Final Presenation-
Prognosis of Diabetic Readmission

Yian Chen
So Man Amanda Au-Yeung
Chun-Wei Tseng

Background & Motivation

- 11.6% of US population had diabetes (2021)
  - Diagnosed: 29.7 million
  - Not been Diagnosed: 8.7 million
- Many diabetic patients still do not receive preventive and therapeutic interventions.
- Diabetic patients can be hospitalized due to infection, stroke, and electroyte disorder, etc.
- Datset Source: UCI Machine Learning Repository

Objectives

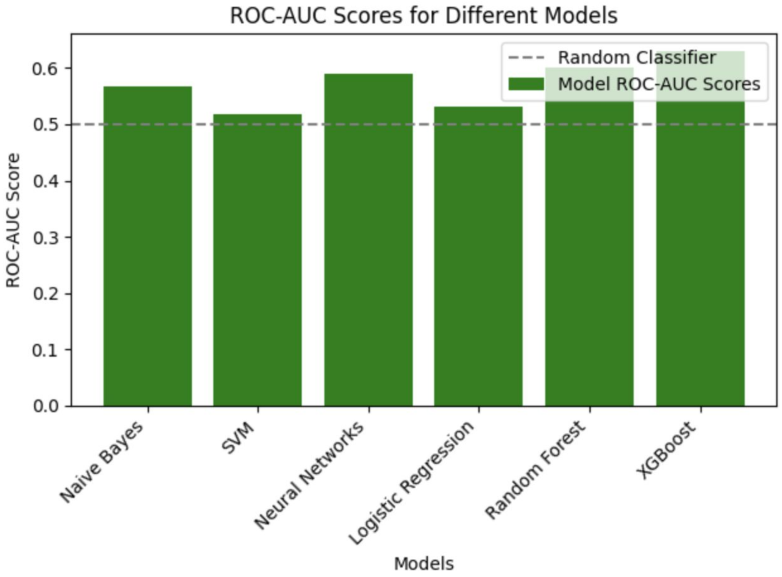
- Predicting combination of features that lead to early hospitalization readmission within 30 days.
- Dataset: features VS target (readmission < 30 days)
- Comparing the accuracy of predictions using different methods
  - Naive Bayes
  - Support Vector Machine (SVM)
  - Neural Network
  - Logistic Regression
  - Random Forest
  - XGBoost

Preprocessing

- Dropped columns with high proportion of missing values
- Dropped medication columns with low correlation to y label and those that have less than 100 patient the medication
- Dropped secondary and additional secondary (diag_2, diag_3) diagnosis
- One-hot encoding and label-encoding categorical features
- Drop rows with NaN values
- Data normalization

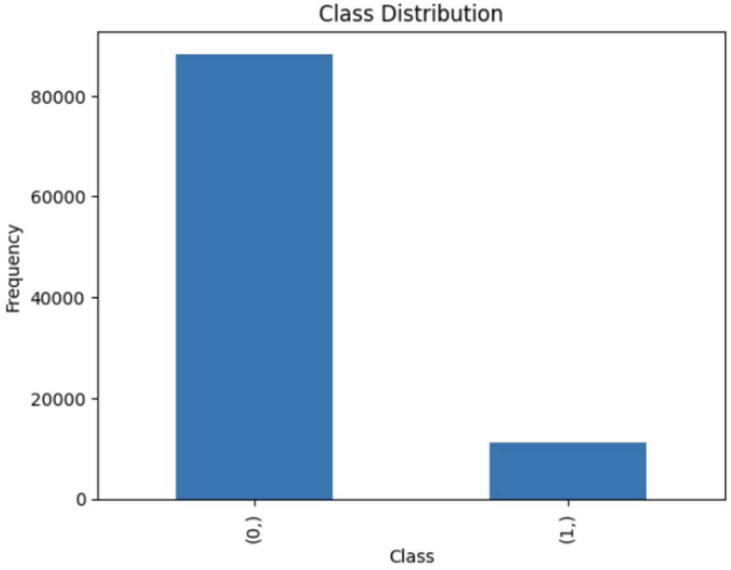
Comparison

- Accuracy:
  - Naive Bayes: 88.76%
  - SVM: 88.76%
  - Neural Networks: 88.76%
  - Logistic Regression: 88.54%
  - Random Forest: 89%
  - XGBoost: 88.78%
- AUC for ROC (Receiver Operating Characteristic):
  - Naive Bayes: 56.75%
  - SVM: 51.87%
  - Neural Networks: 59%
  - Logistic Regression: 51%
  - Random Forest: 60%
  - XGBoost: 62.93%



Balancing Datasets

- The original dataset has imbalance target values
- The majority dataset are negative results - 88.32%
- The minority dataset are positive results - 11.22%
- SMOTE (Synthetic Minority Oversampling Technique)
- Random Undersampling



Conclusion

- Imbalance in data can lead to low in ROC-AUC
- SVM bias towards the majority class in the datasets
- With unbalance data, XGBoost tends to have better generalization capability and less likely to have overfit
- When undersampling our dataset to test, the ROC-AUC of ther models increased but ROC-AUC for XGBoost decreased

Future Direction

- Better updated data, since the dataset is from 1999 - 2008
- Better way to balance data

Citation

- Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore, “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://www.aapc.com/codes/code-search/
- https://keras.io/api/layers/regularization_layers/dropout/
- Han, Jiawei, Jian Pei, and Micheline Kamber. "Data Mining: Concepts and Techniques." Data Mining in Time Series Databases, edited by Mark Last, Abraham Kandel, and Horst Bunke, World Scientific Publishing Co., 2005, pp. 1-34. KEEL Dataset Repository, https://sci2s.ugr.es/keel/keel-dataset/pdfs/2005-Han-LNCS.pdf
- https://www.cdc.gov/diabetes/health-equity/diabetes-by-the-numbers.html