One-year Pediatric Bone Marrow Transplant Survival Prediction

Zoey Yu Brown University Dec 13, 2024

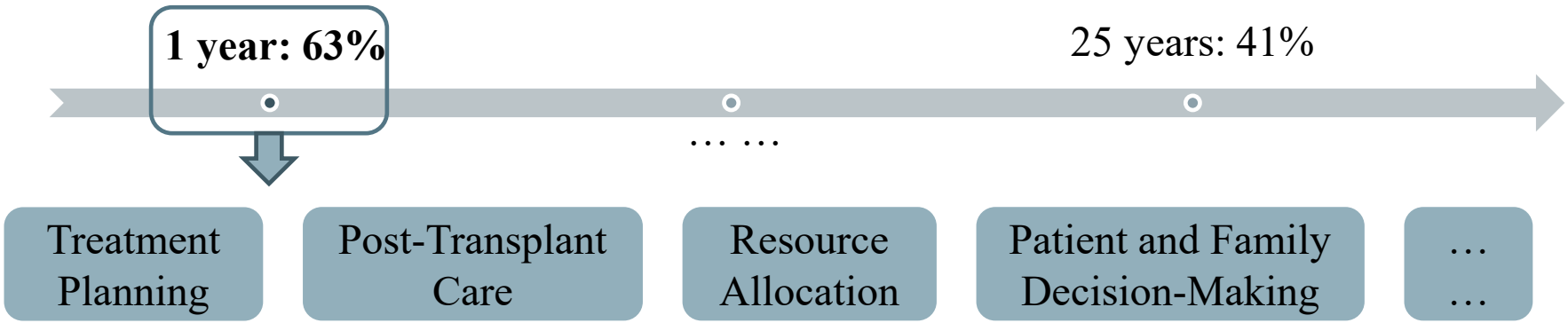
https://github.com/Zoey-Yu98/data1030-midterm.git

IMPORTANCE

Bone Marrow Transport (BMT)

A Post-op 1-yr Survival Status Prediction ➔ Decisions Making

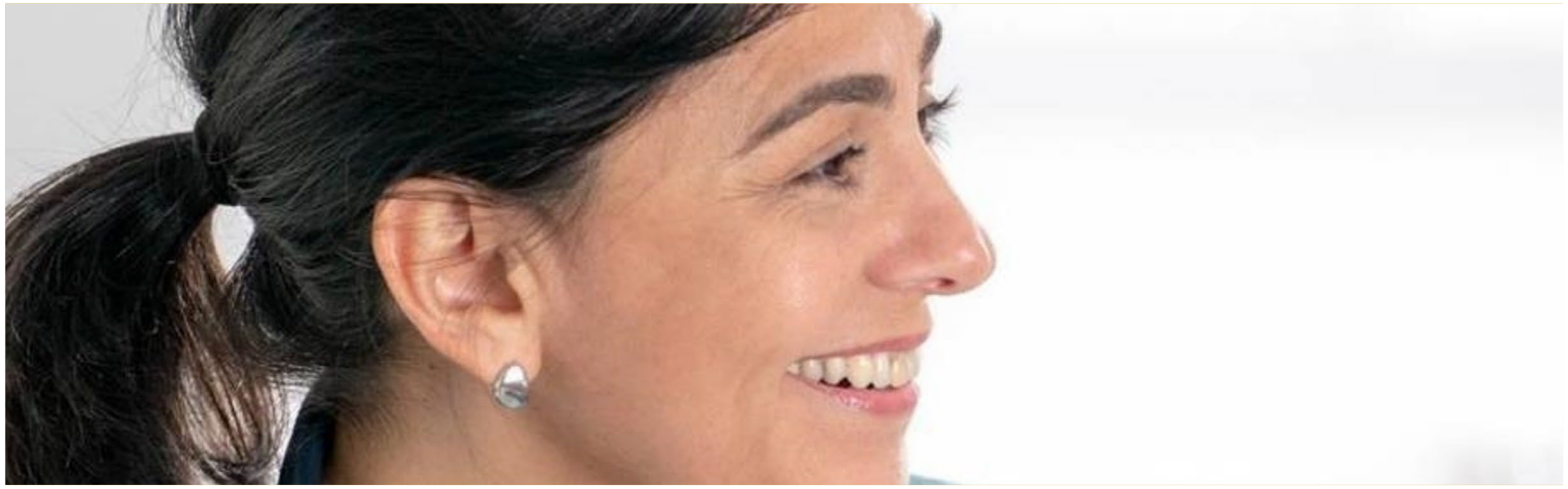
- **BMTs have serious risks and low survival rate**



Biology of Blood and Marrow Transplantation, Volume 12, Issue 4, 422 - 429

DATASET INFO

- Hospital in Poland, from 2000 to 2010
- **187 Pediatric Patients** who underwent BMTs
  - All data are iid
- **37 Attributes**
  - Pre-op Parameters (e.g. Age, Gender, Blood Type)
  - Post-op Assessments (e.g. *G*raft *v*ersus *H*ost *D*isease)
- **Missing data**
  - Fraction of Features: 40 % (15 columns)
  - Fraction of Points: 85 %
- **Binary Classification** Problem
  - 1 Year after Surgery, **Alive** or **Dead**

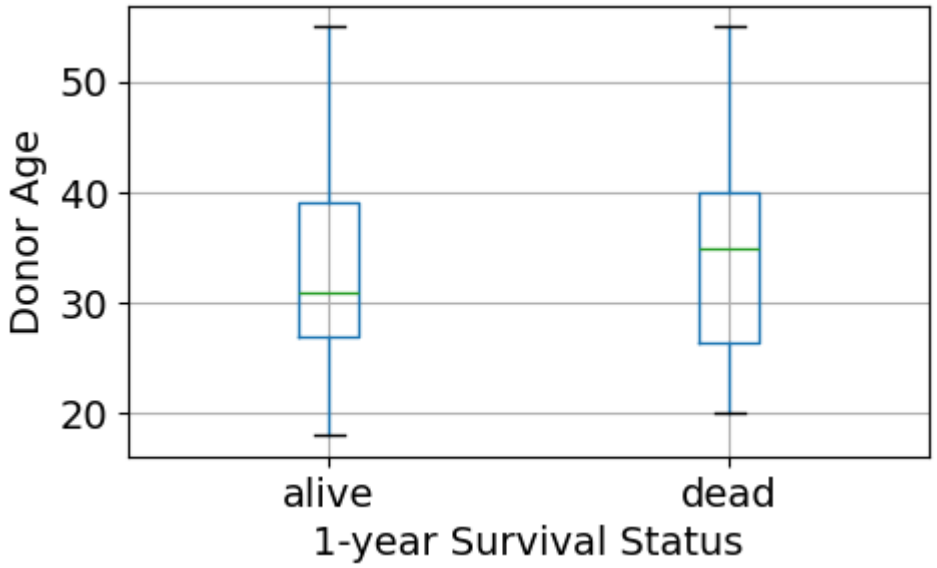


Source: Sikora, M., Wróbel, Ł., & Gudyś, A. (2020). Bone marrow transplant: children [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5NP6Z.

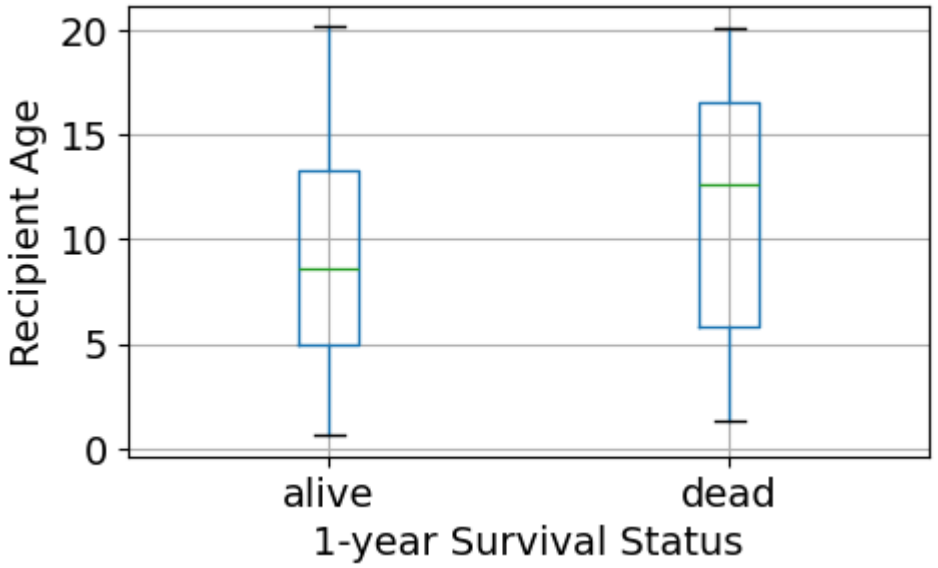
EDA interesting findings:

Pre-op

Donor Age



Recipient Age

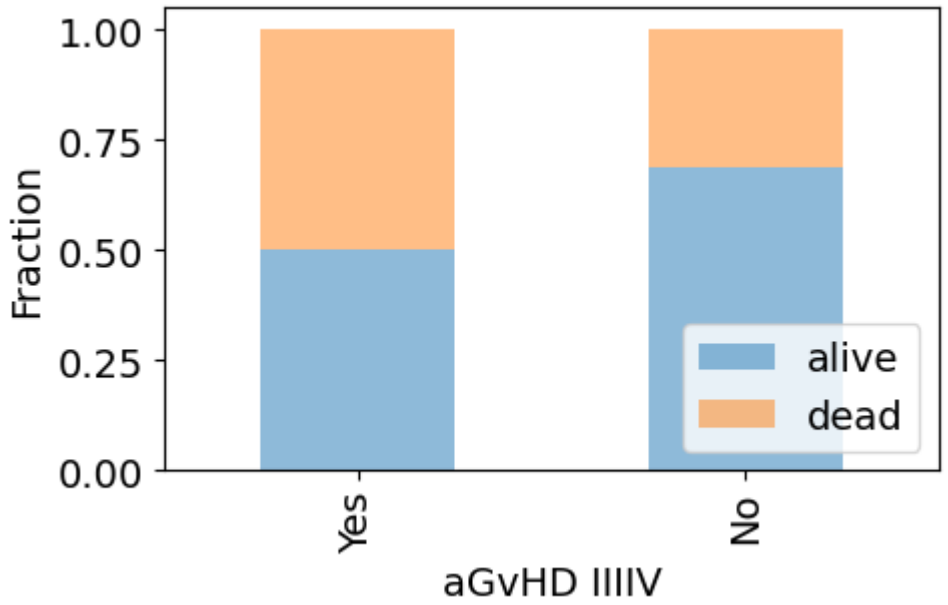


Lower age of either donors or recipient worked better

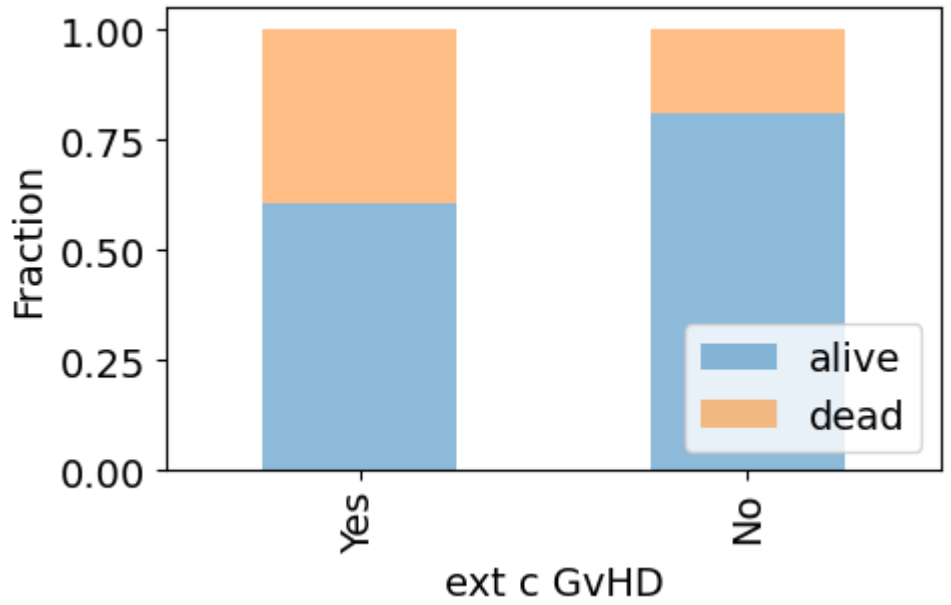
EDA interesting findings:

Post-op

Development of acute GvHD stage III or IV



Development of extensive chronic GvHD



Acute or extensive GvHD increased the risk of death, but it was not always fatal.

*G*raft *v*ersus *H*ost *D*isease (GvHD): one of the most common side effects of BMTs

High Correlation (Redundant) Data Reduction

Donor age, ~~Donor age 35~~;

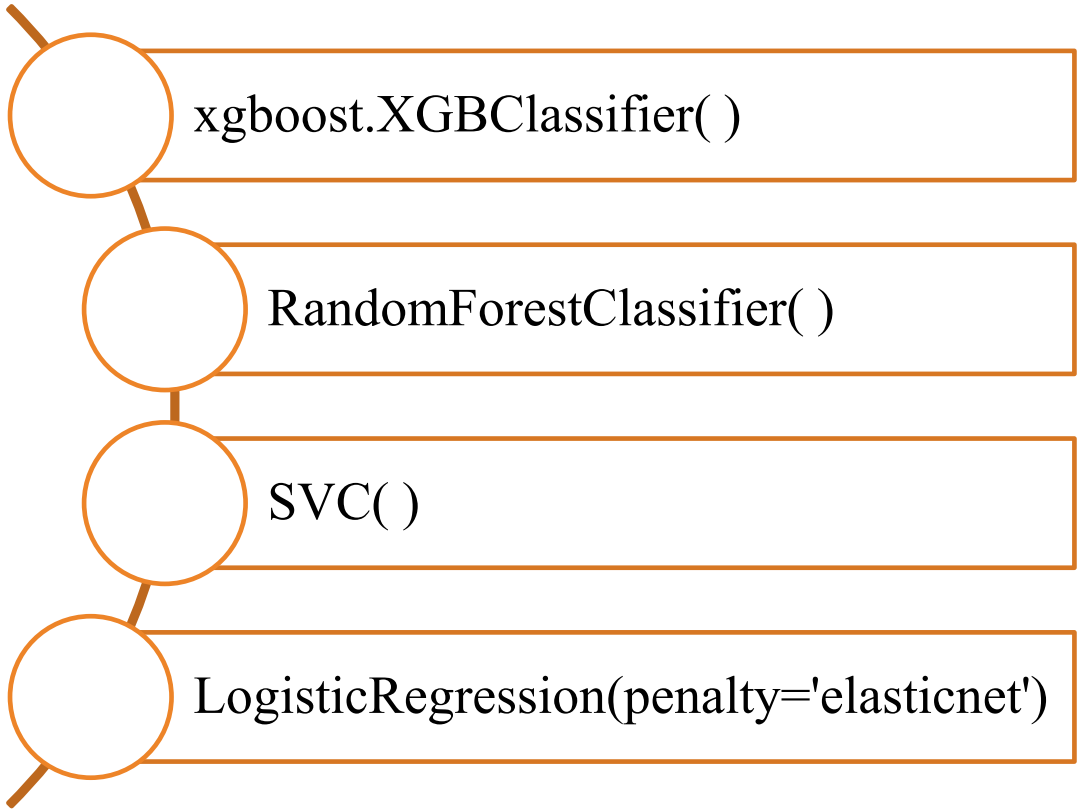
Recipient age, ~~Recipient age 10, Recipient age int~~

OneHotEncoder(**drop="first"**)

X before preprocessing (187, 33)

X after preprocessing (187, 43)

Machine Learning Models



Naturally deal with missing data

Continuous Missing data processing

- Type 1:
  - Time to certain side effects
    - Missing data means no side effects
    - **Transformed to bool**
- Type 2:
  - Recipient HSC before surgery
    - Reasons unknown, only 2 rows
    - **Dropped the 2 rows**

Data Splitting

Step 1: separate out the test set

```
train_test_split(X, y, test_size = 0.2, random_state = random_state*42, stratify = y)
```

Step 2: Cross Validation

```
StratifiedKFold(n_splits=4, shuffle=True, random_state = random_state*42)
```

CV Pipeline (except XGB)

```
make_pipeline(preprocessor, StandardScaler(), machine_learning_model)

GridSearchCV(pipeline, param_grid=param_grid, cv=kfold, scoring='f1',
return_train_score = True,n_jobs=-1,verbose=True)

grid.fit(X_other, y_other)
grid.score(X_test,y_test))
```

CV Pipeline for XGBoost

Loop 1: random state * 10

train_test_split + StratifiedKFold

Loop 2: ParameterGrid(param_grid)

xgboost.XGBClassifier().set_params(**ParameterGrid(param_grid), early_stopping_rounds = 50)

Loop 3: StratifiedKFold

preprocessing (fit X_train, transform X_train and X_CV)

XGB.fit(X_train_prep,y_train,eval_set=[(X_CV_prep, y_CV)])

predict Kth fold

calculate CV score (*4) mean and std

sort out the best model by the best CV score mean

best_model.predict(X_test)

Hyperparameter Tuning

XGB

```
"learning_rate": [0.01, 0.03, 0.1, 0.3],
"reg_alpha": [0e0, 1e-2, 1e-1, 1e0, 1e1, 1e2],
"reg_lambda": [0e0, 1e-2, 1e-1, 1e0, 1e1, 1e2],
"max_depth": [1,3,10,30,100]

'svc__C': [0.01, 0.1, 1, 10, 100],
'svc__gamma': [0.01, 0.1, 1, 10, 100]}

'randomforestclassifier__max_depth': [1, 3, 10, 30, 100]
'randomforestclassifier__max_features': [0.25, 0.5,0.75,1.0]

'logisticregression__C': [0.01, 0.1, 1, 10,100],
'logisticregression__l1_ratio': [0, 0.2,0.4,0.6,0.8, 1]
```

Evaluation Matric (Baseline Calculation)

$$f1 = \frac{2\ \mathrm{TP}}{2\ \mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

Low FP

- Medical Resources are limited and expensive.
- We don't want non-dying patients occupying too many.

Low FN

- Lives are important.
- We don't want dying patients missing the additional chance to be treated.

y_true \ y_pred	0	1
0	TN	FP
1	FN	TP

Stratified train_test_split
y_test_0 = 24
y_test_1 =13

y_true \ y_pred	0	1
0	0	24
1	0	13

$$f1_{baseline} = 0.52$$

Test Scores *10 (Mean And Std)

$f1_{baseline} = 0.52$

XGBClassifier( )

mean	std
0.60	0.11

RandomForestClassifier( )

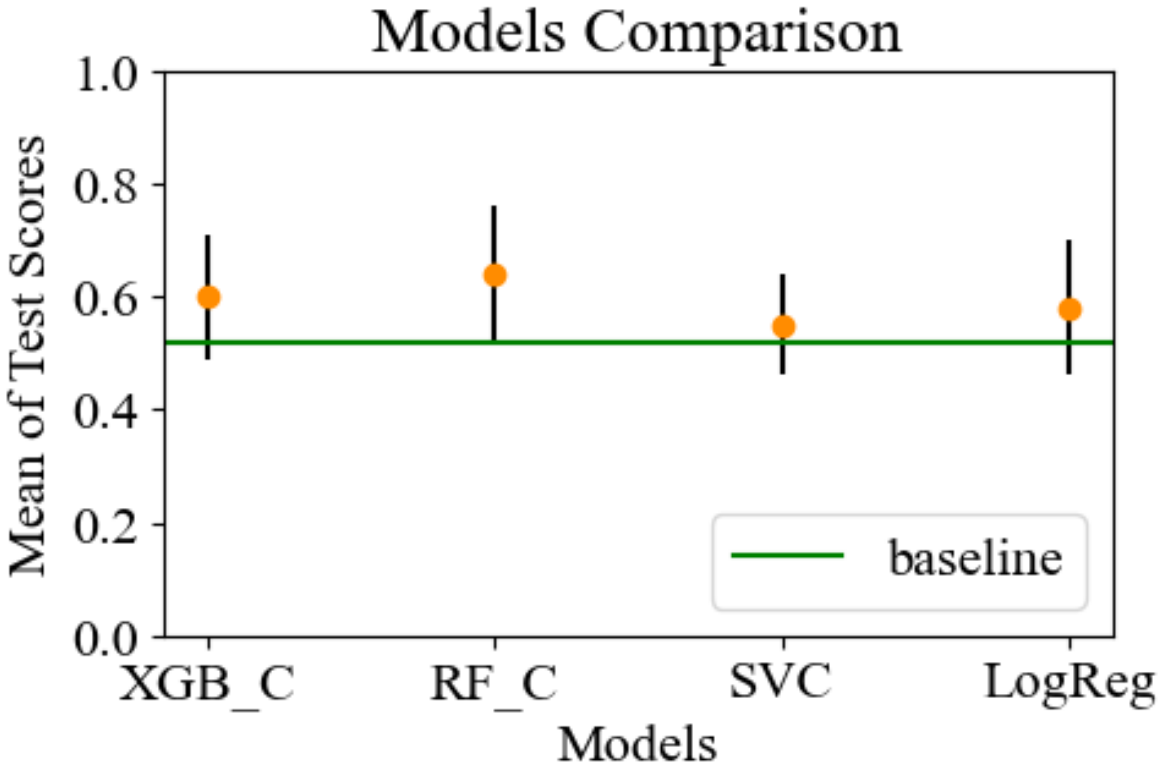
mean	std
0.64	0.12

SVC( )

mean	std
0.55	0.09

LogisticRegression (penalty='elasticnet')

mean	std
0.58	0.12



Test Scores *10 (Mean And Std)

$f1_{baseline} = 0.52$

XGBClassifier( )

mean	std
0.60	0.11

RandomForestClassifier( )

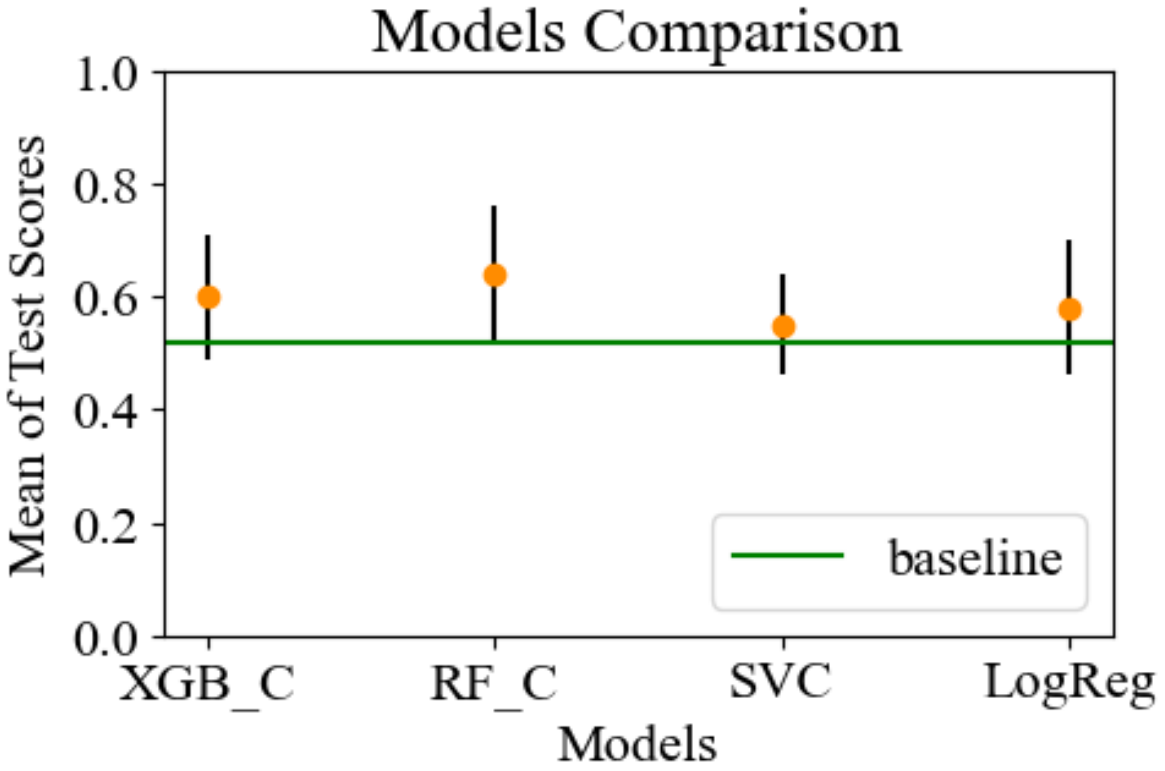
mean	std
0.64	0.12

SVC( )

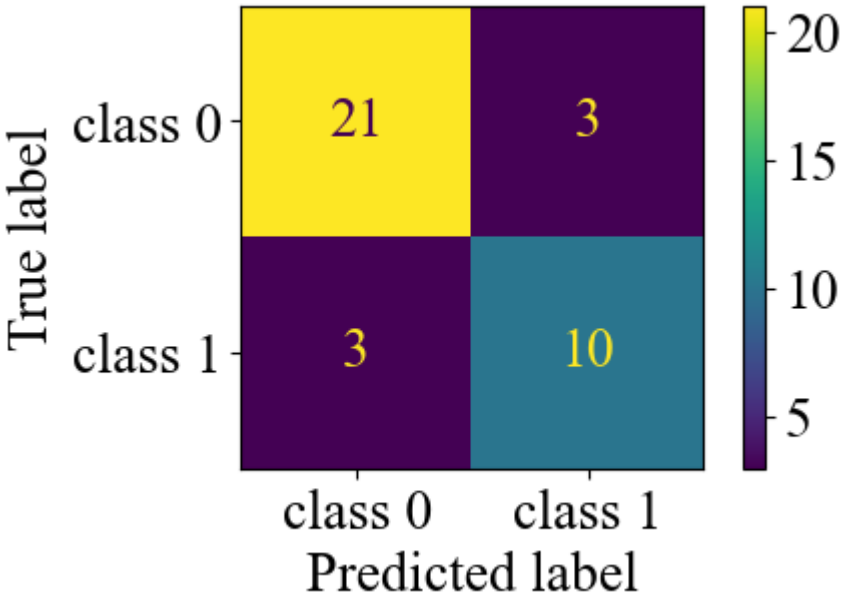
mean	std
0.55	0.09

LogisticRegression (penalty='elasticnet')

mean	std
0.58	0.12



Confusion Matrix



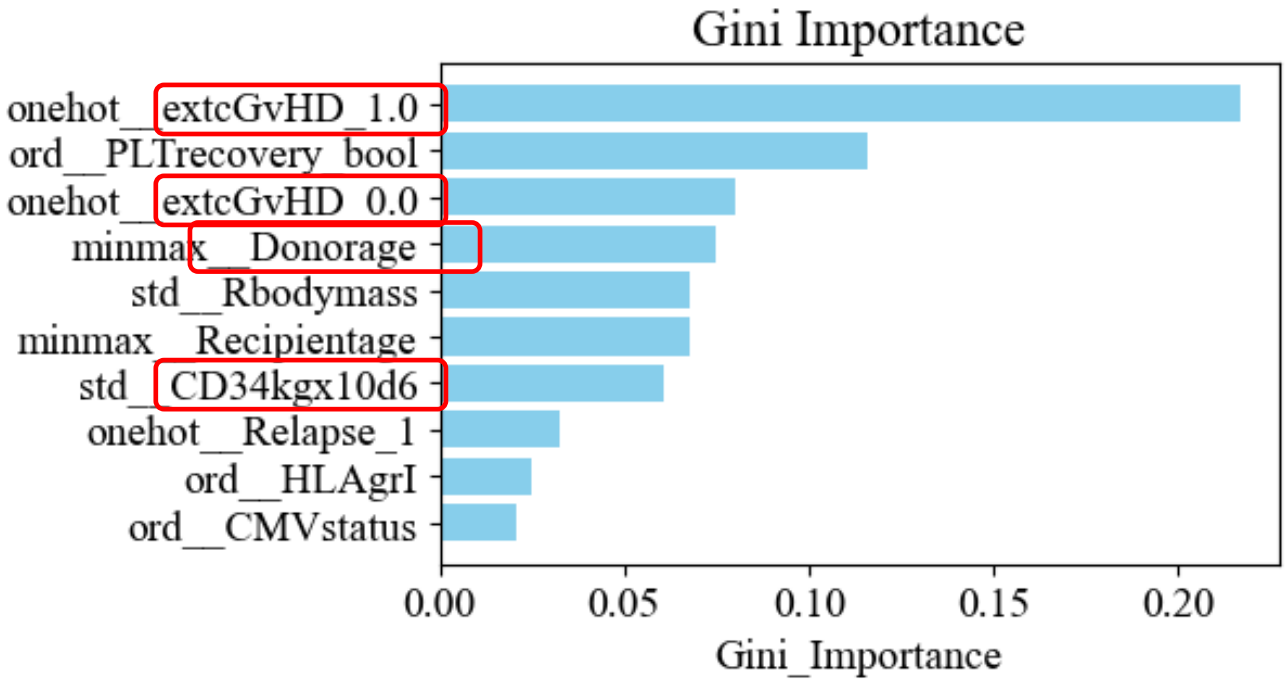
RandomForestClassifier

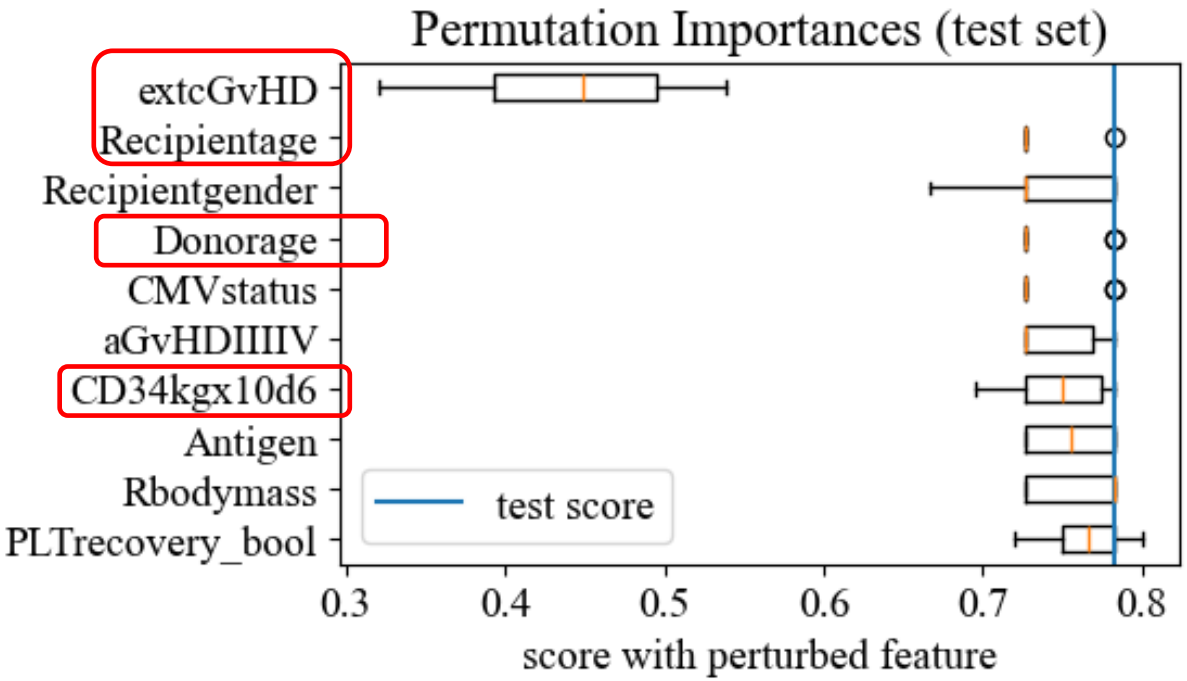
max_depth=3,
max_features=0.5

Test Score: 0.75

Random state = 42*2

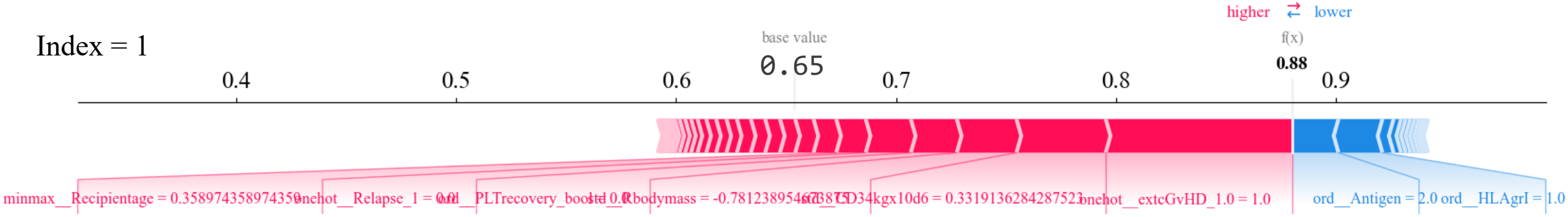
Global Feature Importances (of the Random Forest Classifier)



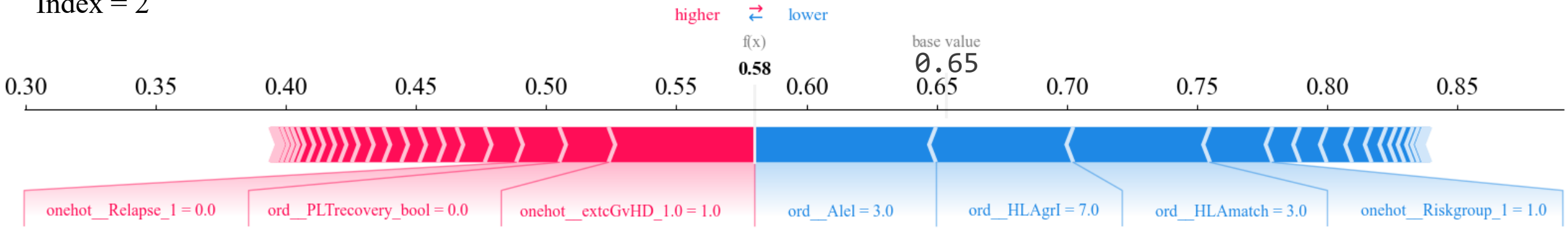


Local Feature Importances (SHAP, For class 0 ‘alive’)

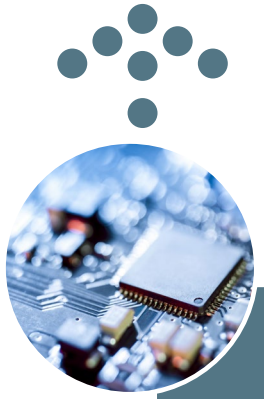
Index = 1



Index = 2



Outlook on Better Predictive Power



More Models

Try more models. Tune the F_beta a little more and see how it affects the result.

Larger Dataset

Search for more data to get a better statistical understanding of the problem. Or potentially collaborate with a hospital.

Domain Knowledge

Consult with medical experts to identify key features relevant to survival outcomes, such as: Graft source, donor-recipient match, pre-transplant health status.

Q&A

Zoey Yu

Dec 13, 2024