Predicting 30-Day Readmissions in Diabetic Patients Using Ensemble Learning with AutoGluon

Baijiang (Noke) Yuan
Institute of Medical Science
University of Toronto
baijiang.yuan@mail.utoronto.ca

April 10, 2025

Table of Contents

Introduction

Background: Diabetes and Readmission

Diabetes readmissions are common, costly, and often preventable.
Patients with severe dysglycemia face high risks of recurrent hospitalizations within 30 days.[a]

References: [a] Mayo Clinic News Network

Background: Burden of Readmission in Diabetes

- Diabetes is a chronic condition marked by impaired regulation of blood glucose due to insulin deficiency or resistance.[a]
- Diabetic patients are nearly twice as likely to be readmitted within 30 days of discharge, with rates ranging from 14.4% to 22.7%.[b,c]
- Hospital readmission is a key quality metric and contributes substantially to healthcare costs.[d]
- In the U.S., diabetes-related hospital costs exceeded $124 billion in 2012, with 30-day readmissions alone accounting for at least $20 billion.[e]
- For diabetic foot ulcers, readmitted patients incur nearly 3x higher costs than non-readmitted patients.[g]

References: [a]Ojo et al. (2023), [b]Rubin (2015), [c]Gregory et al. (2018), [d]Kassin et al. (2012), [e]Karunakaran et al. (2018), [f]Kum Ghabowen et al. (2024), [g]Hicks et al. (2019)

Related Works

- Hai et al. (2023) used LSTM networks on longitudinal EHR data from 36,000+ diabetic patients, achieving AUROC of 0.79 and outperforming traditional models like RF and LR.
- Liu et al. (2024) evaluated 11 models on the UCI diabetes dataset, finding that ensemble methods (RF, XGBoost) with Grey Wolf Optimizer (GWO) for feature selection outperformed DL models.
- Shang et al. (2021) tested RF, Naïve Bayes, and Tree Ensemble on the same dataset, with RF achieving AUROC of 0.661 and SMOTE improving recall in imbalanced data.

Problem Statement and Study Objectives

Problem Statement:
This project aims to develop predictive models to determine the risk of a diabetic patient being readmitted within 30 days of discharge.

The objectives of this study are:

- To develop and evaluate an AutoML-based ensemble framework (AutoGluon) for predicting 30-day readmissions in diabetic patients.
- To compare the performance of ensemble methods with traditional ML models, the DL model, and the foundation model for tabular data.
- To analyze the impact of different feature sets and preprocessing strategies on predictive performance.
- To identify key risk factors contributing to readmission and evaluate their clinical relevance.
- To assess the generalizability of the model through sub-group analysis.

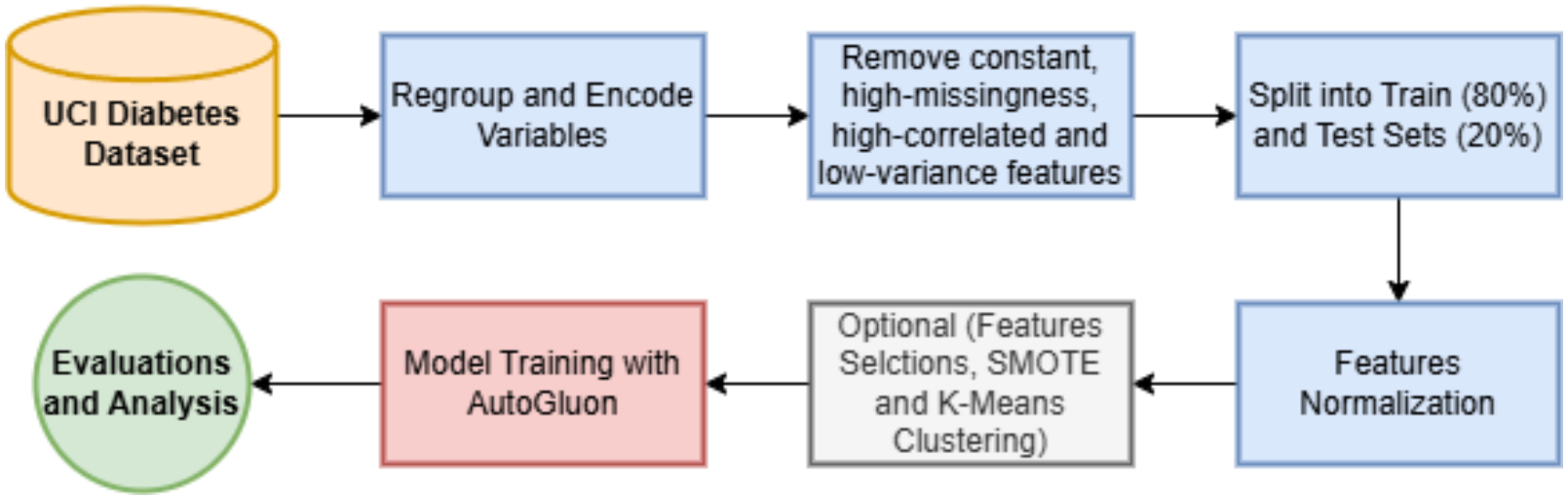
Methods

Cohort Overview

- Dataset: UCI Diabetes 130-US Hospitals dataset[a].
- Population: 100K+ encounters from 70K+ diabetic patients.
- Key features: demographics, diagnoses, lab results, and admission data.
- Outcome: Binary label for 30-day readmission.

References: [a]UCI Diabetes 130-US Hospitals Dataset (2014)

Data Preprocessing



- To reduce correlation bias from multiple encounters, only the encounter with the longest hospital stay per patient was kept, reducing the dataset from 101,766 to 71,518 rows. Variables were regrouped and encoded; features with constant values, high missingness (>0.8), high correlation (>0.9), or low variance (<0.1) were dropped.
- **Two feature settings**: (1) top 10 features only; (2) original features plus K-means cluster labels. **Three preprocessing strategies**: addressed class imbalance using Synthetic Minority Over-sampling Technique (SMOTE), Adaptive Synthetic Sampling (ADASYN), and random undersampling.
- Hyperparameters were optimized using Bayesian optimization.

Example: Grouping of Diagnostic Codes

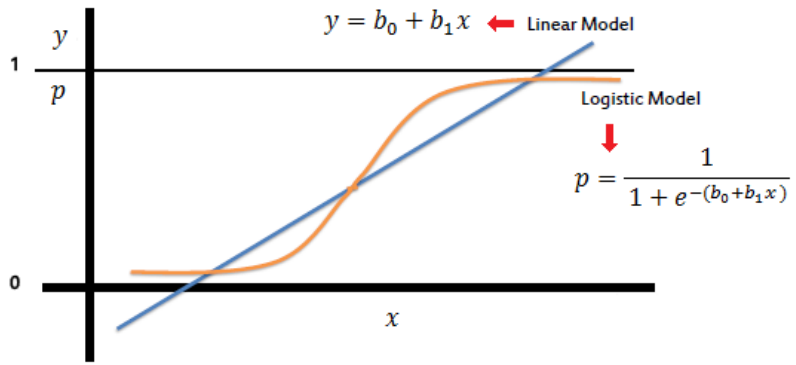
Group Name	**ICD-9 Codes**
Circulatory	390–459, 785
Respiratory	460–519, 786
Digestive	520–579, 787
Diabetes	250.xx
Injury	800–999
Musculoskeletal	710–739
Genitourinary	580–629, 788
Neoplasms	140–239
	780, 781, 784, 790–799
	240–279 (w/o 250)
	680–709, 782
Other (17.3%)	001–139
	290–319
	E–V
	280–289
	320–359
	630–679
	360–389
	740–759

References: Strack et al. (2014).

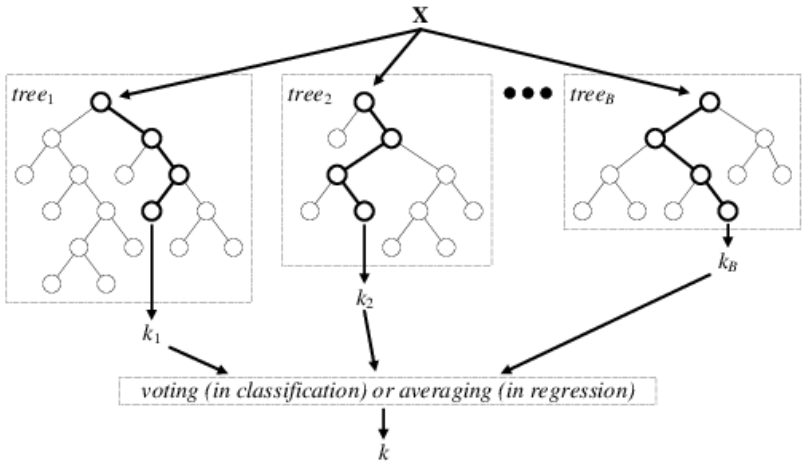
Model Training Overview

- Model development was performed using AutoGluon, an open-source AutoML toolkit.
- Included models:
  - Traditional ML: Logistic Regression, Random Forest, Extra Trees
  - Gradient Boosting: LightGBM, XGBoost, CatBoost
  - Neural Networks: Multi-layer Perceptron (MLP)
  - Foundation Model: TabPFNMix
  - AutoGluon Meta-model: WeightedEnsemble
- Metrics reported: AUROC, AUPRC, Accuracy, Precision, Recall, and F1 Score.

Traditional Machine Learning Models



Logistic Regression[a]

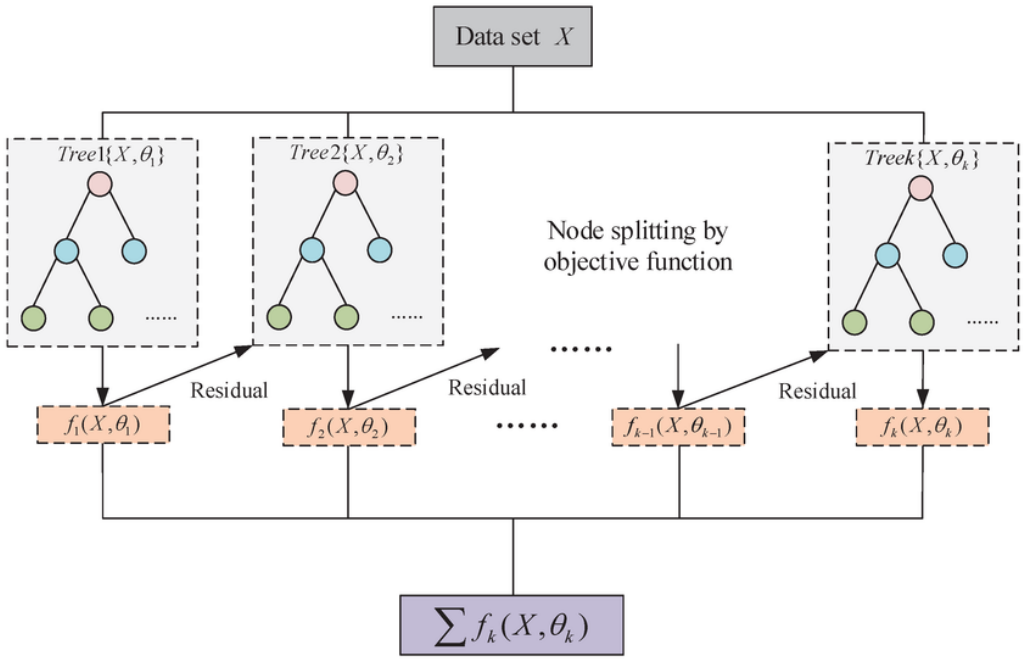


Random Forest[b]

- **Logistic Regression**: A linear model for binary classification based on the log-odds of input features.
- **Random Forest**: An ensemble of decision trees built from bootstrapped samples and random feature splits.

References: [a] www.saedsayad.com/logistic_regression.html, [b] Verikas et al. (2016)

Gradient Boosting Models

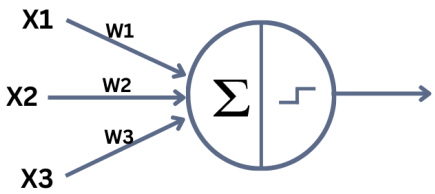


Gradient boosted tree (XGBoost)[a]

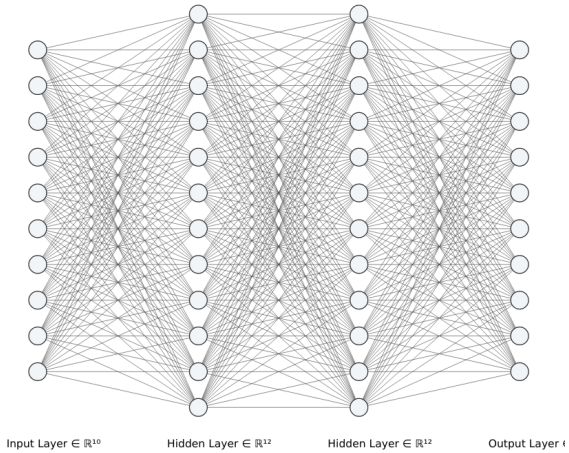
Gradient boosting is an ensemble technique that builds models sequentially, where each new model focuses on correcting the errors made by previous ones. It uses decision trees as weak learners and minimizes a loss function through gradient descent.

References: [a] Guo et al. (2020)

Multi-layer Perceptron (MLP)



Single-layer perceptron



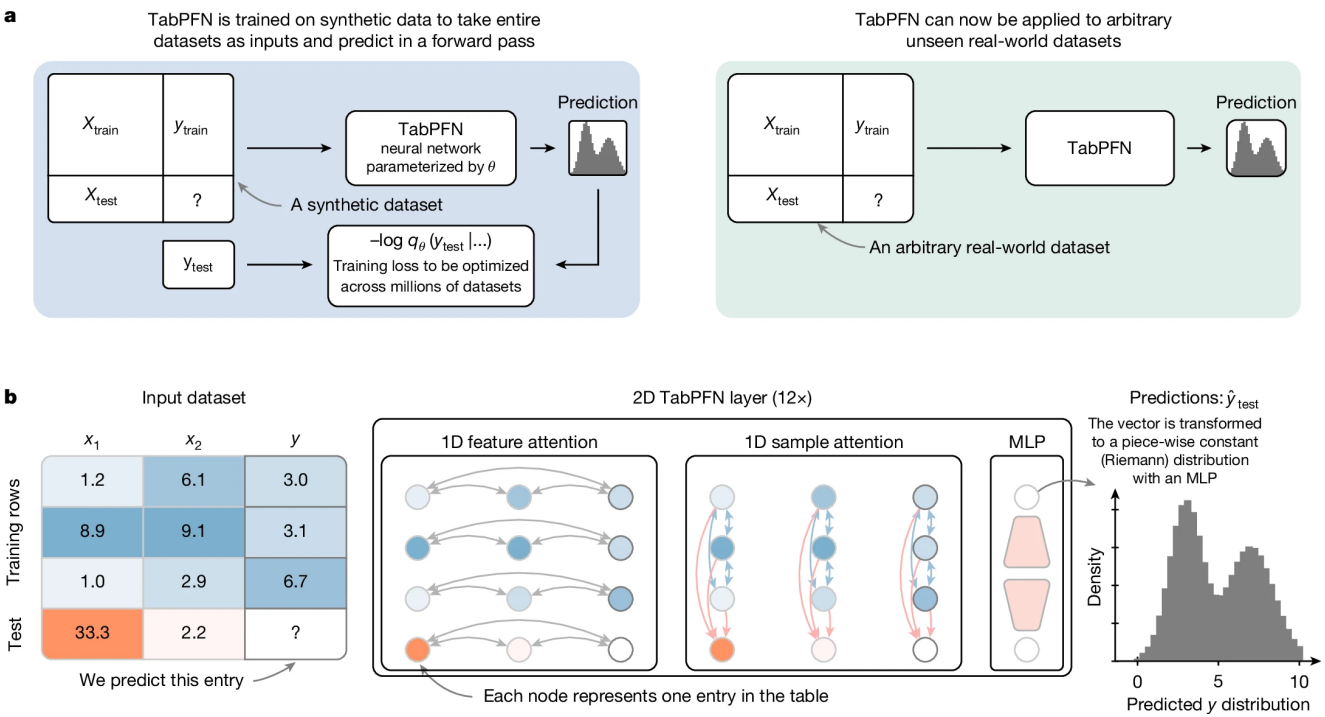
Multi-layer perceptron

Multi-layer Perceptron (MLP) [a]

Multi-layer Perceptron (MLP) is a supervised learning algorithm that learns a function $f : \mathbb{R}^m \to \mathbb{R}^o$, where m is the number of input features and o is the number of output dimensions. It maps inputs to outputs through one or more hidden layers. Non-linear activation functions enable the model to capture complex, non-linear patterns in the data.

References: [a] https://www.quarkml.com/2023/01/multi-layer-perceptron-a-complete-overview.html

TabPFN: Tabular Pretrained Foundation Network

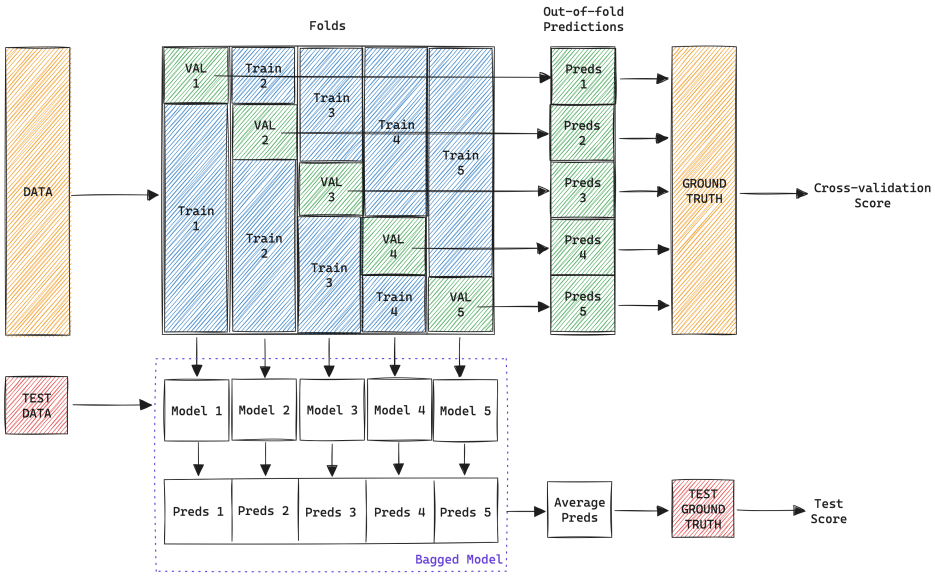


Architecture of TabPFN and its application to unseen tabular datasets [a]

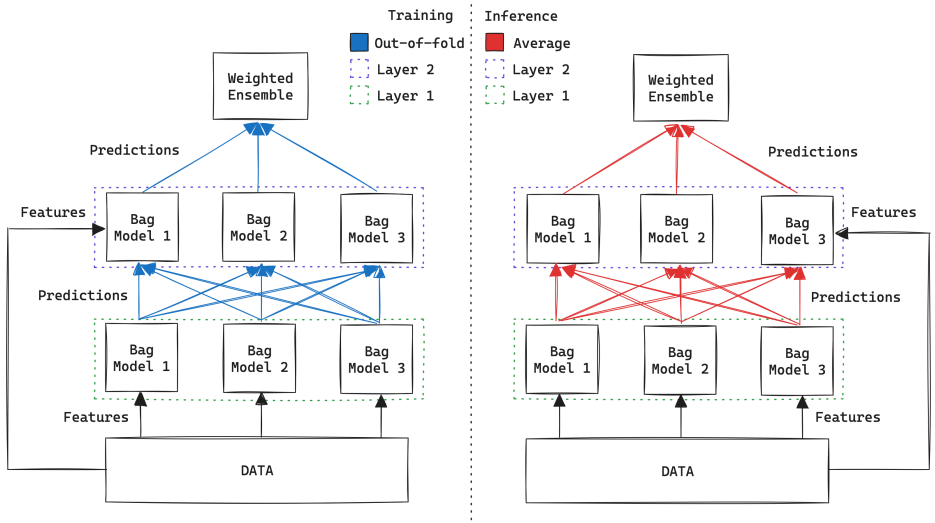
TabPFN is a tabular foundation model pre-trained on synthetic datasets generated from a diverse set of random classifiers. It uses a 12-layer encoder-decoder Transformer (37M parameters) and performs in-context learning to infer predictions without fine-tuning.

References: [a] https://github.com/PriorLabs/TabPFN

Ensemble Learning with AutoGluon



Bagging with Cross-validation



Stacked Ensembling

WeightedEnsembleis a meta-learner that combines predictions from multiple base models using weighted averaging, enhancing overall robustness and predictive accuracy.

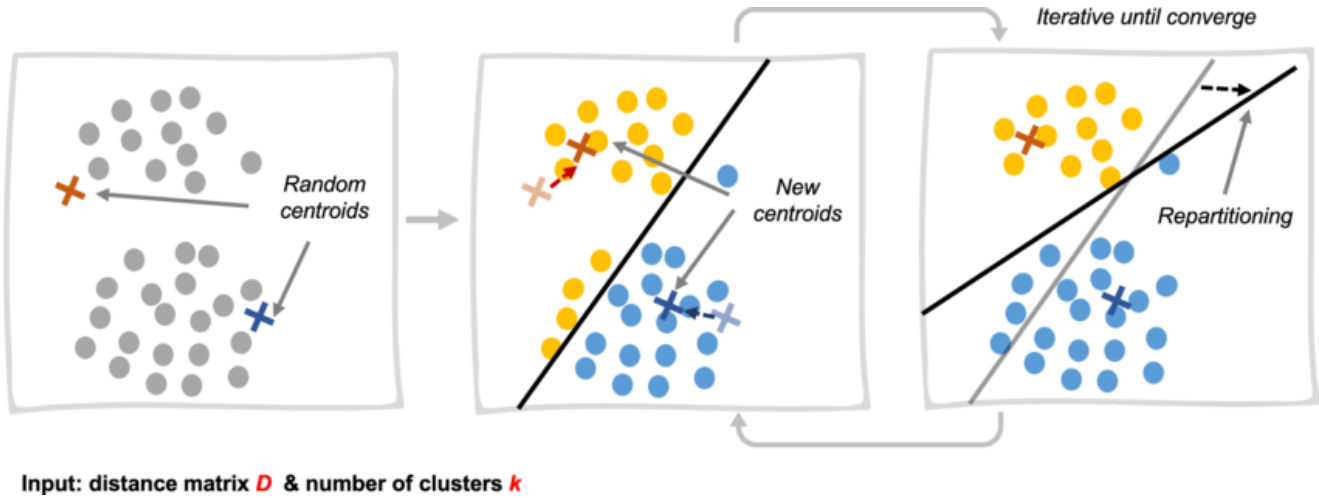
References: https://auto.gluon.ai/stable/tutorials/tabular/how-it-works.html

Explore Different Preprocessing Strategies

- **SMOTE**: Synthetic Minority Over-sampling Technique generates new synthetic samples between existing minority class instances.
- **ADASYN**: Adaptive Synthetic Sampling focuses more on difficult-to-learn minority examples by adaptively generating samples based on data distribution.
- **Random Downsampling**: Balances classes by randomly removing samples from the majority class.

References: https://medium.com/@sayahfares19/how-to-handle-imbalanced-classes-in-machine-learning-a7a42bcb54c2

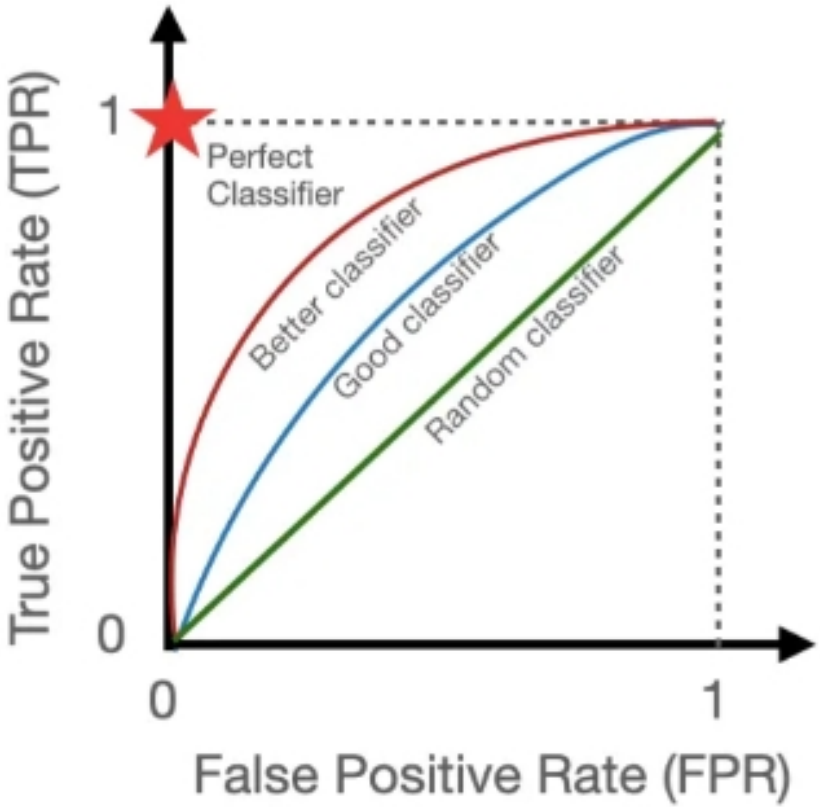
Exploring Different Feature Sets



K-Means is an unsupervised clustering algorithm that partitions data into k clusters by iteratively assigning samples to the nearest cluster centroid and recalculating centroids until convergence.

References: Gao et al. (2023)

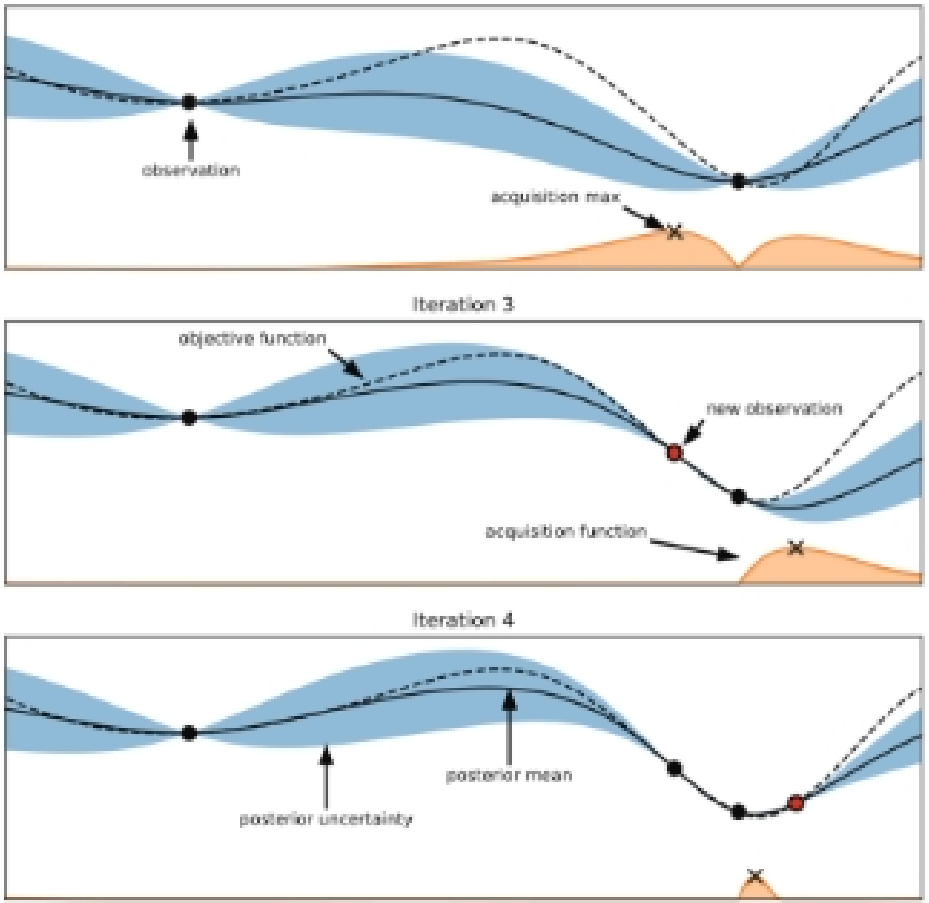
Area under the Receiver Operating Characteristic Curve



ROC is a probability curve of the true positive rate (TPR) against the false positive rate (FPR) at each prediction threshold setting. AUC represents a single scalar value that summarizes the area under the ROC, measuring the model's ability to **distinguish between positive and negative classes accurately.**

Ref: Di Sipio, R. (n.d.). A Quick Guide to AUC-ROC in Machine Learning Models. Towards Data Science.

Bayesian optimization



Bayesian Optimization builds a **surrogate model** of the objective function and uses an **acquisition function** to select hyperparameters to evaluate in the true objective function.

Ref: Cerino, F., et al. (2023). Hyperparameter Optimization of an hp-Greedy Reduced Basis for Gravitational Wave Surrogates. Universe, 10(6)

Performance Metrics

TP (True Positives), TN (True Negatives), FP (False Positives), FN (False Negatives)

- **Sensitivity:**

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

- **Specificity:**

$$\text{Specificity} = \frac{TN}{TN + FP}$$

- **Precision:**

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **F1 Score:**

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

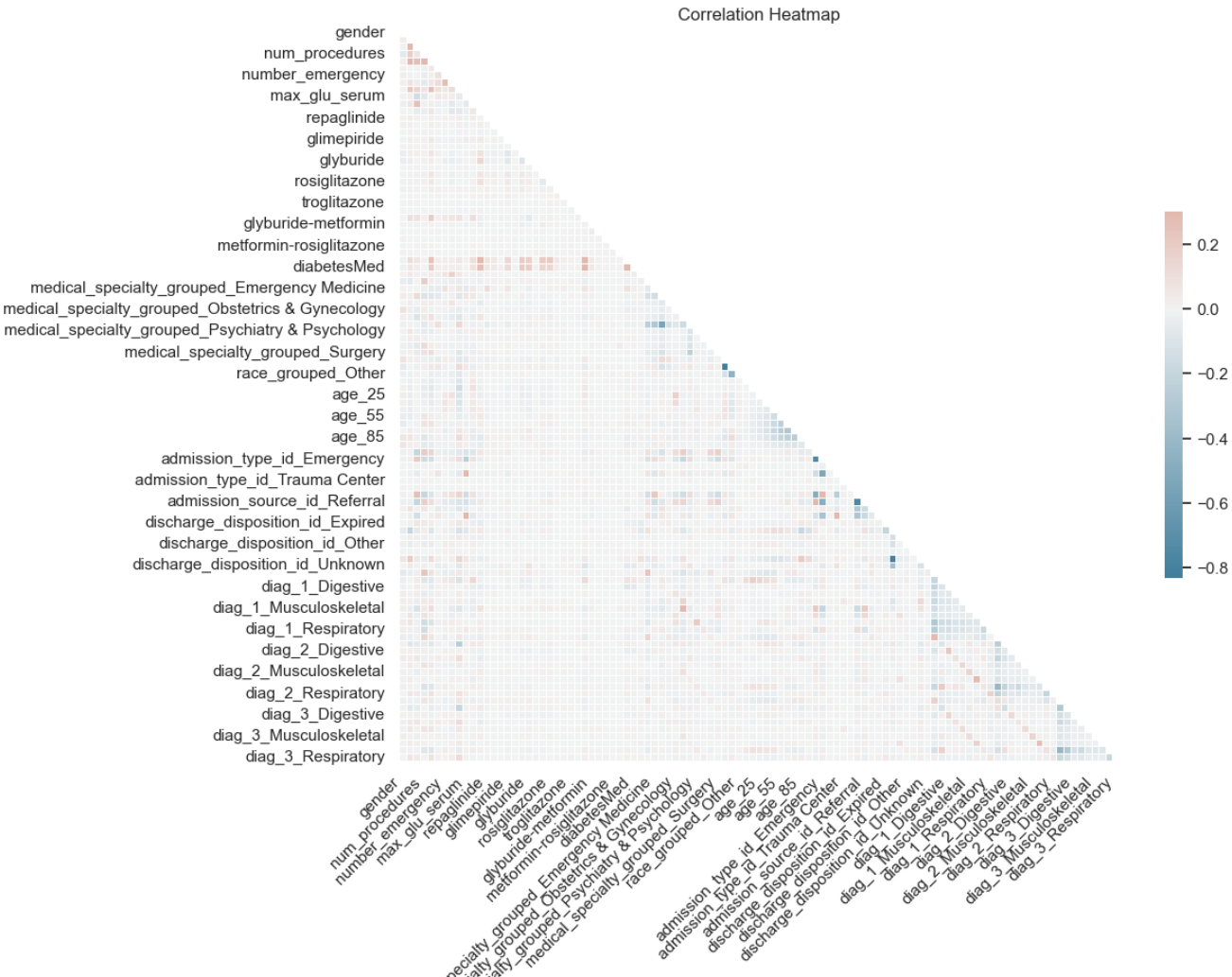
Results

Patient Characteristics

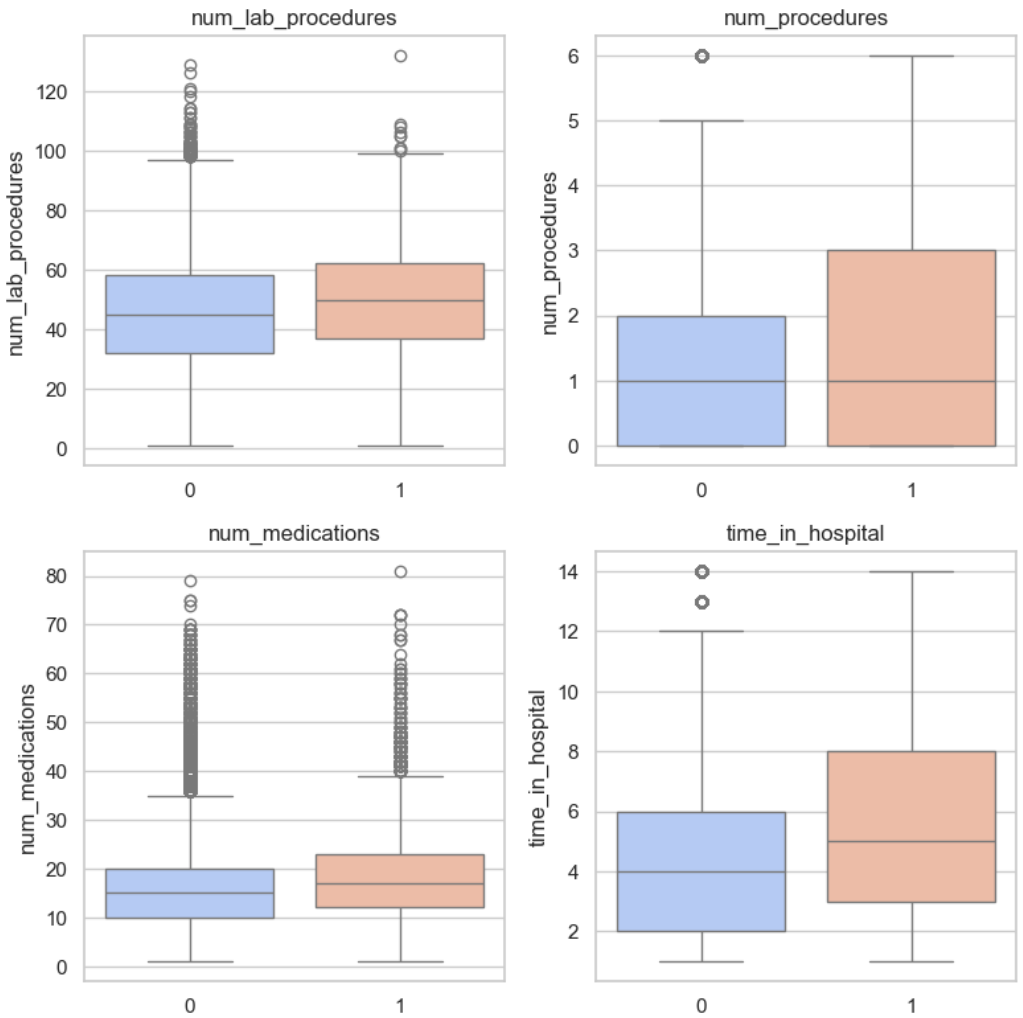
Discharge Disposition	Overall, n (%)	Readmitted = 0, n (%)	Readmitted = 1, n (%)
Expired	1273 (1.8%)	1273 (1.9%)	0 (0.0%)
Home	51956 (72.7%)	48853 (73.5%)	3103 (61.3%)
Hospice	561 (0.8%)	541 (0.8%)	20 (0.4%)
Outpatient	14 (0.0%)	11 (0.0%)	3 (0.1%)
Transfer	14181 (19.8%)	**12452 (18.7%)**	**1729 (34.1%)**
Unknown	3150 (4.4%)	2962 (4.5%)	188 (3.7%)
Other	380 (0.5%)	358 (0.5%)	22 (0.4%)

Note: Higher readmission observed in patients discharged to Transfer and lower in those discharged to Home.

Correlation Matrix



Boxplots by readmission status



Model Comparisons

Model	No Clustering	With Clustering	Top 10 Features	SMOTE	ADASYN	Down Sampling
WeightedEnsemble_L2	**0.699**	**0.699**	0.667	0.637	0.633	**0.695**
LightGBM	0.699	0.697	0.665	**0.657**	**0.658**	0.691
CatBoost	0.698	0.696	**0.668**	0.637	0.637	0.690
XGBoost	0.627	0.623	0.593	0.621	0.624	0.658
ExtraTrees	0.687	0.682	0.666	0.595	0.595	0.678
RandomForest	0.691	0.688	0.662	0.633	0.633	0.688
NeuralNetTorch	0.681	0.685	0.661	0.612	0.628	0.672
TabPFNMix	0.641	0.638	0.644	0.589	0.582	0.682
Logistic Regression	0.692	0.691	0.661	0.582	0.581	0.685

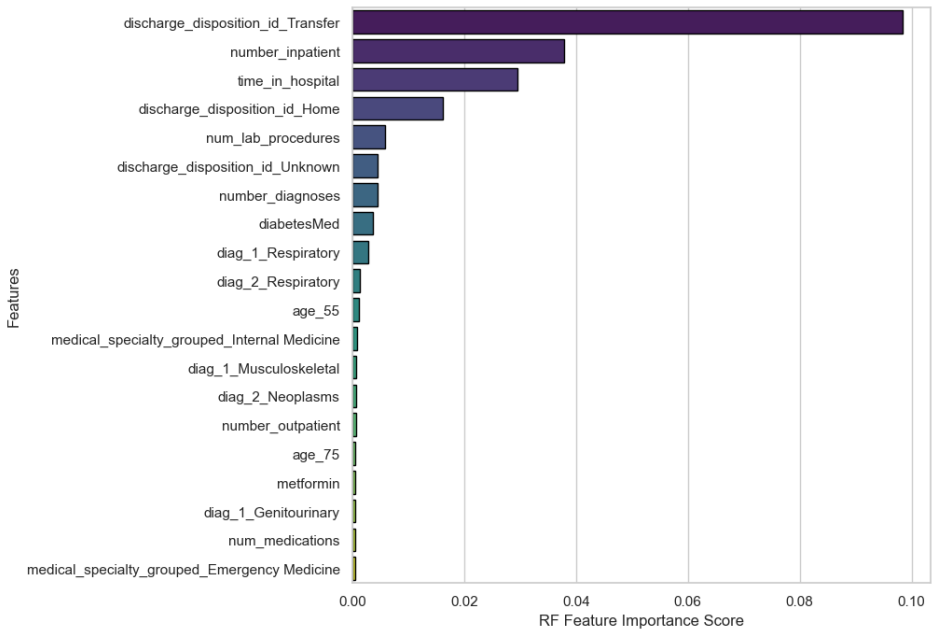
AUROC scores across experimental settings.

Performance Metrics of the Ensemble Model

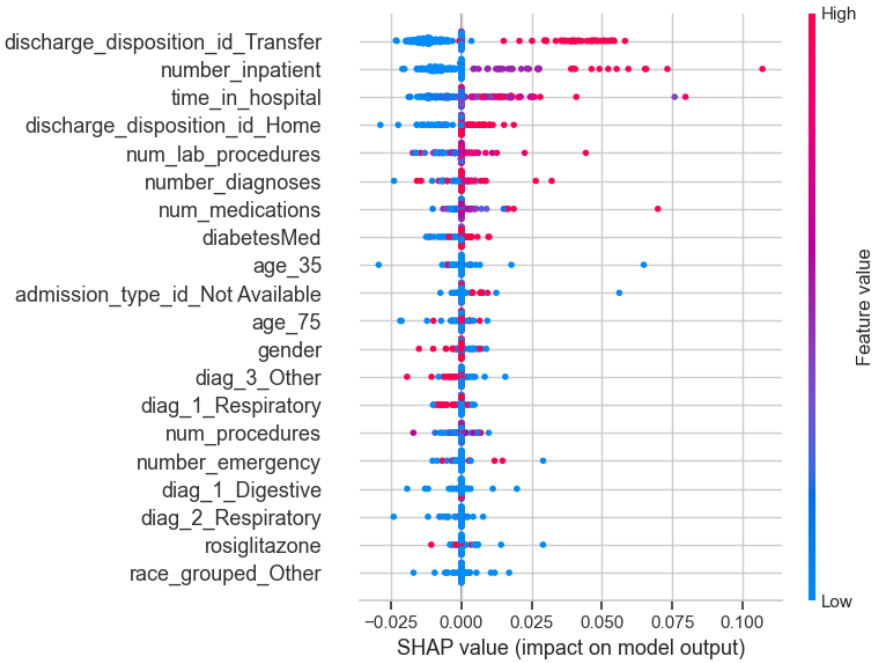
Model Type	Accuracy	Precision	Recall	F1	AUROC
No Clustering	0.7543	0.1428	0.4936	0.2215	**0.6991**
With Clustering	0.7975	**0.1558**	0.4205	**0.2273**	0.6987
Top 10 Features	**0.8266**	0.1502	0.3110	0.2026	0.6672
SMOTE	0.6321	0.1053	0.5597	0.1773	0.6374
ADASYN	0.6180	0.1034	**0.5726**	0.1751	0.6331
Down Sampling	0.7649	0.1442	0.4699	0.2207	0.6946

Comparison of performance metrics across experimental configurations for the ensemble model.

Feature Importances



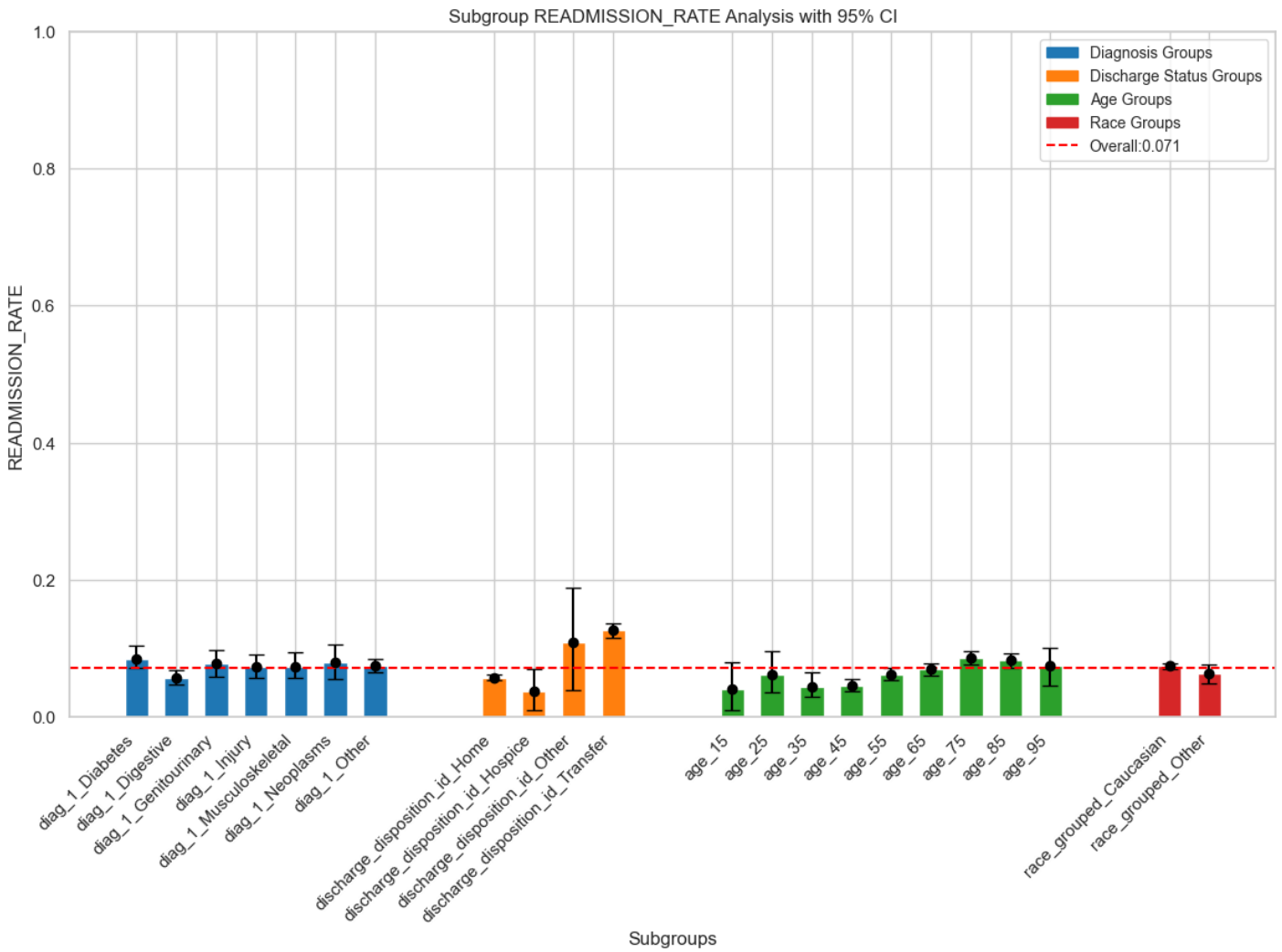
Random Forest Feature Importance



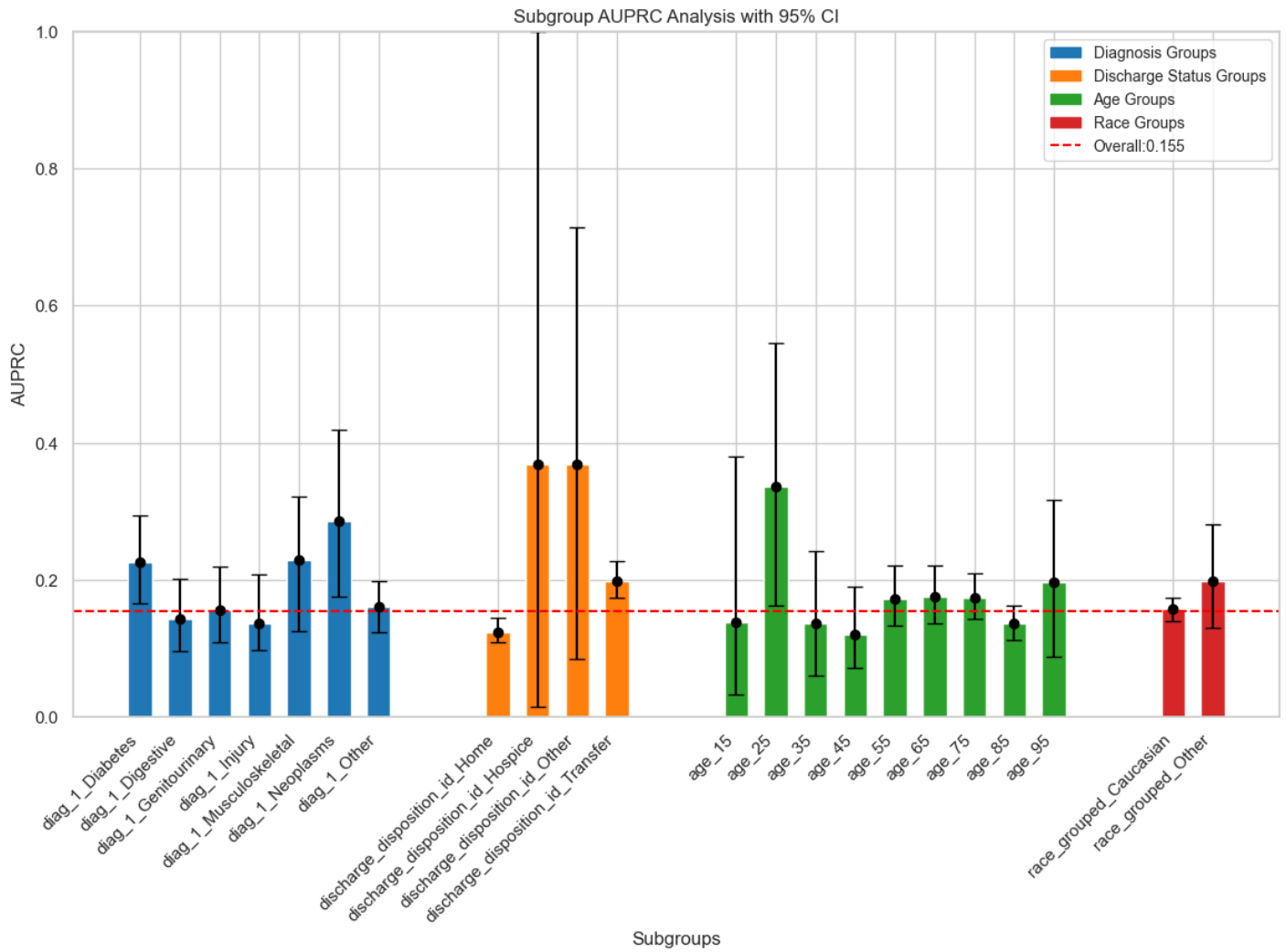
SHAP Feature Importance

- In summary, among the top 20 features, 13 were shared between both methods, and all of the top five were consistent across approaches.
- Patients with more prior inpatient visits, longer hospital stays, and discharge dispositions involving transfers are more likely to be readmitted.

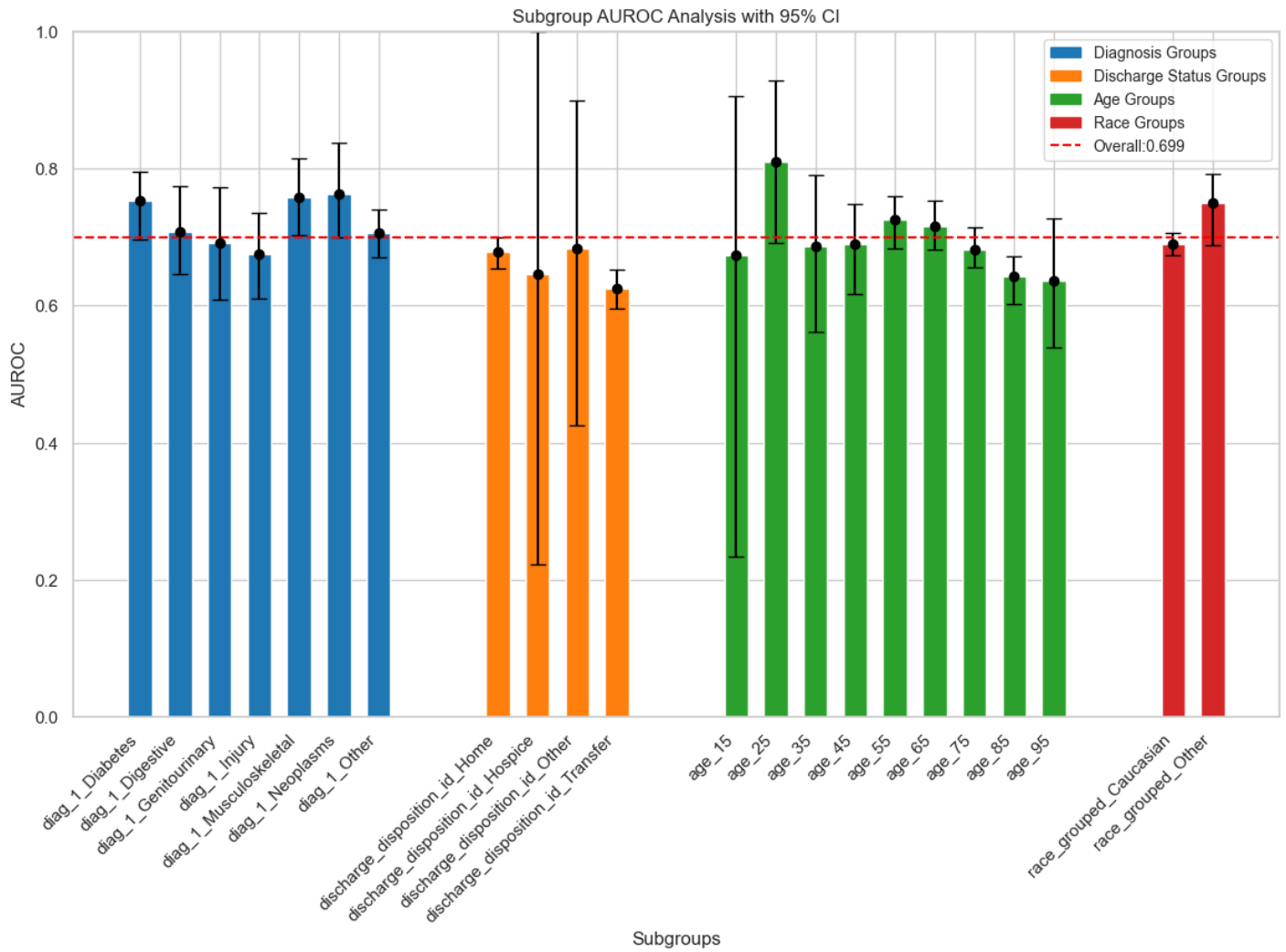
Readmission Prevalence Rate



AUPRC across subgroup



AUROC across subgroup



Discussions

Comparison with Existing Literature

- Results were consistent with prior studies: Liu et al. (2024) reported AUROC = 0.64 using XGBoost, and Shang et al. (2021) reported AUROC = 0.661 using Random Forest. In contrast, our ensemble model achieved a higher AUROC of 0.699.
- The high F1 scores (e.g., 0.84) reported by Liu et al. (2024) could not be reproduced in our study.
- Top Kaggle solutions often applied data balancing techniques (e.g., SMOTE) to the entire dataset before splitting, which inflated the target distribution in the test set.

Impact of Different Feature Sets and Preprocessing

- Adding cluster labels from K-means slightly improved F1 scores, with minimal impact on AUROC.
- Training on only the top 10 most important features yielded strong performance, suggesting feasibility for simplified models in clinical settings.
- SMOTE and ADASYN increased recall by generating synthetic samples, but reduced precision and AUROC—highlighting the trade-offs in imbalanced classification tasks.

Limitations

- The dataset is retrospective and cross-sectional, limiting the model's ability to capture temporal dynamics.
- The study population was predominantly Caucasian, which may reduce generalizability to more diverse populations.
- Some clinically important features, such as `A1Cresult`, had high missingness, potentially affecting model performance.
- Future work should incorporate longitudinal EHR data, behavioral and social determinants of health, and conduct external validation across diverse cohorts.

Conclusions

Key Findings

- This is the first study to apply `AutoGluon` to the task of predicting 30-day readmissions among diabetic patients.
- Nine models were compared, including traditional ML algorithms, deep neural networks, ensemble methods, and a transformer-based foundation model.
- The `WeightedEnsemble` model achieved the highest AUROC (0.699) in 3 out of 6 configurations and ranked among the top two in all settings.
- Gradient boosting models such as LightGBM and CatBoost also showed strong performance.
- Neural network-based models (TabPFNMix and MLP) underperformed compared to traditional ML models.
- Model performance was consistent across demographic and clinical subgroups, supporting its generalizability to diverse populations.

Thank You and Q&A

The source code and full report are available at:

`https://github.com/NokeYuan/Diabetes-Readmission-AutoGluon`

Or scan the QR code on the right.