

Data Science Program

Capstone Report - Spring 2024

Predicting Blood Transfusions for Coronary Artery Bypass Graft Patient

Jenny Hsiao-Tien Tsai,

Jichong Wu,

Puneet Gupta

Supervised by

Amir Jafari

Abstract

A strong abstract sums up your work in very few sentences: (i) state the problem you are addressing; (ii) say why it's an interesting problem, and which issues are hard to tackle; (iii) give your approach towards solving the problem; (iv) say Why and how well your approach solves the problem.

Contents

[1 Introduction 3](#_Toc162457211)

[2 Problem Statement & Project Objectives 3](#_Toc162457212)

[3 Related Work 3](#_Toc162457213)

[4 Solution and Methodology 4](#_Toc162457214)

[5 Results and Discussion 5](#_Toc162457215)

[5.1 Model selection and tuning 5](#_Toc162457216)

[5.2 Results and interpretation 9](#_Toc162457217)

[6 Discussion 9](#_Toc162457218)

[7 Conclusion 9](#_Toc162457219)

[8 Bibliography 9](#_Toc162457220)

# Introduction

Coronary Artery Bypass Graft (CABG) is a common cardiac surgery but continues to have many associated risks, including major bleeding which might need blood transfusion. Previous research has shown that blood transfusion during CABG surgery is associated with an increased risk for mortality after surgery. Specially, post-operative blood transfusion after CABG is associated with higher odds of readmission and heart failure within 30-days.

To lower the risk of mortality after surgery, there is a need to develop models that preoperatively predict which patients will need an intra-operative or post-operative blood transfusion. This will not only help to improve patient selection and patient education, but also physician preoperative awareness and perioperative guidelines for CABG patients. Therefore, the goal of this project is to explore different approaches and find the models that can best make predictions, including feature selection/engineering, classical statistical models, and neural networks.

# Problem Statement & Project Objectives

The objectives of the project are three-fold. The first objective is to develop models that can best predict whether a CABG patient will need blood transfusions. Second, we also look to experiment with various data science techniques to be applied in our models in order to achieve best performance, including feature selection, feature engineering, and synthetic data generation. Lastly, we aim to build a full set of modules and functions to be reused in the future beyond the current project. The modularized codes include but not limited to data preprocessing, feature selection, feature engineering, and modeling.

# Related Work

Research have been conducted to investigate factors that can help to predict major bleeding (Gao, et al., 2022) and the need for red blood cell transfusion after cardiac surgery (Li, et al., 2024). In one of the studies (Tschoellitsch, Bock, Mahecic, Hofmann, & Meier, 2022) that is most relevant to the current project, the researchers employed machine learning models to predict perioperative allogeneic blood transfusion for cardiac patients. The best model (Random Forest) showed good performance (RUC ranged from .76 - .86), however, the study has several limitations. For example, the data was from a single adult cardiac surgery center in Austria with a relatively small sample size (N = 3782), thus the results may not be generalizable to other samples with different demographics or nationalities. Moreover, the studies only predicted allogeneic blood transfusion (i.e., transfusion of more than 10 units of packed red blood cells (pRBC)), while blood transfusion regardless of volume has been associated with many known risks. Lastly, the study only tested the basic machine learning models (e.g., tree-based models), and it is likely that the performance can be significantly improved using more advanced techniques and deep neutral networks.

To address this research gap, the current project will use the national medical database in the U.S. with a large sample size of over 8,000 data points. Additionally, we will predict blood transfusion regardless of volume. Lastly, we will experiment with various approaches in order to optimize the performance, including feature selection, feature engineering, synthetic data generation, and deep neural networks.

# Solution and Methodology

*Data Source and Data Preprocessing*

The data was downloaded from the Participant Use Data File (PUF) on the American College of Surgeons National Surgical Quality Improvement Program ([ACS NSQIP](https://www.facs.org/quality-programs/data-and-registries/acs-nsqip/)). In this project, we focus on the data from 2018 to 2022, which has a total of 8,587 observations and around 294 variables across five datasets.

Add a table to summarize our data?

Add a table to display our feature dictionary?

After data preprocessing, including basic cleanup, imputation (mean for numeric variables and most frequent values for categorical variables), standardization, and encoding, the dataset with 41 features identified as most relevant to the current study was served as our baseline data. The target variable is Occurrences Bleeding Transfusions, which is a binary variable predicting whether the patient needs blood transfusion after surgery. The target can be further categorized into intraoperative vs. postoperative vs. no transfusion, therefore can be transformed into a 3-class variable when needed. With different analysis strategies, these features will be entered into our models to predict the target variable, and we will compare the performance with each other as well as with the benchmarks from previous research.

Add a table to display the 41 selected features

*Add a section on EDA?*

*Analysis Strategy*

Figure X shows the analysis strategy for the current project. After data preprocessing, we first entered the data into eight models, and used the results as our baseline benchmark. Next, we experimented with a different technique and method, and then entered the modified data into our models. In each iteration, we compared the new model performance with the baseline results.

A diagram of a process

Description automatically generated

*Figure X., Analysis strategy for the current project.*

# Results and Discussion

## Model selection and tuning

We set random state as 100, testing size 25%, k-folds 10 in all models for comparison. The target is “OTHBLEED”.

*Iteration #1:* As the first step, we include all possible features with a missing data percentage larger than 50%. We then drop “NOTHBLEED” and “DOTHBLEED” due to high collinearity with the target. NOTHBLEED, number of bleeding transfusions occurrences, is highly correlated with the target OTHBLEED (occurrences bleeding transfusions) and has a Pearson correlation coefficient of -0.99. Similarly, DOTHBLEED, days from operation until bleeding transfusions complication, has a -0.81 Pearson correlation coefficient with the target. This leaves the data to be a 4953 by 127 dimension.

Simple linear imputation is applied to fill in the missing data in order for some models to run without errors. However, data is not standardized in this first round of modeling. Table xx summarizes the setup of iteration #1.

Eight typical classification models were selected to train the data, including Decision Tree, SVM, Gaussian Naive Bayes, Logistic Regression, Gradient Boosting, XGBoost, KNN, and Random Forest. Different parameters and algorithm were also compared within Decision Tree (gini vs entropy) and SVM (linear vs rbf). Table xx indicates that Random Forest and Gradient Boosting have the better results across several evaluation metrics (accuracy score, root mean square error, F1 score, and ROC-AUC score). Figure X combines the ROC plots for all the models which verify the conclusion above on top performing models.

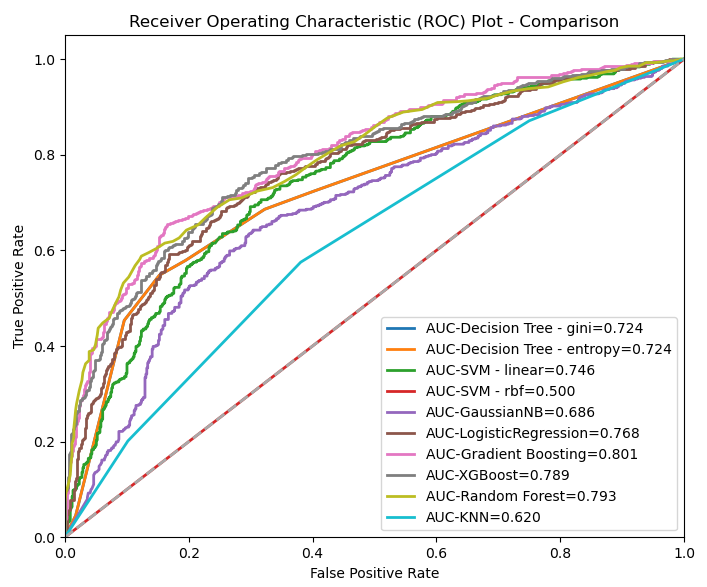
Table: Summary of Iteration #1 Setup

|  |  |
| --- | --- |
| Data year | 2018-2020 |
| Observations | 4953 |
| Features included | 126 |
| Features manually dropped based on expert judgement | NOTHBLEED  DOTHBLEED |
| Data preprocessing techniques applied | Simple imputations |
| Final dataset | [CABG\_2018\_2020\_baseline.csv](https://github.com/jennytsai32/Capstone/blob/master/code/main_code/processed_data/2018_2020/CABG_2018_2020_baseline.csv) |

Table: Model Results from Iteration #1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model Name | Parameters | Accuracy | RMSE | F1 (macro avg) | ROC-AUC |
| Decision Tree – gini | max\_depth=3  min\_samples\_leaf=5 | 68.12 | 0.56 | 0.68 | 0.72 |
| Decision Tree – entropy | max\_depth=3  min\_samples\_leaf=5 | 68.12 | 0.56 | 0.68 | 0.72 |
| SVM – linear | C=1.0, gamma=auto | 69.49 | 0.55 | 0.69 | 0.75 |
| SVM – rbf | C=1.0, gamma=0.2 | 55.53 | 0.67 | 0.36 | 0.50 |
| Gaussian Naive Bayes |  | 52.95 | 0.69 | 0.50 | 0.69 |
| Logistic Regression |  | 71.75 | 0.53 | 0.71 | 0.77 |
| Gradient Boosting | n\_estimators=300  learning\_rate=0.05 | 73.93 | 0.51 | 0.73 | 0.80 |
| XGBoost | n\_estimators=100  eta=0.3 | 72.56 | 0.52 | 0.72 | 0.79 |
| KNN | n\_neighbor=3 | 59.97 | 0.63 | 0.60 | 0.62 |
| Random Forest | n\_estimators=300  feature\_importances=20 | 73.93 | 0.51 | 0.73 | 0.79 |

Figure xx: ROC Plot with 10 Selected Models from Iteration #1



*Iteration #2:*

Table: Summary of Iteration #2 Setup

|  |  |
| --- | --- |
| Data year | 2018-2022 |
| Observations | 8587 |
| Features included | 126 |
| Final dataset | [CABG\_2018\_2022\_baseline.csv](https://github.com/jennytsai32/Capstone/blob/master/code/main_code/processed_data/2018_2022/CABG_2018_2022_baseline.csv) |

*Iteration #3:*

Table: Summary of Iteration #3 Setup

|  |  |
| --- | --- |
| Data year | 2018-2022 |
| Observations | 8587 |
| Features included | 41 |
| Features manually dropped based on expert judgement | HEIGHT  WEIGHT  ETHNICITY\_HISPANIC |
| Features kept based on expert judgement | ASACLAS  RACE\_NEW  SEX  OTHERCPT1 |
| Data preprocessing techniques applied | Standardization  Cross validation (10-folds) |
| Final dataset | [CABG\_5yr\_preselect41.csv](https://github.com/jennytsai32/Capstone/blob/master/code/main_code/processed_data/2018_2022/CABG_5yr_preselect41.csv) |

*Iteration #4:*

Table: Summary of Iteration #4 Setup

|  |  |
| --- | --- |
| Data year | 2018-2022 |
| Observations | 8587 |
| Features included | 40 |
| Data preprocessing techniques applied | PCA |
| Final dataset | TBD |

*Iteration #5:*

Table: Summary of Iteration #5 Setup

|  |  |
| --- | --- |
| Data year | 2018-2022 |
| Observations | 8587 |
| Features included | 41 |
| Data preprocessing techniques applied | AutoFeat |
| Final dataset | [CABG\_autofeat\_top20.csv](https://github.com/jennytsai32/Capstone/blob/master/code/main_code/processed_data/2018_2022/CABG_autofeat_top20.csv) |

*Iteration #6:*

Table: Summary of Iteration #6 Setup

|  |  |
| --- | --- |
| Data year | 2018-2022 |
| Observations | 8587 |
| Features included | 41 |
| Data preprocessing techniques applied | TPOT |
| Final dataset | [CABG\_5yr\_preselect41.csv](https://github.com/jennytsai32/Capstone/blob/master/code/main_code/processed_data/2018_2022/CABG_5yr_preselect41.csv) |

*Iteration #7:*

Table: Summary of Iteration #7 Setup

|  |  |
| --- | --- |
| Data year | 2018-2022 |
| Observations | 8587 |
| Features included | 41 |
| Data preprocessing techniques applied | Synthetic data generation – Bayesian networks |
| Final dataset | [CABG\_synthetic\_Bayesian.csv](https://github.com/jennytsai32/Capstone/blob/master/code/main_code/processed_data/2018_2022/CABG_synthetic_Bayesian.csv) |

*Iteration #8:*

Table: Summary of Iteration #8 Setup

|  |  |
| --- | --- |
| Data year | 2015-2022 |
| Observations | 13534 |
| Features included | 41 |
| Data preprocessing techniques applied | N/A |
| Final dataset | [CABG\_8yr\_preselect41.csv](https://github.com/jennytsai32/Capstone/blob/master/code/main_code/processed_data/2015_2022/CABG_8yr_preselect41.csv) |

*Iteration #9:*

Table: Summary of Iteration #9 Setup

|  |  |
| --- | --- |
| Data year | 2015-2022 |
| Observations | 13534 |
| Features included | 41 |
| Data preprocessing techniques applied | Synthetic data generation – Bayesian networks |
| Final dataset | TBD |

## Results and interpretation

* Best models and comparison with other literatures
* Post-training analysis: what are important features and their impacts in blood transfusion

# Discussion

* Our code vs. “one-line code” package (e.g. Caret, TPOT)
* 2 labels vs 3 labels in the Target – will it improve model performance?
* Features that can be dropped, and those should keep
* Additional ways of feature selection and feature engineering, and the comparison
* Synthetic data generation can significantly improve the models

# Conclusion

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.

# Bibliography

Gao, Y., Liu, X., Wang, L., Wang, S., Yu, Y., Ding, Y., . . . Ao, H. (2022, July 28). Machine learning algorithms to predict major bleeding after isolated coronary artery bypass grafting. *Front Cardiovasc Med.*

Li, Q., Lv, H., Chen, Y., Shen, J., Shi, J., Zhou, C., & Yan, F. (2024, April). Development and validation of a machine learning prediction model for perioperative red blood cell transfusions in a cardiac surgery. *International Journal of Medical Informatics, 184*.

Tschoellitsch, T., Bock, C., Mahecic, T., Hofmann, A., & Meier, J. (2022, September). Machine learning-based prediction of massive perioperative allogeneic blood transfusion in cardiac surgery. *European Society of Anaesthesioloty and Intensive Care, 39*(9), 766-773.