MA 679 Final Project

Jianing Yi
Chenxuan Xiong
Yuchen Huang

Abstract

This study explores the efficacy of various machine learning models in predicting 30-day hospital readmissions using comprehensive patient data from 2018 to 2020. Focusing on tracheostomy and mastoditis cases, we implemented rigorous data preprocessing, feature selection, and modeling techniques including logistic regression, Lasso regression, gradient boosting, and neural networks to address the challenge of predicting readmissions. We tried multiple methods to deal with the imbalance including resampling, weight class. The models' effectiveness was assessed based on various metrics including accuracy, AUC, and precision, which revealed predictive capabilities under imbalance condition. This project not only contributes to the ongoing efforts in healthcare analytics but also underscores the complexities and nuances involved in predicting hospital readmissions.

Introduction

Addressing the challenge of 30-day hospital readmissions is critical for improving patient care and reducing healthcare costs. Our project utilizes advanced machine learning techniques and comprehensive patient data from 2018 to 2020 to develop predictive models. Focusing on high-risk procedures and diagnoses, such as tracheostomy and mastoditis, we aim to identify key factors that influence readmission risks. This report details our approach, from data preprocessing to deploying several predictive models, including logistic regression and neural networks, to enhance hospital management and patient outcomes.

Objective

We aim to develop a model that predicts whether a patient will be readmitted within 30 days, using all available patient information as input to forecast readmission likelihood.

Method

(Detailed explanation of how you did the work. Your sample selection, your choice of the model, how you tuned the model, etc.)

- **Data Cleaning**
  - **Filtering and merging**

In our analysis, we use data from 2018 and 2019 as our training set and 2020 as our test set. We mainly focus on two procedures (laryngectomy and tracheostomy) and one diagnosis (mastoditis). We filtered these treatments by searching target ICD-10 codes in procedure and diagnosis columns and selecting rows with any of our target code. After filtering, we have nine datasets (3*3) for three years' laryngectomy, tracheostomy and mastoditis. And we merged the hospital and severity data into core data for more information.

- **30-days readmission indicator:**
  To calculate the readmission events of patients to the hospital, I first compute the interval between two admissions for the same patient. This is done by calculating the difference between the two occurrences of 'NRD_DaysToEvent' for each patient, and then subtracting the duration of the hospital stay (denoted as 'loss'). According to the definition of a readmission, if this interval value is thirty days or less, the readmission value (denoted as readmit_flag) corresponding to the second admission is marked as 1, indicating a readmission event. Conversely, if the interval exceeds thirty days, the readmission value is set to 0. Additionally, if a patient is lost to follow-up after discharge, the readmission value for that admission is directly recorded as zero.

- **Melting**
  To investigate the effect of comorbidity, we selected 10 diagnoses and procedures that have the largest difference between readmit and non-readmit rate and melt them into new columns with binary indicators.

- **Preliminary data and feature selection**
  In our filtered data, we found there are no 30-days readmission for laryngectomy. We did our analysis on tracheostomy and mastoditis.
  Columns we dropped manually:
  1. Columns contain keys for identification.
  2. Original records for diagnosis and procedure (DX and PR)
  3. Number of day from admission to procedure (PRDAY)
  4. The columns that are in 2020 and 2019 but not in 2018.

- **EDA**
  - We used EDA to meticulously examine and identify some compelling features within our selected procedures.
  - **`ZIPINC_QRTL`:**

We linked the variable `ZIPINC_QRTL`, which represented the patients' income divided into quartiles, to the death rate to gain a clearer perspective on their relationship.

- **ICD 10 Code:**
  We calculated the proportions of readmitted and non-readmitted patients across different ICD 10 codes. The ICD-10 codes serve as standardized classifications for both diagnoses and medical procedures. From this analysis, we identified ten diagnosis codes that exhibited the largest disparities between readmission and non-readmission rates. This approach helped us pinpoint the most critical diagnoses affecting patient readmissions, allowing for targeted investigations into the factors influencing these outcomes.

- **Modeling:**
  - **Logistic regression**
    - Oversampling:
      Given the extreme imbalance in the data, we implemented an oversampling technique to achieve a more balanced distribution between the classes. This method enhances the robustness of our analysis by equalizing the number of observations across different categories.
    - In addition to analyzing the ten ICD-10 codes, we manually selected additional features to incorporate into our model.
    - For our predictive model, any predictions with a value of 0.5 or higher are classified as readmissions, while predictions below 0.5 are classified as non-readmissions.

  - **Lasso Logistic Regression**
    - Similar to logistic regression, with calculated best lamda.

  - **Gradient Boosting Classifier**
    - Oversampling: To address the significant data imbalance, we applied an oversampling technique, which effectively equalized the distribution of observations among different classes.
    - To tune the parameters, we divide the data into training sets and validation sets. We measure the model based on the classification report where we compare the f1- score and ROC AUC scores since these indicates the performance of the model when the dataset is extremely unbalanced.

        - We deploy a threshold of 0.5 to do the classification. A probability higher than 0.5 will be classified as readmissions, otherwise, it will be considered non-readmissions.

    - **NN**
        - Feature selection：
          We used feature selection based on univariate chi-square analysis to select 40 most significant features.
        - Structure:
          For Tracheostomy
            - 1 input layer accept data with a certain number of features
            - 6 Dense layer contained 64 neurons with ReLu activation function
            - 3 dropout layer with 0.5 dropout rate
            - 1 output layer contains 1 neurons with sigmoid activation function and bias

          For Mastoditis:
            - 1 input layer accept data with a certain number of features
            - 2 Dense layer contained 128 neurons with ReLu activation function and 1 Dense layer contained 32 neurons
            - 2 dropout layer with 0.5 and 0.7 dropout rate
            - 1 output layer contains 1 neurons with sigmoid activation function and bias

          The bias is defined as: log(positive/negative)
        - Training:
            - The model is trained for 100 epchos with batch size 2048. Large batch size makes sure each batch contains at least one positive sample.
            - Add early stopping monitors validation accuracy to avoid overfitting.
            - Train the model with a class weight to deal with the imbalance problem.
                - The weight for class 0: (1 / negative) * (total / 2)
                - The weight for class 1: (1 / positive) * (total / 2)

- **Evaluation Matric**
    - **Accuracy:** Accuracy indicates the proportion of correct predictions among the total predictions made by a model. It shows how often the model's predictions match the true labels. But when the dataset has a significant imbalance between classes, accuracy can be misleading.

- **AUC:** AUC considers all possible thresholds, providing a holistic view of a model's performance. And gives a sense of a model's ability to balance true positive rate and false positive rate.
- **Precision:** Precision measures the accuracy of positive predictions. It's crucial when the positive class is rare, and the goal is to ensure the accuracy of predictions for that class. Also a false positive has significant cost.

Results

Readmission Prediction for Mastoditis

Model	Logistic Regression	Lasso Logistic Regression	Gradient Boosting Classifier	CNN
Accuracy	0.0847	0.0833	0.6799	0.9031
AUC	0.6056	0.6045	0.6547	0.5819
Precision	0.037	0.037	0.0587	0.0599

Readmission Prediction for Tracheostomy

Model	Logistic Regression	Lasso Logistic Regression	Gradient Boosting Classifier	CNN
Accuracy	0.6719	0.6353	0.83	0.9730
AUC	0.6644	0.6411	0.7369	0.6905
Precision	0.00424	0.00391	0.00758	0.0134

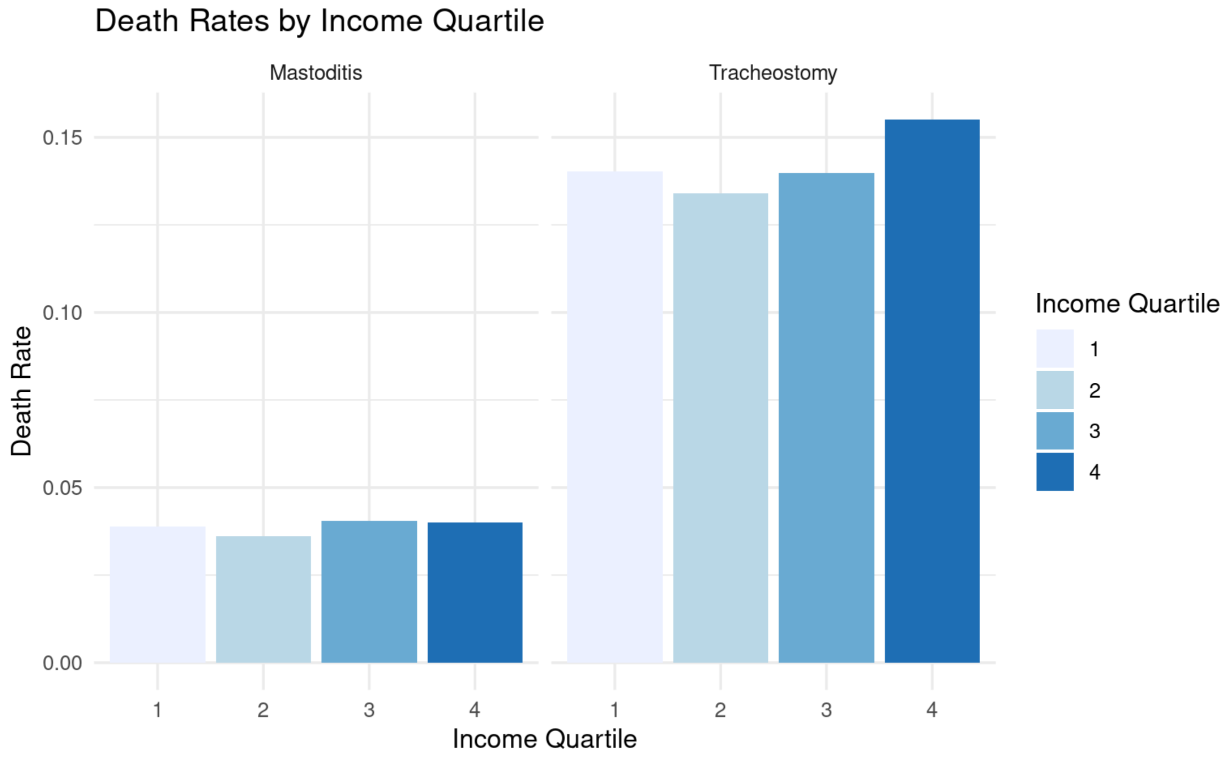
Conclusion

- Despite our efforts, none of the models we tested effectively predicted readmission rates due to their low accuracy and precision. However, the Gradient Boosting Classifier emerged as the most effective model, demonstrating the highest levels of accuracy, precision, and stability among all the models we evaluated.
- For an unbalanced dataset, we may consider the Gradient Boosting Classifier model. Besides, we should consider resampled techniques, for instance, we can address the class imbalance directly by either oversampling the minority class. Techniques such as Oversampler can be used to enhance the performance of the model.
- Limitations:
  Although we utilized both automatic and manual feature selection techniques to identify suitable predictors, the residual plots reveal that our models may still be missing some underlying trends. Given this situation, exploring causal inference methods could potentially address these shortcomings and improve our model's performance.

Appendix

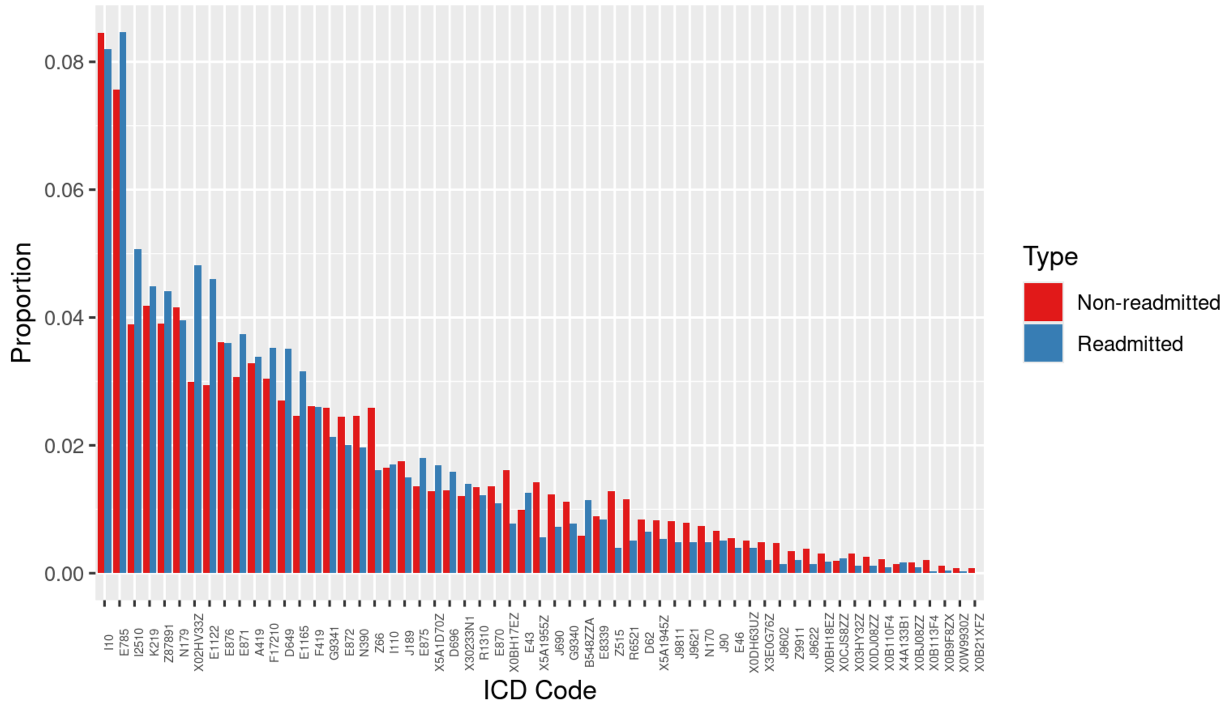
Everything that does not fit in the main report including your many EDA figures, model check figure, tables, code outputs, etc.

EDA

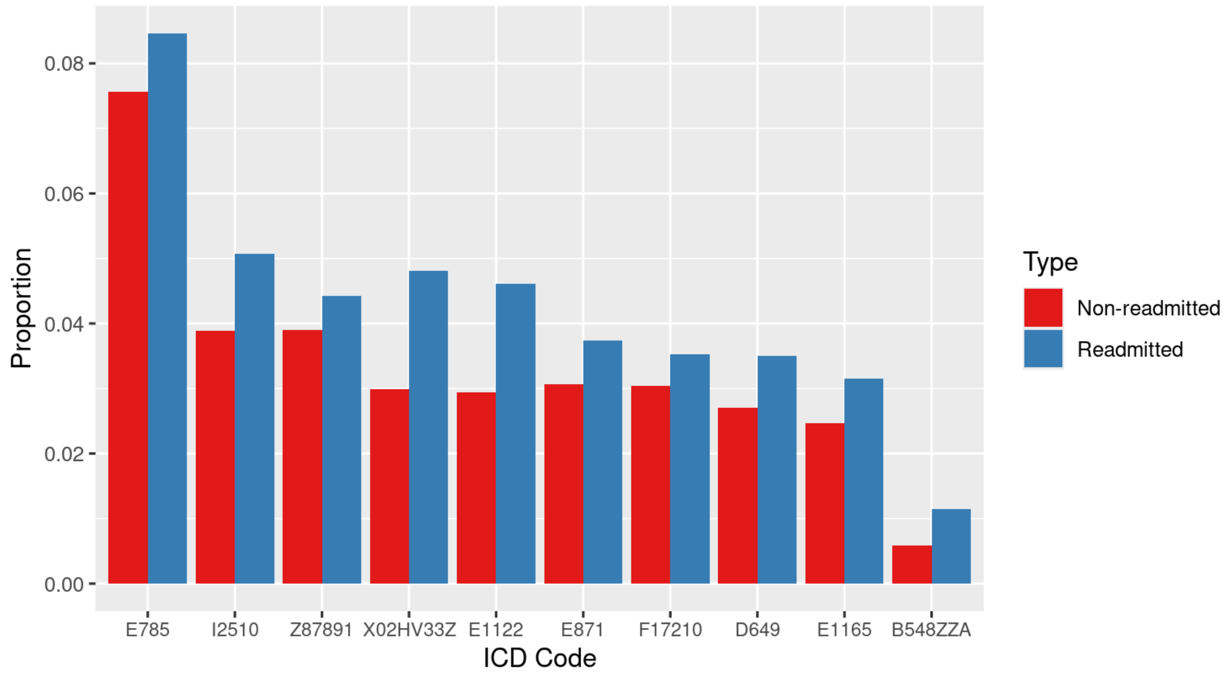


Analysis shows that the death rate does not significantly differ across various income quartiles. This lack of variation suggests that `ZIPINC_QRTL`, which categorizes patients by income, may not be a reliable indicator for assessing the risk factors associated with mortality rates.

ICD 10 Code Proportion for Mastoditis



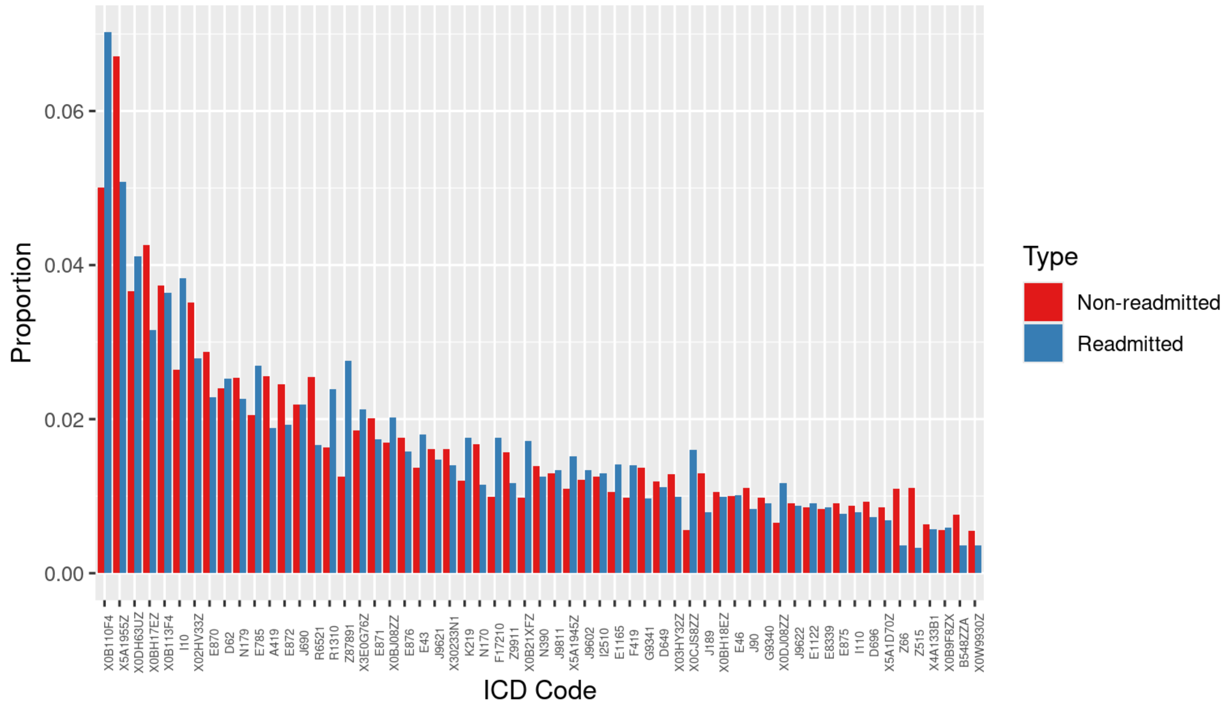
ICD 10 Code Proportion difference for Mastoditis



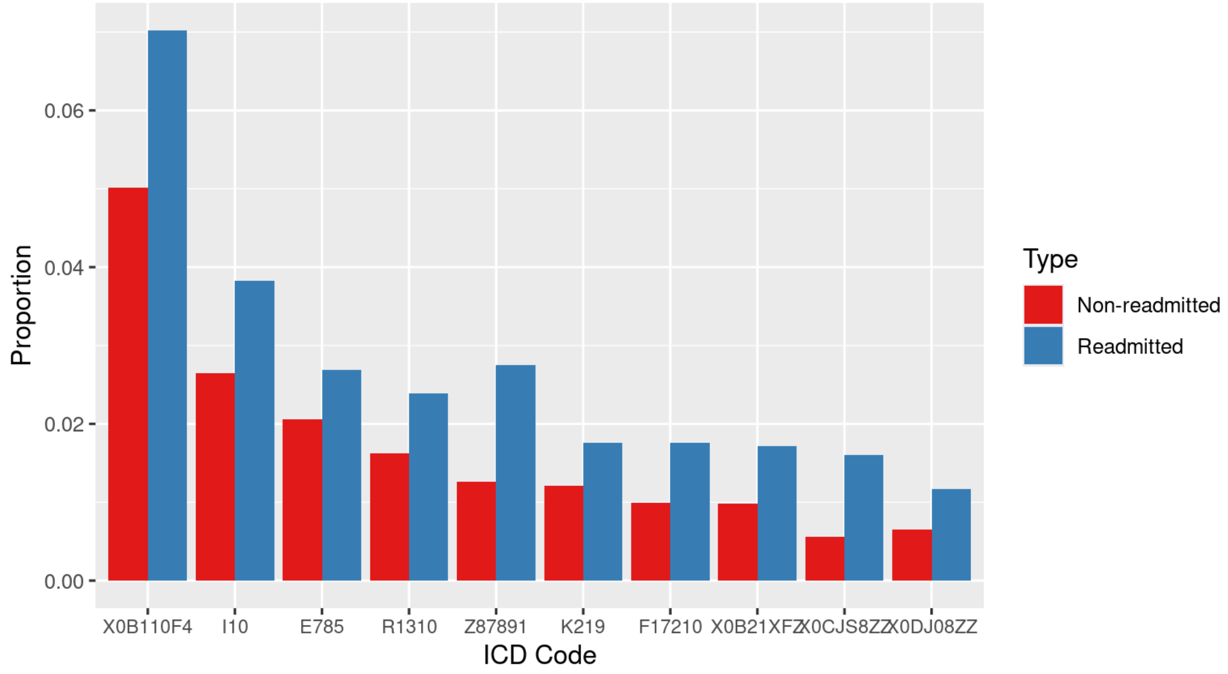
The difference is calculated by `readmitted proportion` - `non-readmittedproportion`

```
[1] "For Mastoditis, The top 10 ICD10 codes of proportion difference between readmitted and
non-readmitted are: X02HV33Z, E1122, I2510, E785, D649, E1165, E871, B548ZZA, Z87891,
F17210."
```

ICD Code Proportion for Tracheostomy



ICD 10 Code Proportion difference for Tracheostomy



The difference is calculated by `readmitted proportion` - `non-readmitted proportion`

```
[1] "For Tracheostomys, the top 10 ICD10 codes of proportion difference between readmitted
and non-readmitted are: X0B110F4, Z87891, I10, X0CJS8ZZ, F17210, R1310, X0B21XFZ, E785,
K219, X0DJ08ZZ."
```

Acknowledgement

We extend our heartfelt thanks to Dean and Anand for their generous contribution of HCUP data, which was instrumental in the completion of this research. Their willingness to share their resources and expertise has greatly enhanced the quality and depth of our work. We are truly grateful for their support.

Attribution

Clearly state who did what.

- Yuchen Huang:
    - EDA
    - Model: Logistic, Lasso
- Chenxuan Xiong:
    - Preprocessing: Filtering and merging, Melting and Preliminary feature selection
    - Model: NN
- Jianing Yi:
    - Preprocessing: Adding 30-days readmission indicator
    - EDA
    - Model: Gradient Boost