Predicting Hospital Readmission

By Christabelle Gontani, Haley Choe, Shaivi Shah, Sarah Ung, Alicia Huynh

1. Motivation

Unplanned hospital visits create a significant financial and public-health burden. Emergency Room (ER) visits alone cost the U.S. healthcare system billions of dollars annually, and nearly half of the U.S. adults report delaying care due to cost barriers. This creates a reinforcing cycle wherein patients who cannot afford preventative or timely care experience health deterioration that eventually results in expensive, unavoidable ER visits. Hospital readmissions, defined as return visits shortly after discharge, are widely used as a proxy for poor care quality and deteriorating health. Early identification of these high-risk patients can enable targeted interventions.

Our project's goal is to build predictive models that predict and flag patients at high risk of readmission. A well-performing classification model would allow hospitals and insurance to allocate limited care coordination resources more efficiently and prevent avoidable hospitalizations. They could reduce the operational and financial strain on ER departments and improve patients' quality of life, particularly for vulnerable or high-risk populations.

2. Data

2.1 Data Source

We used a Predicting Hospital Readmissions dataset from Kaggle, which records 10 years of hospital admissions involving diabetes diagnoses. Patient information covers age group, length of stay, number of procedures, laboratory tests, medications administered, and prior healthcare use. Clinical context is provided through primary and secondary diagnosis categories and diabetes-specific lab results. The target variable 'readmitted' allows for the dataset to be used for modeling readmission risk. Together, these features offer a structured view of patient characteristics and patterns of healthcare utilization that are relevant for predicting readmission. The distribution of the readmission outcome is shown in Appendix Figure A1.

This dataset was chosen for modeling since it is high-dimensional but structured and it covers long-term clinical utilization patterns. All records in the dataset are de-identified and contain no sensitive patient data. A complete list of variables and definitions is provided in Appendix Table A1.

2.2 Key Descriptives

Our dataset consists of 25,000 diabetes-related hospital admissions, with near-balanced classes *(47% readmitted, 53% not)*. We will not have severe class imbalance issues in our classification models, making our performance metrics more reliable. However, this situation is not often reflected in healthcare datasets in the real world.

Numeric features show that most hospital stays are short *(median 4 days)* but involve substantial lab work and medications *(mean 43.2 'n_lab_procedures', mean 16.3 'n_medications')*. Prior utilization variables are heavily right-skewed, with many patients having no previous inpatient, outpatient or emergency visits. Age distribution skews older, with the majority of patients between 60-80 years old *(51% of our sample)*. The most common primary diagnoses include circulatory, respiratory, digestive and diabetes-related conditions. A1C and glucose testing is performed infrequently *(around 83.8% and 94.5% 'not tested' respectively)*, suggesting that while clinically meaningful, the variable will be less informative due to sparse measurement.

2.3 Preprocessing

We performed several preprocessing steps before modeling with Logistic Regression, CART and Random Forest. We encoded the target variable into a binary indicator (1 = readmitted, 0 = not readmitted) and created an 80/20 stratified train–test split to preserve the outcome distribution. We then separated numerical and categorical features and applied one-hot encoding to all categorical variables, dropping the first category to avoid multicollinearity. All core 16 predictors were consistently included across our models. Encoded columns were aligned across train and test sets and we standardized numerical variables for logistic regression. We also added an intercept term for Statsmodels since it does not automatically add a bias term, and ensured all features were numeric for modeling. Our dataset had very few missing values handled by encoding, wherein "Missing" became a dummy category.

3. Analytics Models

This analysis applied logistic regression, a CART decision tree, and a Random Forest classifier using the same predictor set and preprocessing steps. Categorical variables were one-hot encoded, numerical features were standardized where appropriate, and a stratified train–test split preserved the readmission distribution. Logistic regression served as the baseline model, the decision tree introduced nonlinear structure, and the Random Forest extended the tree approach through ensembling and stratified 5-fold cross-validated hyperparameter tuning. Model performance was evaluated using ROC–AUC, precision, recall, and the confusion matrix to provide a clearer picture of classification outcomes than overall accuracy.

3.1 Logistic Regression

Logistic regression estimated the probability of hospital readmission using the complete predictor set. The target variable was coded as a binary indicator, categorical variables were one-hot encoded, and numerical features were standardized before fitting the model. A stratified split maintained class proportions across the datasets, and the model was trained through maximum likelihood. Evaluation relied on ROC–AUC, precision, recall, and the confusion matrix. The variables that contributed most to the model included prior inpatient, emergency, and outpatient visits, diabetes medication use, and older age categories, while a higher number of procedures showed a small negative relationship with readmission.

The model revealed stable relationships across predictors, but its ability to distinguish between readmitted and non-readmitted patients was limited by the similarity between the two groups. On the test set, the model reached an ROC–AUC of 0.636 with recall at 40% and precision at 62%, indicating only moderate separation between outcome classes. Its specificity of roughly 78% reflects stronger performance in identifying patients who were not readmitted. These results suggest that although the model detects broad utilization patterns tied to readmission, the available predictors do not provide sufficient detail for stronger classification performance. Future work could adjust prediction thresholds for different intervention strategies or incorporate additional patient information not present in the current dataset. The overlap between predicted probabilities for readmitted and non-readmitted patients is shown in Appendix Figure A2, which helps explain the model's moderate level of separation.

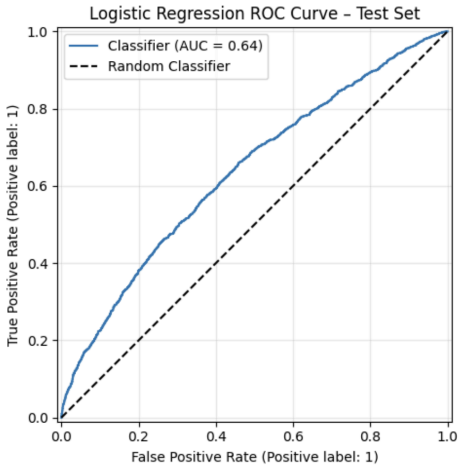


Figure 1. The ROC curve for the logistic regression model on the test set, with an AUC of 0.64.

3.2 CART (Decision Tree)

The CART model used the same predictors and preprocessing steps applied to the other methods, with a stratified split to preserve the distribution of the readmission outcome. Performance was evaluated using ROC–AUC, precision, recall, and the confusion matrix. The decision tree represented nonlinear patterns and interactions among predictors that are not directly reflected in a linear model. Feature importance values indicated that prior inpatient visits had the greatest influence on the tree structure, with smaller contributions from outpatient visits, emergency visits, and the diabetes medication indicator. Appendix Figure A3 presents the full ranking of feature importances for the CART model.

Given their sensitivity to training variation, decision trees offer more limited stability without tuning. The tuned CART model achieved a test ROC–AUC of 0.634, with precision at 0.58 and recall at 0.56, meaning that it performed moderately and relied mostly on a few strong predictors. The analysis could be strengthened through systematic hyperparameter selection or by extending the method through ensemble techniques, which generally produce more reliable generalization.

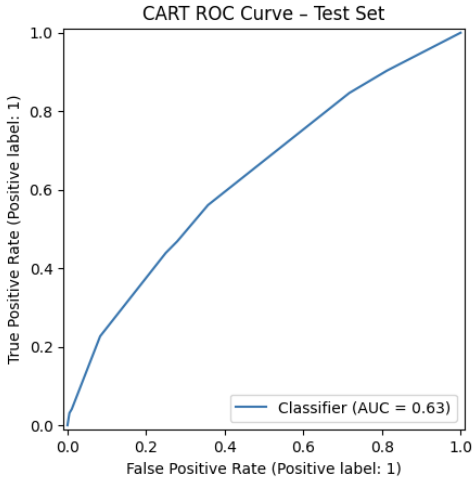


Figure 2. The ROC curve for the CART decision tree on the test set, with an AUC of 0.63.

3.3 Random Forest Model

The Random Forest classifier incorporated the full predictor set using a preprocessing pipeline that one-hot encoded categorical variables while passing numerical variables through unchanged. Training was performed on a stratified split to maintain class balance. Hyperparameters, including the number of trees, depth constraints, and minimum sample requirements, were selected through GridSearchCV with stratified 5-fold cross-validation, using ROC–AUC as the tuning criterion. Model performance was then evaluated using the same metrics applied to the other methods. The variables with the greatest influence on the model included prior inpatient visits, the number of medications, outpatient visits, laboratory procedures, emergency visits, and older age groups. The complete set of feature importances for the Random Forest model is shown in Appendix Figure A4.

By averaging multiple CART trees and incorporating cross-validated tuning, the Random Forest reduced variance and provided greater stability than a single untuned decision tree. On the test set, the model achieved an ROC–AUC of 0.653 with precision and recall near 0.60, indicating that it performed better than the CART model but still misclassified many patients. Further improvements could be achieved by exploring additional hyperparameter configurations or expanding the predictor set to gain a more complete view of factors associated with readmission.

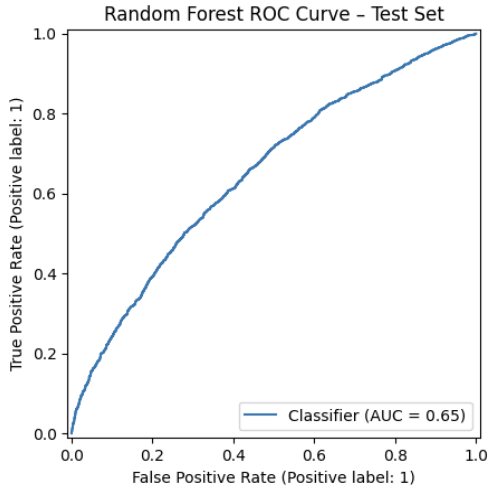


Figure 3. The ROC curve for the Random Forest model on the test set, with an AUC of 0.65.

3.4 Results and Comparisons

Logistic regression offered the clearest interpretation of how individual predictors relate to readmission risk, identifying consistent patterns influenced by prior inpatient, outpatient, and emergency visits. Its performance remained moderate, with an ROC–AUC near 0.64 and a recall of about 40%, reflecting considerable overlap between the two outcome groups. The CART model represented nonlinear structure but was less stable without additional tuning. The Random Forest model, by contrast, incorporated cross-validated hyperparameter selection and the variance reduction gained from combining many decision trees, producing more reliable generalization. Appendix Figure A5 provides a visual comparison of ROC curves across all three models.

The three models feature different strengths rather than in overall outcomes. Logistic regression contributes interpretability, the decision tree reveals interaction-based splits, and the Random Forest provides improved robustness under the same feature constraints. Evaluating these methods together offers a more complete view of the predictive structure in the dataset.

3.5 Conclusion

The analytic models indicate that the predictors in the dataset contain meaningful but limited information for identifying patients at higher risk of readmission. Logistic regression clarified how individual variables relate to the outcome, while the decision tree and Random Forest captured additional nonlinear structure, with the ensemble approach showing the most stable generalization due to cross-validated tuning. Although none of the models achieved strong separation between readmitted and non-readmitted patients, evaluating them together reveals where the current features are informative and where the data impose constraints. Expanding the predictor set or applying more extensive model selection procedures would likely improve predictive performance and support a more detailed assessment of readmission risk.

4. Impact and Future Work

A deployed model like ours could support targeted, preventive care by allowing hospitals or insurers to proactively reach out to patients flagged as high-risk, assign care coordinators or transitional-care programs, and provide tailored patient education.

However, deployment also raises important ethical and operational considerations. False negatives, which refer to patients incorrectly classified as low-risk, may not receive necessary follow-up, potentially resulting in preventable readmissions. Conversely, false positives could lead to unnecessary interventions, straining healthcare resources and placing undue burden on patients. Additionally, because the dataset lacks socioeconomic, racial, or ethnicity variables, the model may systematically under- or over-predict risk for certain subpopulations. Without explicit fairness evaluation, biases could inadvertently amplify existing healthcare disparities. Ensuring that the variables used in predictive models are transparent and interpretable is critical for clinician trust and responsible use in decision-making.

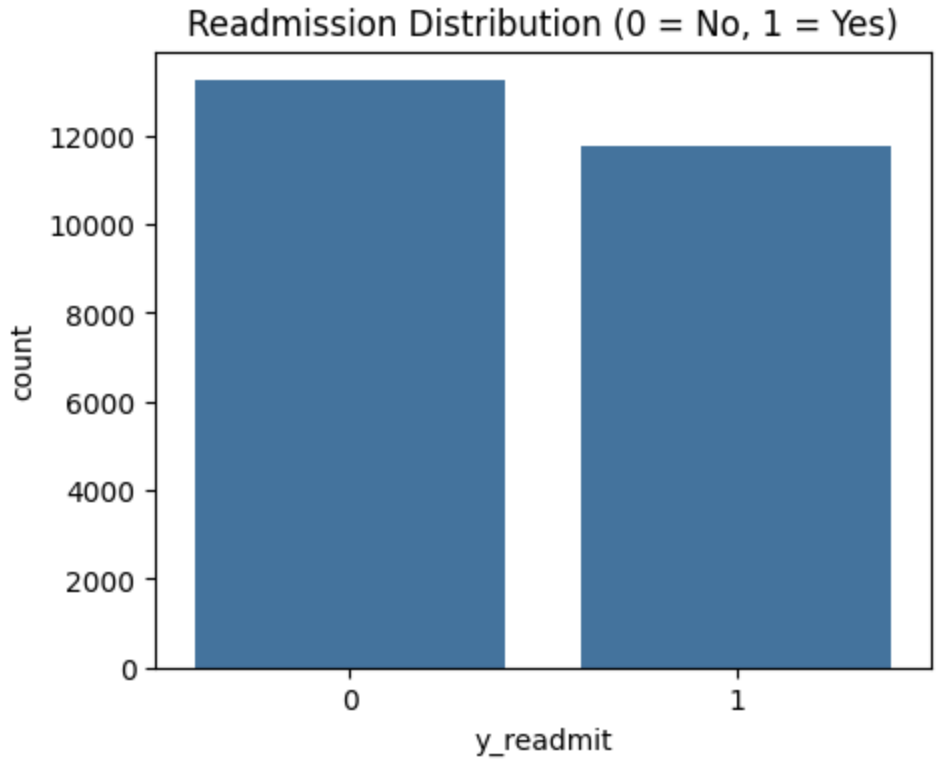
For future work, we could use prior hospitalization history over time for temporal modeling, using variables like sequence data and time since last discharge, to capture dynamic risk. If demographic or socioeconomic data becomes available, rigorous fairness assessments should be conducted and fairness-aware modeling approaches considered. Beyond this, probability thresholds should be calibrated to real-world resource constraints and intervention costs, rather than relying on a default 0.5 cutoff. Any clinical deployment should involve close collaboration with practitioners to refine feature sets, validate predictions, and ensure that the model aligns with ethical standards and the practical realities of hospital care delivery.

5. Appendix

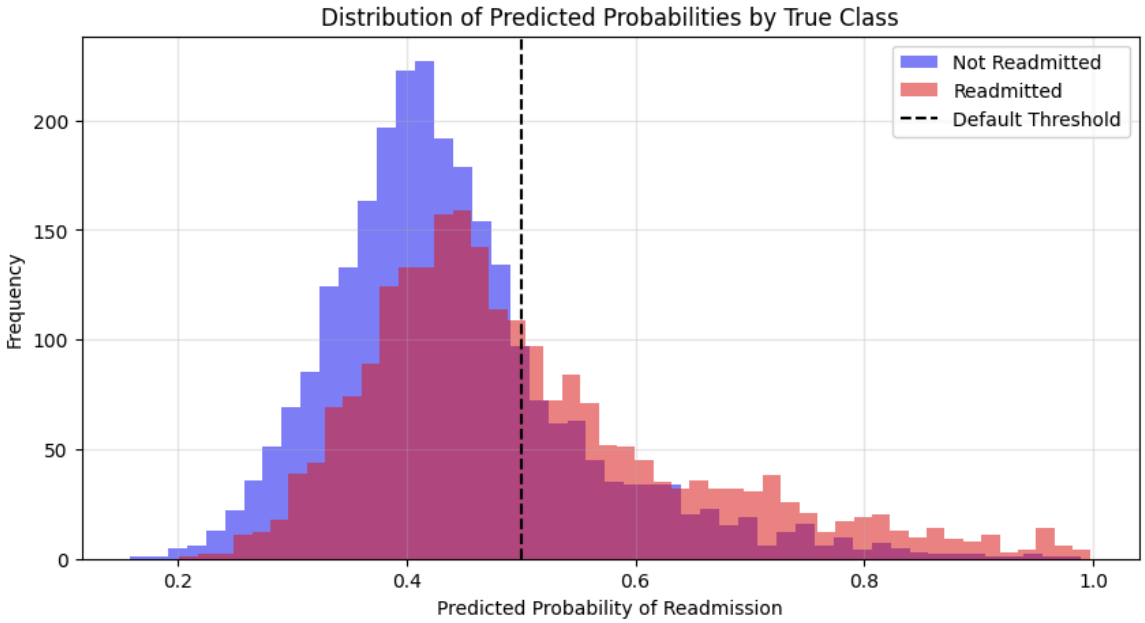
View code for project here.

Feature	Description
age	Age bracket of the patient
time_in_hospital	Length of stay in days (1–14)
n_procedures	Number of procedures performed during the hospital stay
n_lab_procedures	Number of laboratory procedures performed during the hospital stay
n_medications	Number of medications administered during the hospital stay
n_outpatient	Number of outpatient visits in the prior year
n_inpatient	Number of inpatient admissions in the prior year
n_emergency	Number of emergency room visits in the prior year
medical_specialty	Specialty of the admitting physician
diag_1	Primary diagnosis category (Circulatory, Respiratory, Digestive, etc.)
diag_2	Secondary diagnosis category
diag_3	Additional secondary diagnosis
glucose_test	Glucose serum result: high (>200), normal, or not performed
A1Ctest	A1C level result: high (>7%), normal, or not performed
change	Whether diabetes medication was changed (yes/no)
diabetes_med	Whether diabetes medication was prescribed (yes/no)
readmitted	Whether the patient was readmitted (yes/no)

Appendix Table A1. 'hospital_readmissions.csv' Variable Descriptions



Appendix Figure A1. A bar chart showing the number of patients who were readmitted versus not readmitted, illustrating the near-balanced outcome.



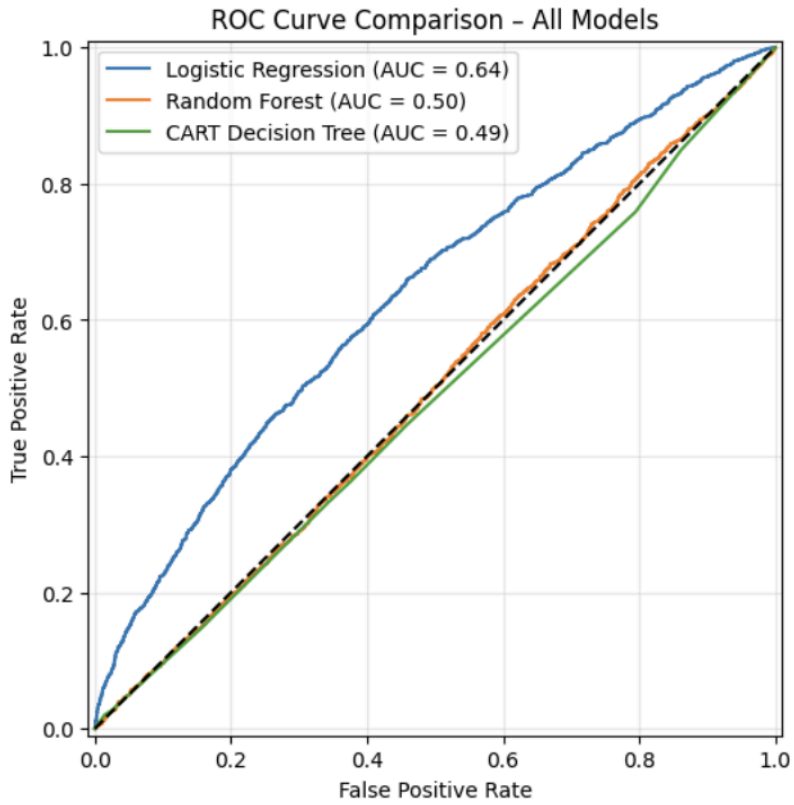
Appendix Figure A2. A histogram of predicted readmission probabilities for each true class, showing substantial overlap between readmitted and not-readmitted patients.

```
Top 20 features by importance:
                                    feature  importance
52                               n_inpatient    0.807468
51                              n_outpatient    0.087709
53                               n_emergency    0.044366
45                            diabetes_med_no    0.039782
1                                 age_[50-60)    0.020675
0                                 age_[40-50)    0.000000
6               medical_specialty_Cardiology    0.000000
7         medical_specialty_Emergency/Trauma    0.000000
8   medical_specialty_Family/GeneralPractice    0.000000
9         medical_specialty_InternalMedicine    0.000000
10                medical_specialty_Missing    0.000000
11                  medical_specialty_Other    0.000000
12                medical_specialty_Surgery    0.000000
13                       diag_1_Circulatory    0.000000
14                          diag_1_Diabetes    0.000000
15                        diag_1_Digestive    0.000000
16                           diag_1_Injury    0.000000
17                          diag_1_Missing    0.000000
2                                 age_[60-70)    0.000000
3                                 age_[70-80)    0.000000
```

Appendix Figure A3. The top twenty features ranked by importance in the CART model, with inpatient visits as the most influential predictor.

```
Top 20 Feature Importances:
                                feature  importance
52                          n_inpatient    0.242683
50                       n_medications    0.086125
51                          n_outpatient    0.080630
48                     n_lab_procedures    0.077590
53                          n_emergency    0.058345
47                     time_in_hospital    0.050540
49                         n_procedures    0.036501
45                      diabetes_med_no    0.013430
46                     diabetes_med_yes    0.012491
4                           age_[80-90)    0.011924
19                         diag_1_Other    0.011700
3                           age_[70-80)    0.011573
10            medical_specialty_Missing    0.011475
1                           age_[50-60)    0.010893
35                         diag_3_Other    0.009859
9   medical_specialty_InternalMedicine    0.009809
21                    diag_2_Circulatory    0.009687
11              medical_specialty_Other    0.009666
29                    diag_3_Circulatory    0.009604
13                    diag_1_Circulatory    0.009511
```

Appendix Figure A4. The top twenty features ranked by importance in the Random Forest model, led by inpatient visits, number of medications, outpatient visits, and lab procedures.



Appendix Figure A5. A comparison of ROC curves for logistic regression, Random Forest, and CART to show differences in model performance.