香 港 大 學

THE UNIVERSITY OF HONG KONG

Statistics Course

STAT 3612: Statistical Machine Learning

Group Project Proposal

30-day All-Cause Hospital Readmission Prediction

Group Member:

Yao Zixuan 3035845148

Wang Ziyu 3035777547

Fang Jiahe 3035772482

Huang Yining 3035662522

Huang Zixun 3035844522

1. Objective

The primary goal of the research proposal is to predict hospital readmission rates by leveraging the Electronic Health Records (EHRs) data available in the MIMIC-IV v1.0 database (Johnson et al., 2023). The proposed task revolves around a binary classification problem, with the target variable denoting the readmission state (Y/N), and a well-predicted result can optimize hospitals' medical resource allocation.

2. Proposed Variables

After research on data types, data formats, and special data entries such as sequential medical inspection information on one patient, we have categorized the variables into target variable and predictor variables.

2.1 Target Variable

The readmission state is the target variable of this project. The output of the target variable can be 'Yes' or 'No'. In the context of dataset, if a patient was admitted three times to the hospital the readmission status is 'Yes' for the first two admission records, and the readmission status is 'No' for the last admission. Additionally, irrespective of prior admission instances, we take the mortality at hospital as a readmission occurrence and label it as 'Yes'.

2.2 Predictor Variables

We select an initial set of 174 features from demographics, comorbidities, laboratory test results, and medications administered during hospitalizations. We further propose incorporating the predictor variables that are 'Length of Saty' and 'Number of Previous Admissions'.

The "Length of Stay" variable represents the duration between the admission and discharge times of a patient. A longer length of stay may serve as an effective indicator of a more severe condition, potentially influencing the likelihood of readmission. By including this variable, we aim to investigate the relationship between length of stay and the probability of readmission.

The "Number of Previous Admissions" variable denotes an aggregated count of the times a patient has been readmitted. It assumes that numbers of previous admissions would affect readmission rate. By considering this variable, we aim to explore the influence of prior readmissions on the likelihood of future readmissions.

3. Exploratory Data Analysis

In this research our Exploratory Data Analysis (EDA) mainly consists of two parts, data preprocessing and data visualization.

In data preprocessing, we will perform missing value imputation and feature scaling to process the raw data. Additionally, noticing the class imbalance across multiple features, we will utilize synthetic minority over-sampling technique (SMOTE) and weight adjustment to solve the class imbalance problems (Glemaitre, 2023).

In data visualization part, we will apply various visualization techniques to depict the distribution of variables. Pairwise correlation analysis will also be performed to examine the relationships between variables.

4. Feature Engineering

We will consider the meanings of different features and choose appropriate metrics (mean, median, or latest value) for time series sensors accordingly. Constant-value features and highly correlated variables will be excluded to address multicollinearity. Moreover, we will use methods like Principal Component Analysis (PCA), forward/backward stepwise selection, or Lasso regression to reduce dimensionality and select significant features. In addition, we will employ feature engineering techniques such as polynomial regression, step functions, regression splines, and smoothing splines to capture complex relationships within the data.

5. Model Selection

Various classification algorithms will be employed, including Logistic Regression, Discriminant Analysis, Naive Bayes, K-Nearest Neighbors, Support Vector Machine, Classification Tree, Bagging, random forests, Boosting (XGBoost, Gradient Boosting, and Adaboost), and deep learning neural network models such as RNN and transformer (Simhayev, Rasul, & Rogge, 2023). We will use hyperparameter optimization techniques like random/grid search and automated hyperparameter tuning like Bayesian optimization to improve model performance (Pandian, 2022). Regularization techniques such as L1/L2 regularization and decision tree pruning will also be applied to prevent overfitting and enhance generalizability. Model assessment techniques like cross-validation (preferably 5 or 10-fold CV) will be used.

6. Performance Analysis

The performance of the models will be evaluated using appropriate metrics, such as accuracy, precision, recall, and F1 score. Additionally, the Area Under the Curve (AUC) metric will be utilized to assess the models' ability to discriminate between readmission and non-readmission instances. AUC values greater than 0.8 will be considered acceptable.

Reference:

Glemaitre, Imbalanced Learn, (2023), *GitHub Repository*, https://github.com/scikit-learn-contrib/imbalanced-learn

Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L. A., & Mark, R. (2023). MIMIC-IV (version 2.2). *PhysioNet*. https://doi.org/10.13026/6mm1-ek67.

Pandian, S. (2022, February 22). A Comprehensive Guide on Hyperparameter Tuning and Its Techniques. *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2022/02/a-comprehensive-guide-on-hyperparameter-tuning-and-its-techniques/

Simhayev, E., Rasul, K., & Rogge, N. (2023, June 16). Yes, Transformers are Effective for Time Series Forecasting (+ Autoformer). *Hugging Face*. Retrieved from https://huggingface.co/blog/autoformer