THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Medical AI Research: Predicting postoperative blood transfusion for CABG patients

Capstone Preliminary Presentation, March 19, 2024

Team: Jenny Tsai [1] & Jichong Wu [1]
Collaborator: Dr. Puneet Gupta [2]
Advisor: Professor Amir Jafari [1]

[1] George Washington University
[2] George Washington University Hospital

Overview

Introduction

Data Preprocessing

Analysis Strategy

Modeling & Results

Conclusions

Next Steps

Introduction

Coronary Artery Bypass Graft (CABG) is a common cardiac surgery

- May cause major bleeding which needs blood transfusion

Blood transfusion is associate with:

- Higher risks of mortality after surgery
- Higher odds of readmission and heart failure within 30 days

Research Gap

Previous research*:

- A single cardiac surgery center in Austria
- N = 3782 (2010-2019)
- Random Forest:

  RUC: 0.76-0.86

In the current project:

- US national database [ACS NSQIP]
- N = 8587 (2018-2022)
- Basic models + Neural Networks + Feature engineering/selection

*Tschoellitsch et al. (2022)

Objectives

1. **Develop models that can best predict which patients need blood transfusion**
    - Improve patient selection and education
    - Enhance physician preoperative awareness
    - Inform periop guidelines for CABG patients

2. **Experiment with different DS techniques** (e.g., feature selection, feature engineering, synthetic data) applied in basic and advanced models **to achieve best outcomes**

3. **Develop a full set of modules that can be reused in the future,** which covers preprocessing, feature selection and feature engineering, and modeling

Data Preprocessing

Datasets

- Participant Use Data File (PUF) on the American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP)
- Year 2018 - 2022 (N = 8587, # of features = 294)

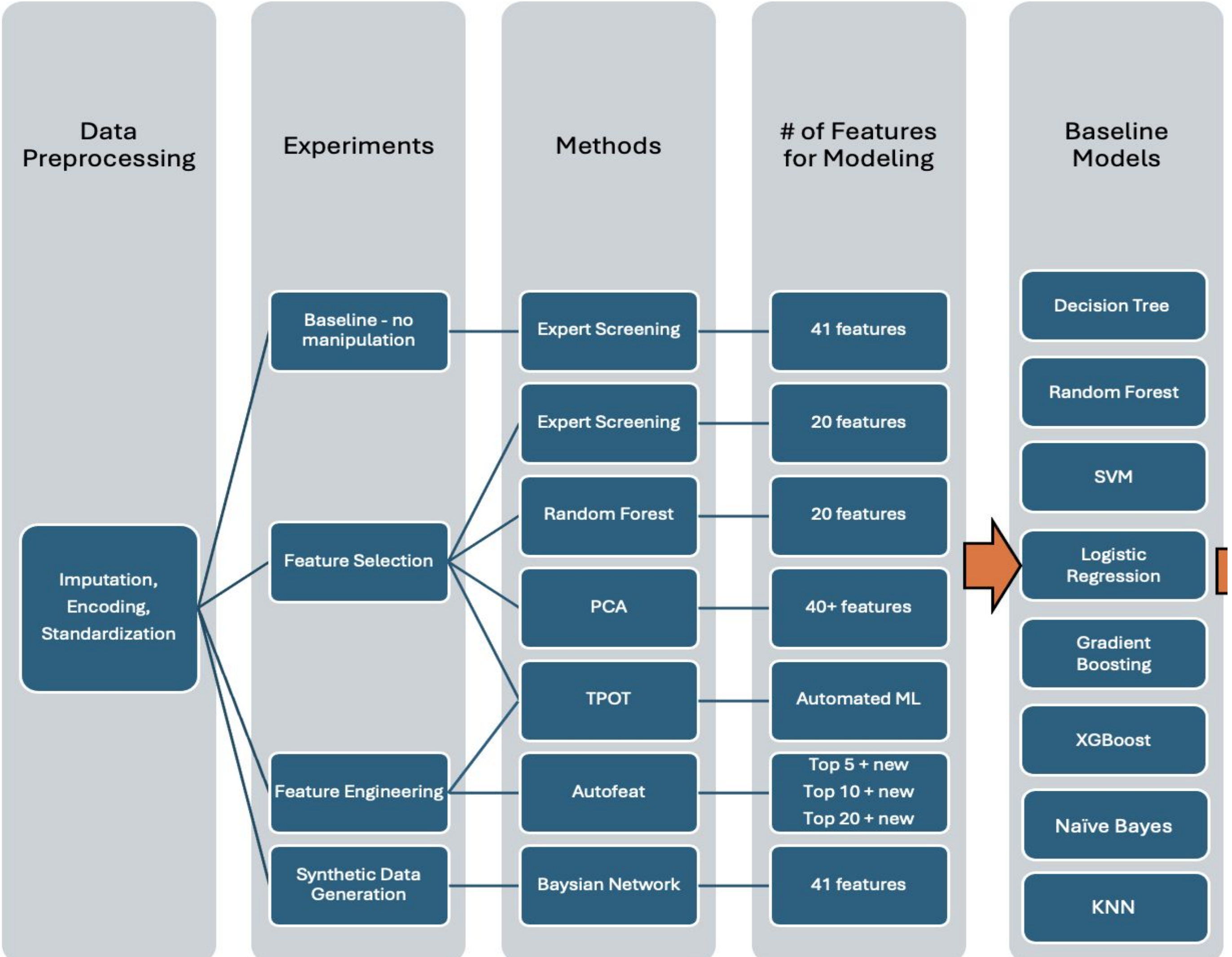
Key preprocessing steps:

1. Basic clean-up (e.g., recode values, correct data type)
2. Remove columns with over 50% missing values
3. Impute with mean (numeric) and most frequent values (categorical)
4. Standardize all numeric features
5. Remove post-operative and irrelevant features by expert

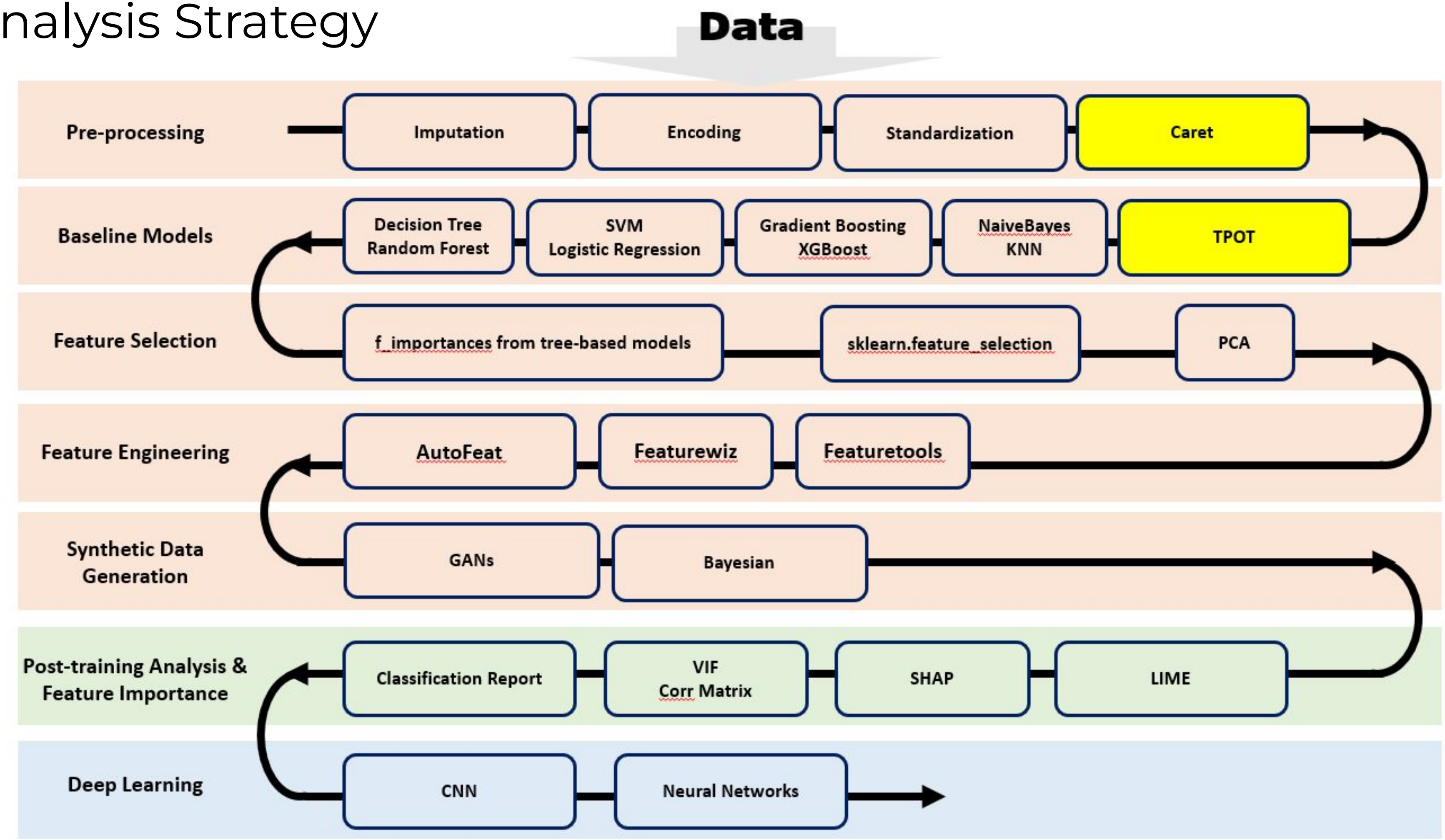
Final dataset size:

N = 8587, # of features = 41

Analysis Strategy



Analysis Strategy

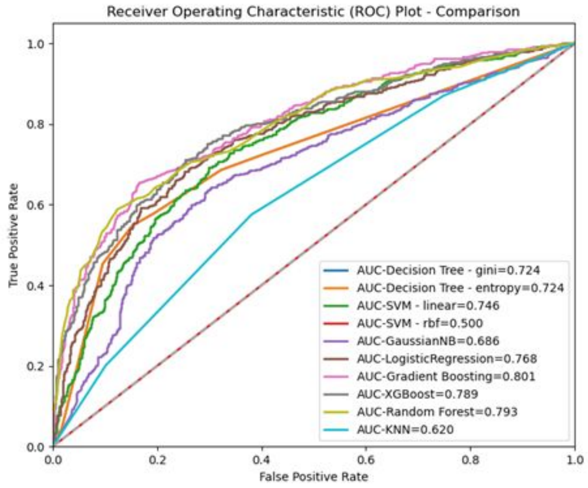


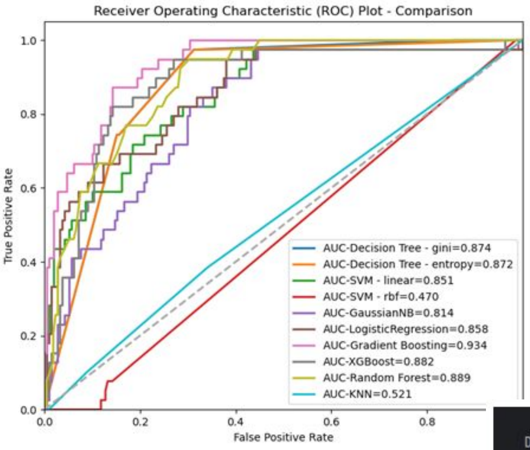
Modules and Utilities

class - data pre-processing		class - baseline models	class - feature selection		class - feature engineering	utility
class name	methods		class name	methods		
datasci	.size()	DecisionTree	PCA	PCA_Reduced_Feature()	featurewiz	file_compare()
	.recode()	RandomForest		Reduced_Feature_Space_Plot()	featuretools	glossary()
	.missingReport()	SVM		Reduced_Feature_Space_Heatmap()		Model_Predict()
	.remove_all_nan_columns()	LogReg		Explained_Variance_Ratio()		Model_Report()
	.impute_all()	GradientBoosting		Reduced_Feature_Space_Plot()		Model_Accuracy()
	.imputation()	XGB		Reduced_Feature_Space_Heatmap()		Model_Mean_Accuracy()
	.standardize()	NaiveBayesGaussianNB		PCA_New_df()		Model_RMSE()
	.eda()	KNN				Model_F1()
	.featureSelection()	TPOT				Model_Confusion_Matrix()
						Plot_Confusion_Matrix()
						Plot_Decision_Tree()
						Model_ROC_AUC_Score()
						Plot_ROC_AUC()
						Plot_Random_Forest_Feature_Importances()
						Model_Results_Table()
						Plot_ROC_Combined()
						Calc_Plot_VIF()
						Calc_Top_Corr()
						Plot_Heatmap_Top_Corr

Model Results

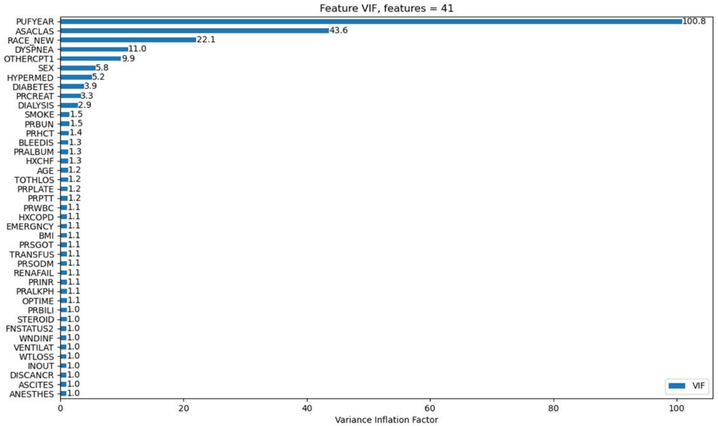
Model Name	Parameters	Target	Test Size	Accuracy	RMSE	F1-score (macro avg)	ROC-AUC score
Decision Tree - gini	max_depth=3, min_samples_leaf=5	OTHBLEED	0.25	68.119451	0.564629	0.679578	0.724499
Decision Tree - entropy	max_depth=3, min_samples_leaf=5	OTHBLEED	0.25	68.119451	0.564629	0.679578	0.724499
SVM - linear	C=1.0, gamma=auto	OTHBLEED	0.25	69.491525	0.552345	0.687647	0.746498
SVM - rbf	C=1.0, gamma=0.2	OTHBLEED	0.25	55.528652	0.666868	0.357032	0.500000
GaussianNB		OTHBLEED	0.25	52.945924	0.685960	0.496722	0.685814
LogisticRegression		OTHBLEED	0.25	71.751412	0.531494	0.712662	0.767803
Gradient Boosting	n_estimators=300, learning_rate=0.05	OTHBLEED	0.25	73.930589	0.510582	0.734304	0.801463
XGBoost	n_estimators=100, eta=0.3	OTHBLEED	0.25	72.558515	0.523846	0.721558	0.789397
KNN	n_neighbors=3	OTHBLEED	0.25	59.967716	0.632711	0.596555	0.619534
Random Forest	n_estimators=100, 20_features	OTHBLEED	0.25	73.930589	0.510582	0.729318	0.792982

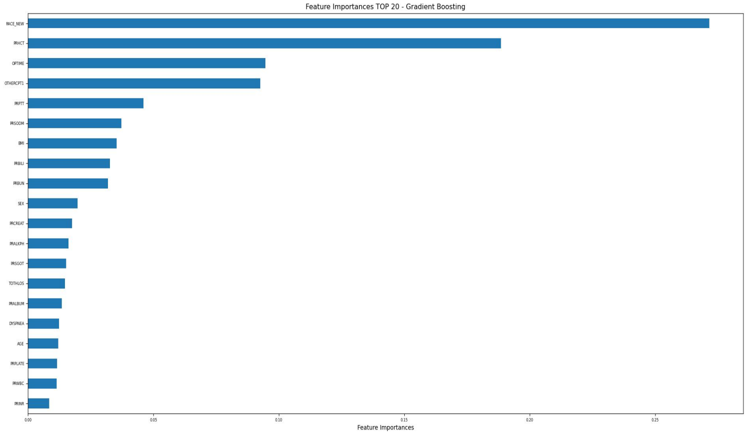


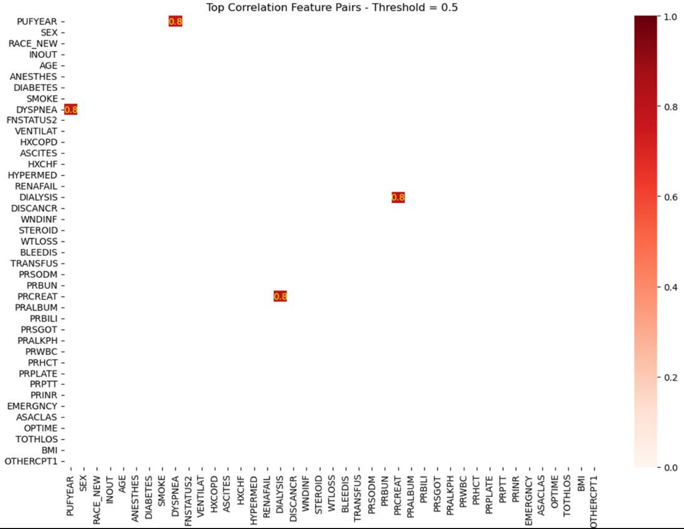


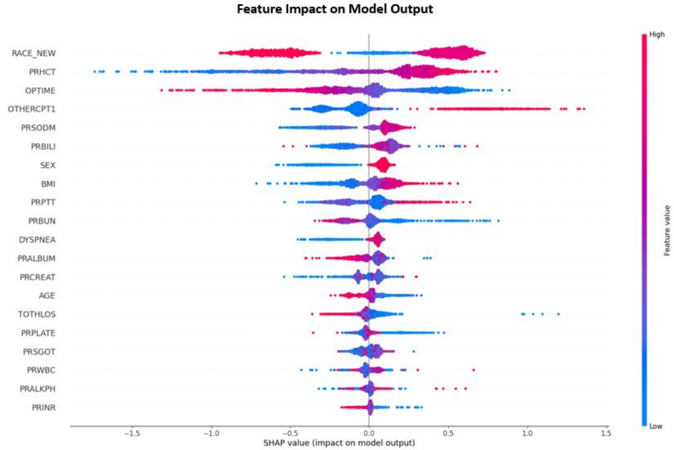
Model Name	Parameters	Target	Test Size	Accuracy	Mean Accuracy (10 folds)	RMSE	F1-score (macro avg)	ROC-AUC score
Decision Tree - gini	max_depth=3, min_samples_leaf=5	OTHBLEED	0.25	83.2	86.9	0.409878	0.737500	0.874408
Decision Tree - entropy	max_depth=3, min_samples_leaf=5	OTHBLEED	0.25	83.2	86.9	0.409878	0.737500	0.872159
SVM - linear	C=1.0, gamma=auto	OTHBLEED	0.25	86.8	86.0	0.363318	0.735331	0.851136
SVM - rbf	C=1.0, gamma=0.2	OTHBLEED	0.25	84.4	83.9	0.394968	0.457701	0.469741
GaussianNB		OTHBLEED	0.25	72.0	79.5	0.529150	0.628639	0.814315
LogisticRegression		OTHBLEED	0.25	87.2	86.7	0.357771	0.751738	0.858306
Gradient Boosting	n_estimators=300, learning_rate=0.05	OTHBLEED	0.25	89.6	90.7	0.322490	0.789071	0.933649
XGBoost	n_estimators=100, eta=0.3	OTHBLEED	0.25	85.2	88.6	0.384708	0.737775	0.881638
KNN	n_neighbors=3	OTHBLEED	0.25	78.4	81.6	0.464758	0.502872	0.521388
Random Forest	n_estimators=100, 20_features	OTHBLEED	0.25	88.0	85.4	0.346410	0.715909	0.888747

Post-training Analysis









Conclusions

1. PUFYEAR, ASACLAS, RACE_NEW may cause multicollinearity.
2. **Gradient Boosting**, XGBoost, Random Forest perform the best.
3. **Synthetic data generation** techniques (DataSynthesizer using Bayesian networks) significantly improve model performance.
4. **Race, days from preoperative labs to operation, operation time**, other procedure, BMI, sex, length of hospital stay, shortness of breath, age, preoperative blood test measures
5. **days from preoperative labs to operation, operation time may have a negative impact on model results while other procedure, BMI have a positive effect.**

Next Steps

Improve model performance by:

- Add more samples (older datasets from 2015-2017)
- Recategorize target variable (intra vs. postop blood transfusion)
- Other ways to generate synthetic data (e.g., realtabformer)
- Conduct post-training analysis (continued) to study impacts of features to model performance
- Neural networks (e.g., CNN, transformers)

Thank You!

THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC