = 6 - maximum depth of each decision tree
- eta = 0.1 - learning rate (controls how quickly the model learns)
- subsample = 0.8 - fraction of rows sampled for each tree
- colsample_bytree = 0.8 - fraction of columns used by each tree

Results & Findings Explained

Logistic Regression

The Logistic Regression model correctly predicted about 76% of patient outcomes, however it missed more than half of the actual readmissions and was better at identifying who would not be readmitted than who would. It provided an overall good baseline to start with, but did not perform well enough for real-world decision-making. The model achieved a 0.7742 AUC score, showing that the model has some predictive power, but needs further improvement as it only provides moderate-level accuracy. Its precision score of 56% also needs improvement, as when the model predicts a readmission, it is only correct about half the time. The model also only caught about 4 out of every 10 actual readmissions (recall) and an F1-score of 0.4577 reflects an imbalance between precision and recall. The Logistic Regression model served as a baseline, showing moderate performance and highlighting the need for more complex models to better identify at-risk patients.

Random Forest

The Random Forest model was able to make smarter predictions and correctly predicted about 90% of all patient outcomes. It caught over 60% of the patients who were readmitted within 30 days and when it predicted who would be readmitted, it was always right in this test. This shows that the model was extremely confident with high precision, but with a low recall, it caught less than 61% of actual admissions, missing about 39% of true cases and is a concern in clinical settings as it would be missing at-risk patients. An F1-score of 0.7593 indicates a strong balance of precision and recall, and has a high discriminatory power with an AUC score of 0.9261. While the model achieved perfect precision, it was conservative by avoiding false alarms at the expense of missing some true admissions. However, it did outperform the Logistic Regression model significantly, making it a strong candidate for predictive employment.

XGBoost

Overall, the XGBoost model outperformed the previous two and showcased better balance overall. The XGBoost model predicted about 90% of patient outcomes correctly, and caught slightly more true readmissions than the Random Forest Model while still maintaining very high precision (recall/sensitivity 64.5%). With an AUC score of 0.8654, it may not have performed as well as the Random Forest model, but it still achieved a desirable score to show its strong ability to discriminate between classes. The model also had very few false positives (recall of 98.8%) and demonstrated the best trade-off between precision and recall, outperforming the previous two models in terms of overall balance. Its gradient boosting structure allowed it to adapt well to subtle patterns in the data.

ROC Curves

Because Logistic Regression was only to provide a baseline, the ROC Curves shown in Figure 3 are for Random Forest and XGBoost as comparison. ROC curves help to visualize how good a model is at spotting future readmissions and the higher the curve means better performance. The Random Forest Model had the strongest curve, meaning it was confident at separating at-risk patients from those not likely to return. XGBoost had a slightly more balanced curve, and was nearly as good as the Random Forest, but more inclusive in catching true admissions. Advanced models Random Forest and XGBoost are effective at recognizing patterns that signal hospital readmission, with the Random Forest model performing slightly better overall in this test.

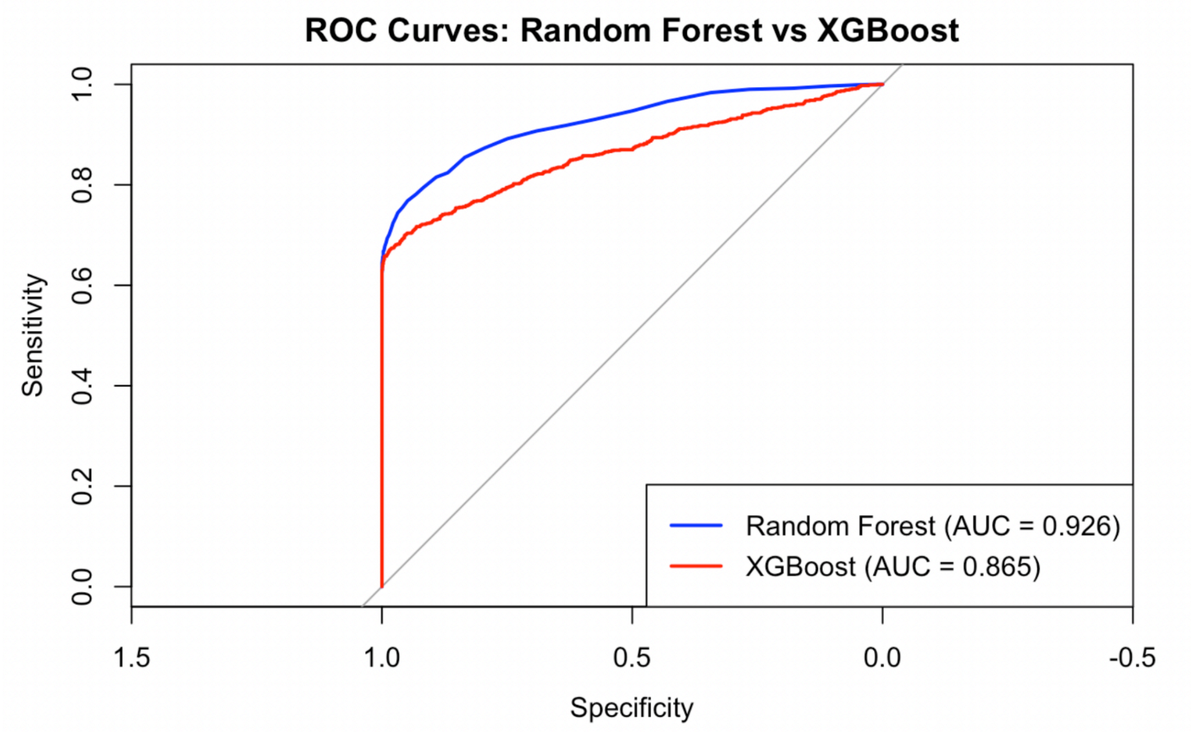

Figure 3: ROC Curves for Random Forest and XGBoost Models

Feature Importance

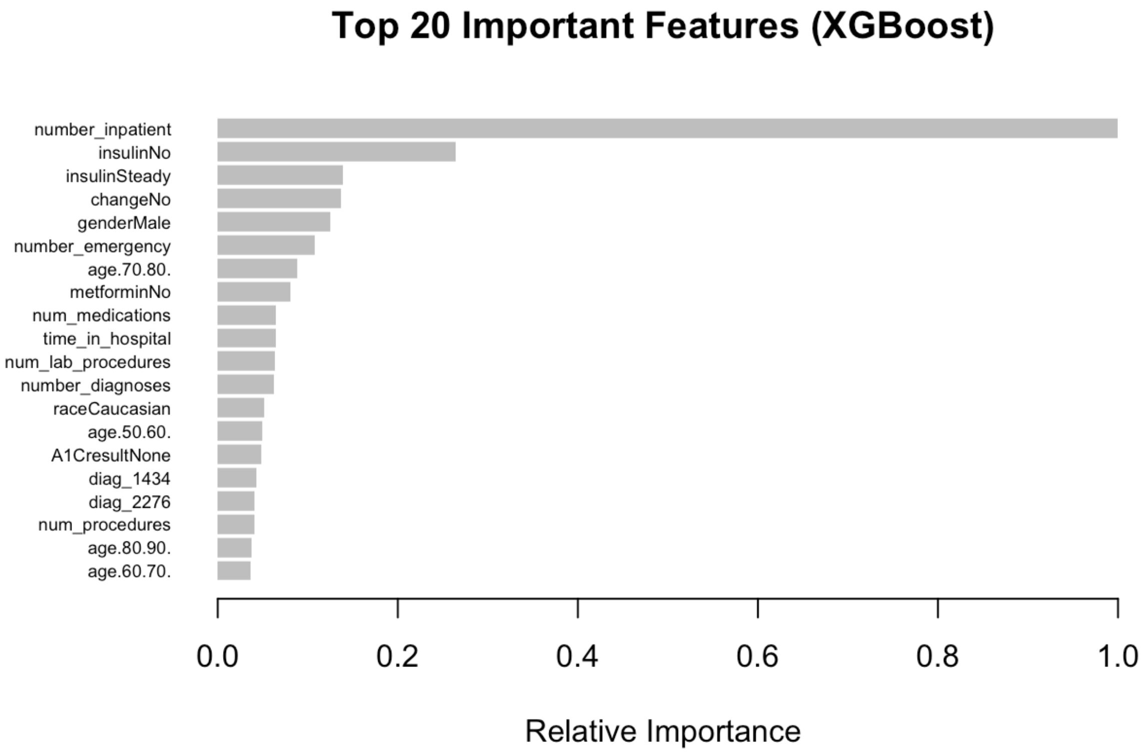

Figure 4: Top 20 Important Features for XGBoost Model

To interpret the XGBoost model's decision-making, I examined feature importance scores that represent how often and how effectively each feature contributed to splitting decisions across the model's trees. The most important features included number_ipatient, number_emergency, and time_in_hospital - all indicators of a patient's recent healthcare utilization. Medication patterns, such as insulinNo, metforminNo, and changeNo, also ranked highly, which suggests that treatment intensity and stability were strong predictors of readmission risk. Demographic indicators such as age group and gender, appeared but were less influential than care utilization factors. Notably, my direct comorbidity flags for hypertension and CKD did not appear amongst the top predictors, implying that how a patient interacts with the healthcare system is more predictive of readmission than diagnosis flags alone.

Conclusion

Changes and Limitations

A few changes and considerations were made throughout the lifetime of this project, the main one being my initial scope. Initially, the project was to be based on building a predictive model that estimated the likelihood of readmission within 30 days for diabetic patients who had either one or both comorbidities of hypertension and CKD. This proved to be too narrow of a scope and would not give us accurate readings on patients themselves, therefore expanding the scope to include all relevant types of features that could attribute to hospital readmission.

As highlighted in the risks section, the largest risks to this project is the utilization of an outdated dataset and ICD diagnosis codes, but other limitations include missing medication adherence and outpatient records, no easily-accessible diabetic-focused datasets to test on, and the introduction of possible synthetic bias by using SMOTE for our class imbalance. These are key things to consider during implementation and may require fine-tuning or further testing on more current data to ensure the integrity of the model.

Final Remarks

Key takeaways from this project include that the ensemble models, Random Forest and XGBoost, outperformed the baseline model, Logistic Regression. XGBoost offered the best balance of recall and precision, and feature importance revealed indirect indicators that contributed to hospital readmissions within 30-days. Class balancing was also important in detecting real patterns related to the readmission within 30-days target. Overall, this final project on predicting hospital readmissions within 30-days amongst diabetic patients provided great experience with applying learned knowledge to build these models and a step towards creating better patient care experiences.

References

Cios, K., Clore, J., DeShazo, J., Gennings, C., Olmo, J., Strack, B., Ventura, S. (2014). “Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records,” BioMed Research International, vol. 2014, Article ID 781670, 11 pages, 2014. https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008