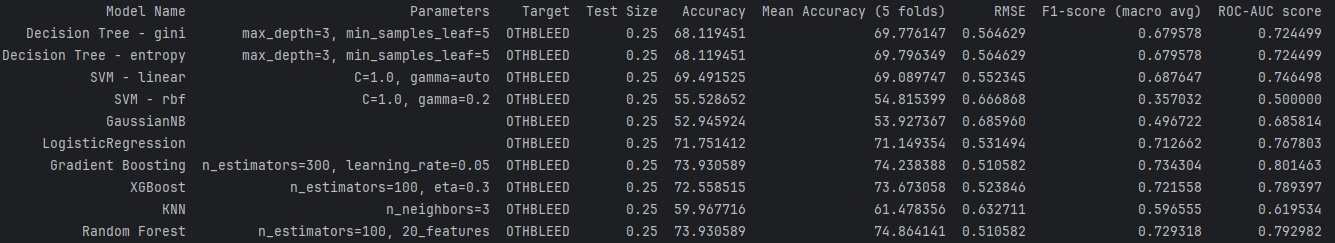
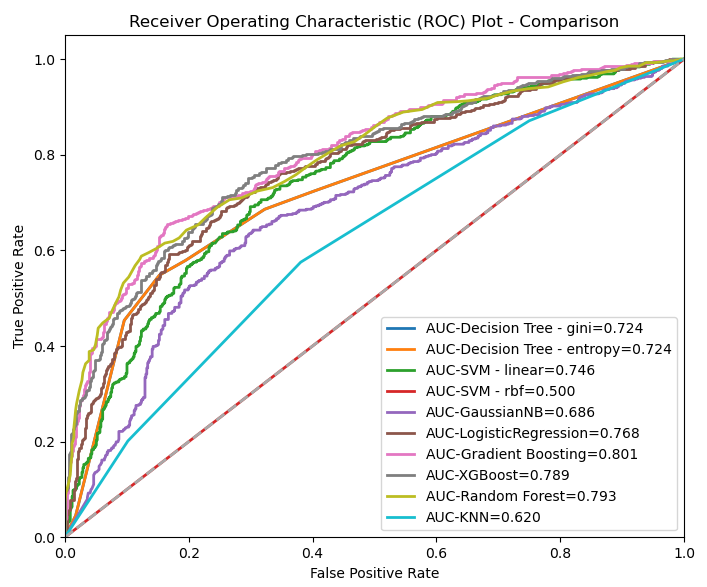
**Model Notes**

# Iteration 1: benchmark

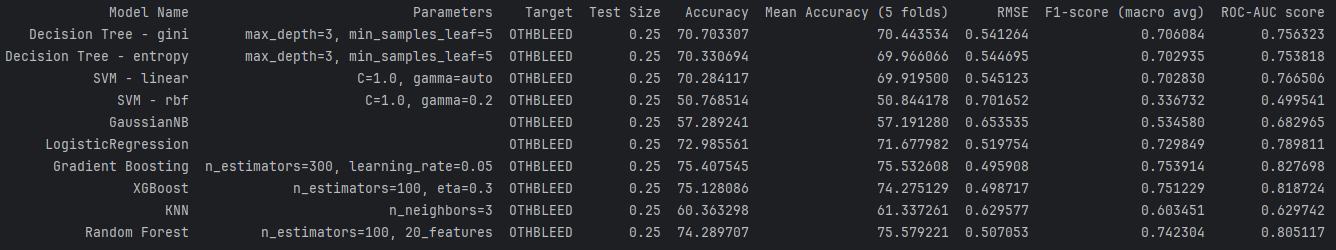
* **Year 2018-2020**
* all features with less than 50% missing
* dropped ‘NOTHBLEED’, and ‘DOTHBLEED’ due to high collinearity with the target
  + NOTHBLEED(r= -0.99) : Number of Bleeding Transfusions Occurrences
  + DOTHBLEED(r=-0.81): Days from Operation until Bleeding Transfusions Complication
  + OTHBLEED (the target): Occurrences Bleeding Transfusions
* **features: 126**, observations: 4953 (shape: 4953x129)
* imputations applied, no standardization
* dataset: [CABG\_2018\_2022.csv](https://github.com/jennytsai32/Capstone/tree/master/code/main_code/processed_data/2018_2022) (slice only 2018-2020)

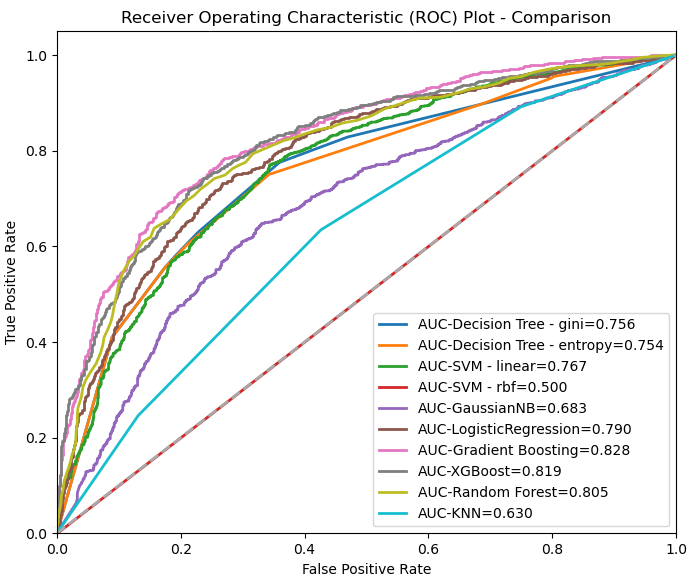




# Iteration 2: more recent data – extended to 2022

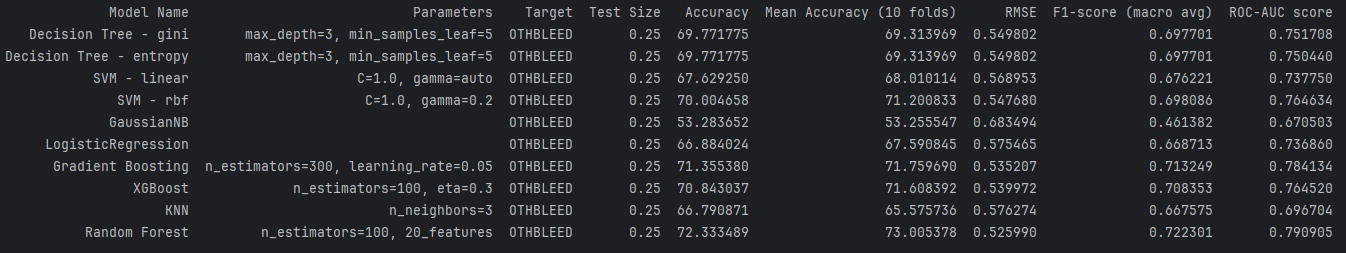
* Year 2018-2022
* all features with less than 50% missing
* **features: 126**, observations: 8587 (shape: 8587x129)
* imputations, no standardization
* dataset: dataset: [CABG\_2018\_2022.csv](https://github.com/jennytsai32/Capstone/tree/master/code/main_code/processed_data/2018_2022)

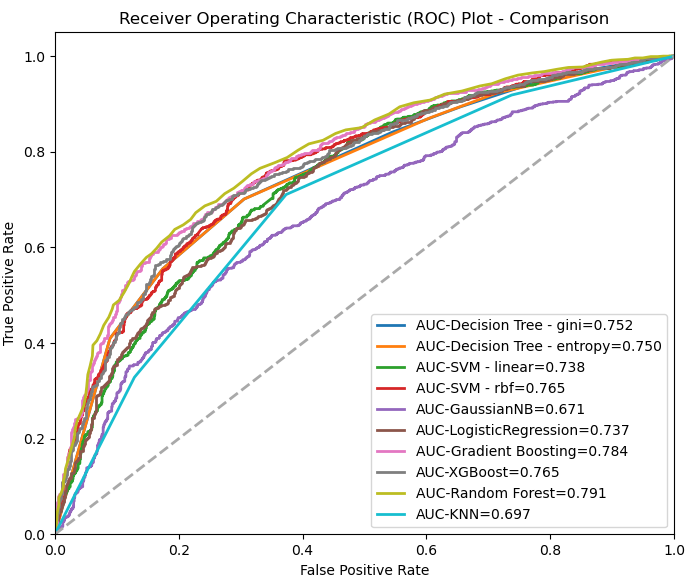




# Iteration 3: selected features; standardization; cross-validation

* Year 2018-2022
* **features: 40**, observations: 8587 (shape: 8587x41)
  + Dr. Gupta selected 43 features
  + 'HEIGHT','WEIGHT','ETHNICITY\_HISPANIC' was dropped for obvious multicollinearity issue (see graph below)
* imputations, standardization, 10-folds cross-validation
* dataset: [GABG\_5yr\_preselect40.csv](https://github.com/jennytsai32/Capstone/tree/master/code/main_code/processed_data/2018_2022)





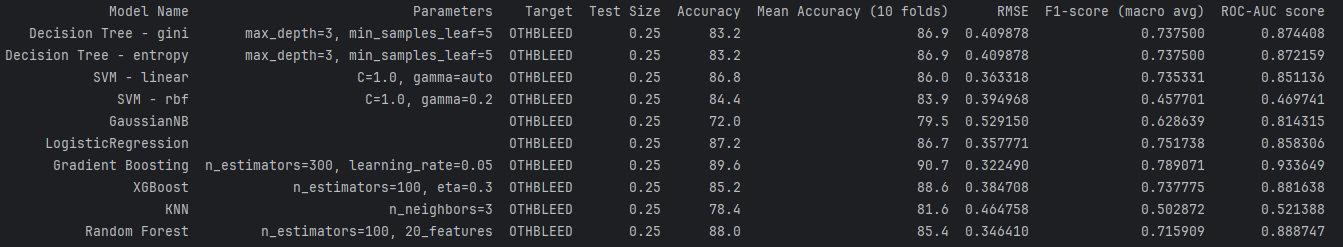
# Iteration 4: AutoFeat

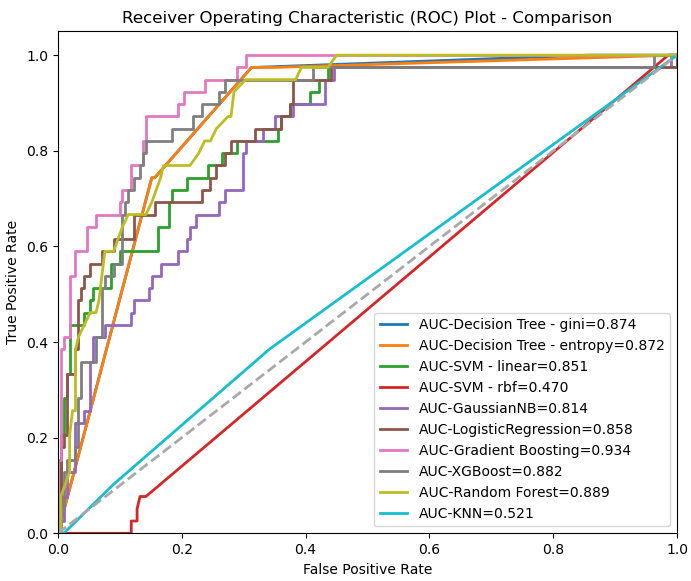
* Year 2018-2022
* imputations, standardization, 10-folds cross-validation
* **features: 40**, observations: 8587 (shape: 8587x41)
  + derived from f\_importances results in the Random Forest model
* used **AutoFeat** package

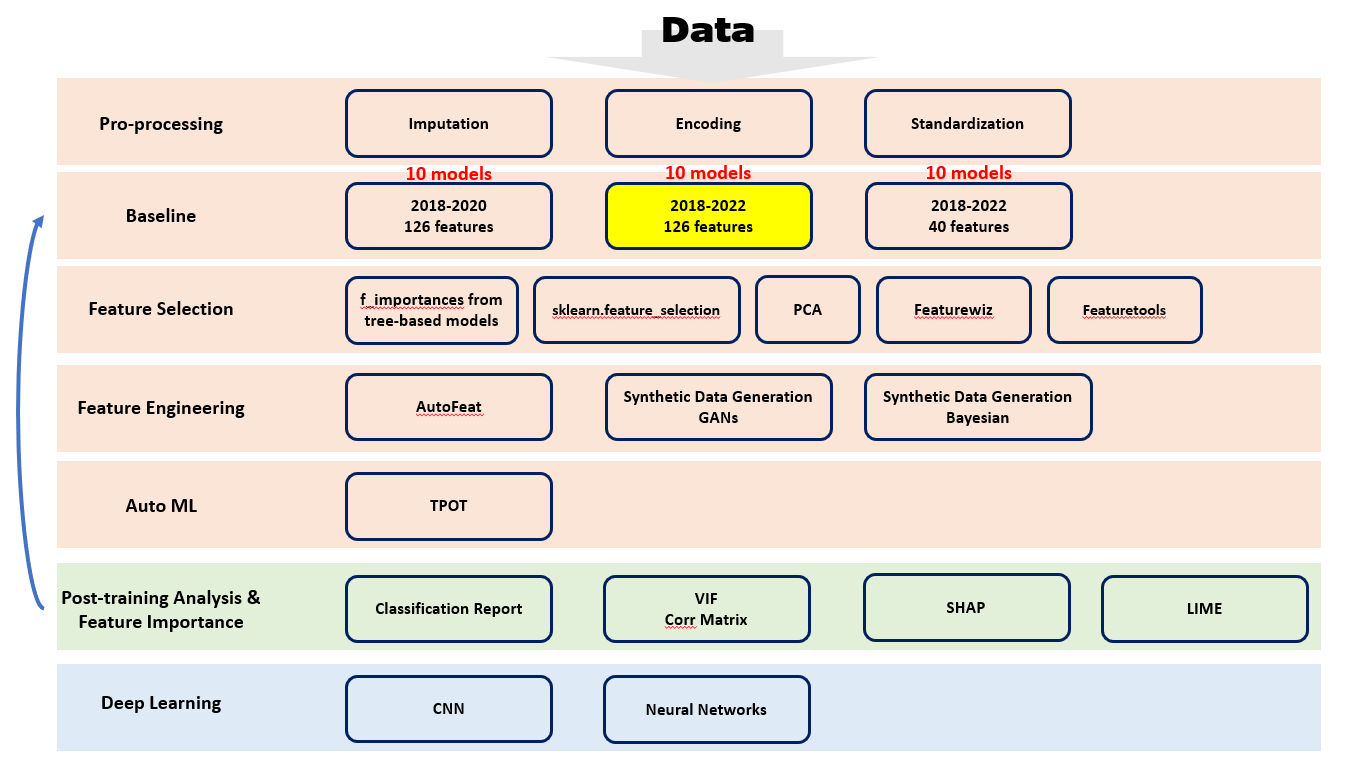
# Iteration 5: Synthetic data generation – GANs

# Iteration 6: Synthetic data generation – Bayesian

* Year 2018-2022
* imputations, standardization, 10-folds cross-validation
* **features: 40**, observations: 1000 (shape: 1000x41)
  + derived from f\_importances results in the Random Forest model
* used **DataSynthesizer** package (display\_bayesian\_network)

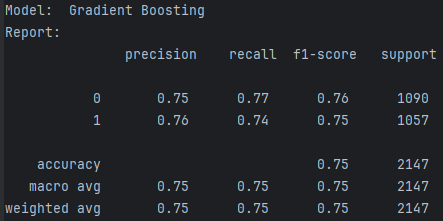






# Post-training Analysis

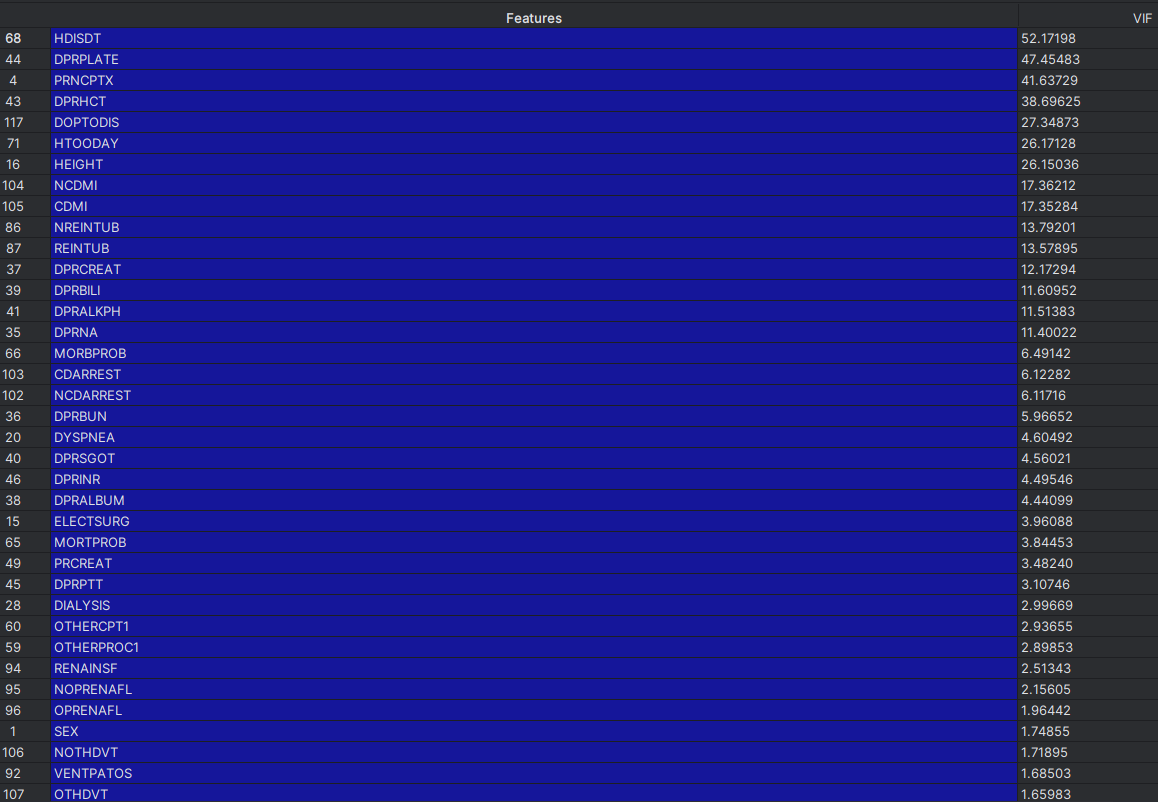
**Step 1: Classification Report – Gradient Boosting**

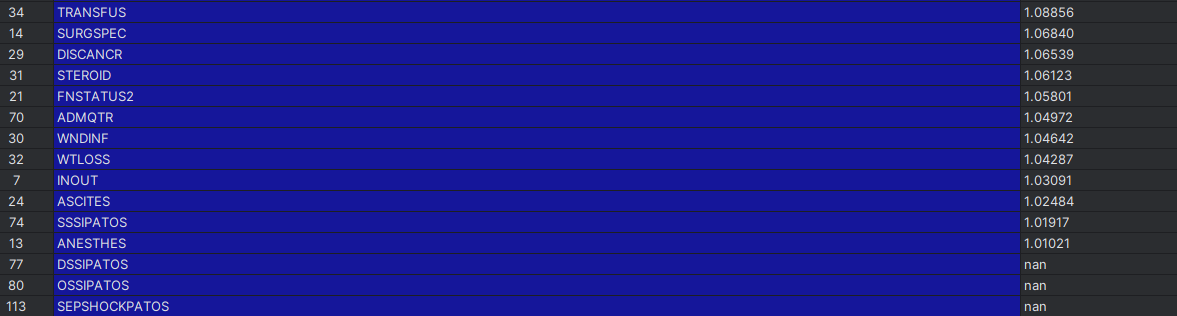
****

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Target** | **y\_train** | **y\_test** |
| **0** | 4366 | 3164 | 1090 |
| **1** | 4221 | 3276 | 1057 |

**Step 2: VIF**

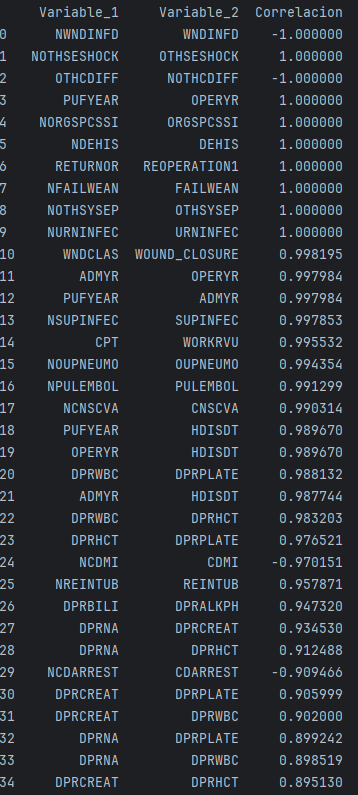
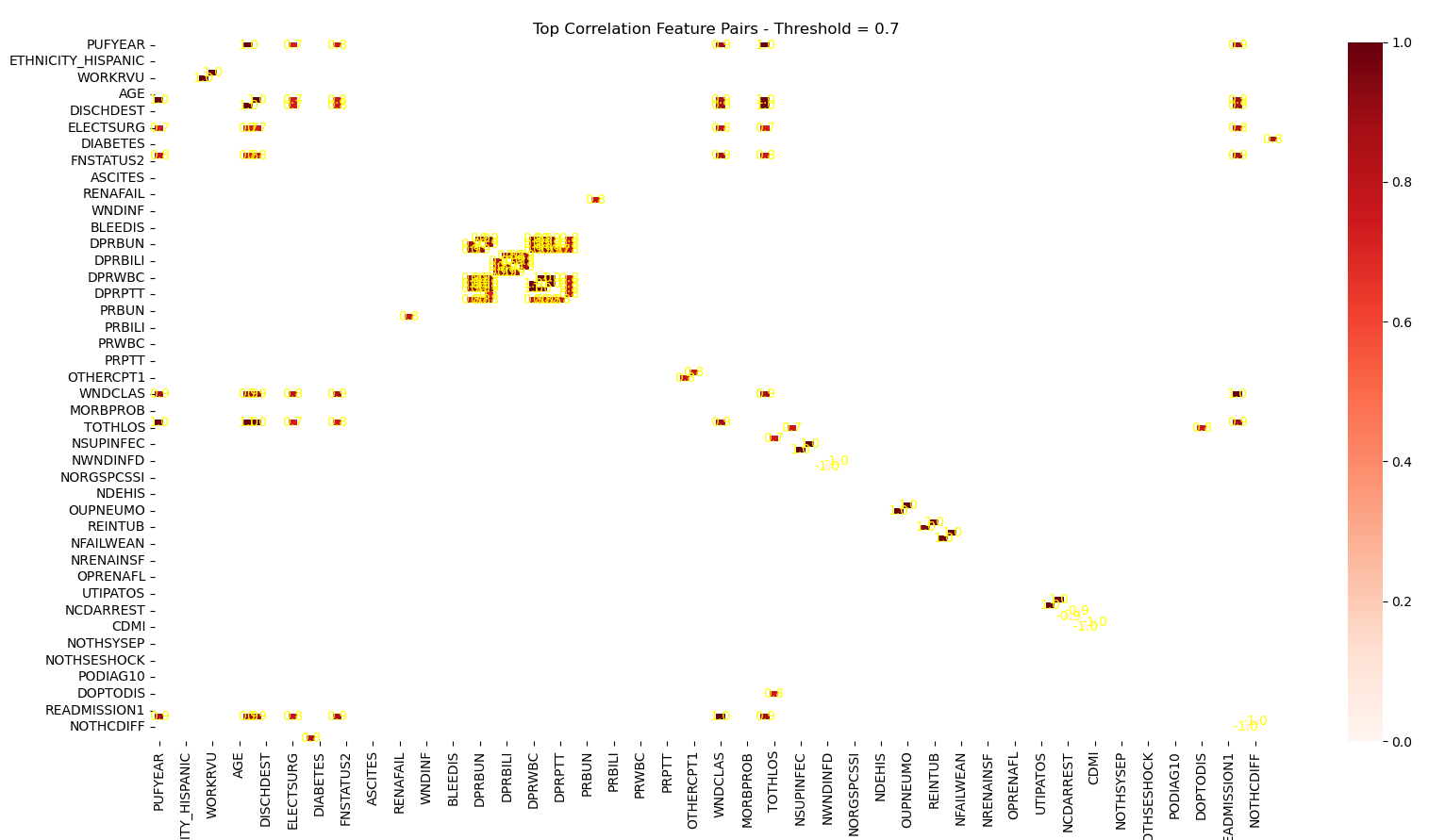
Features (126) VIF Ranking

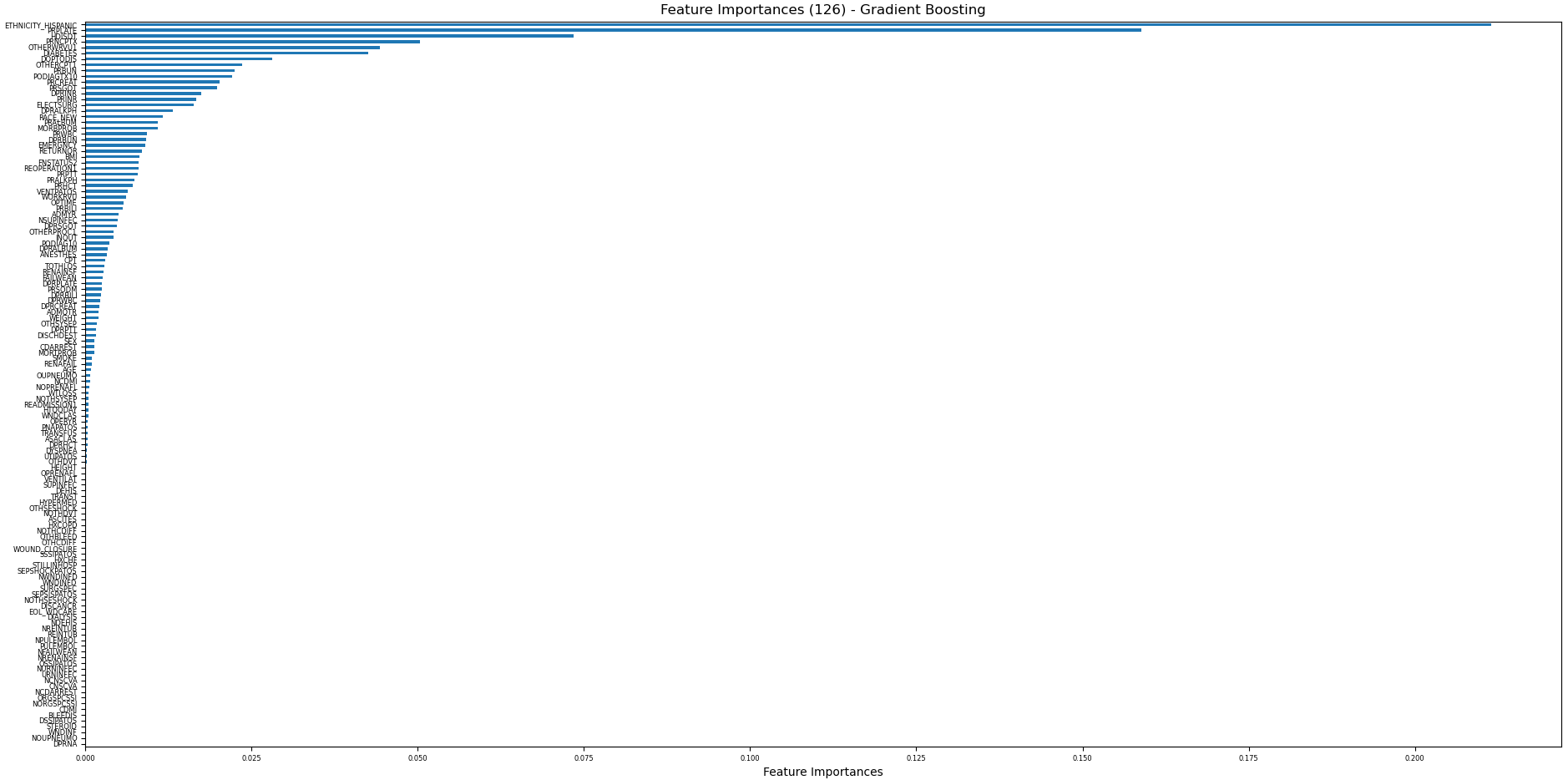
 

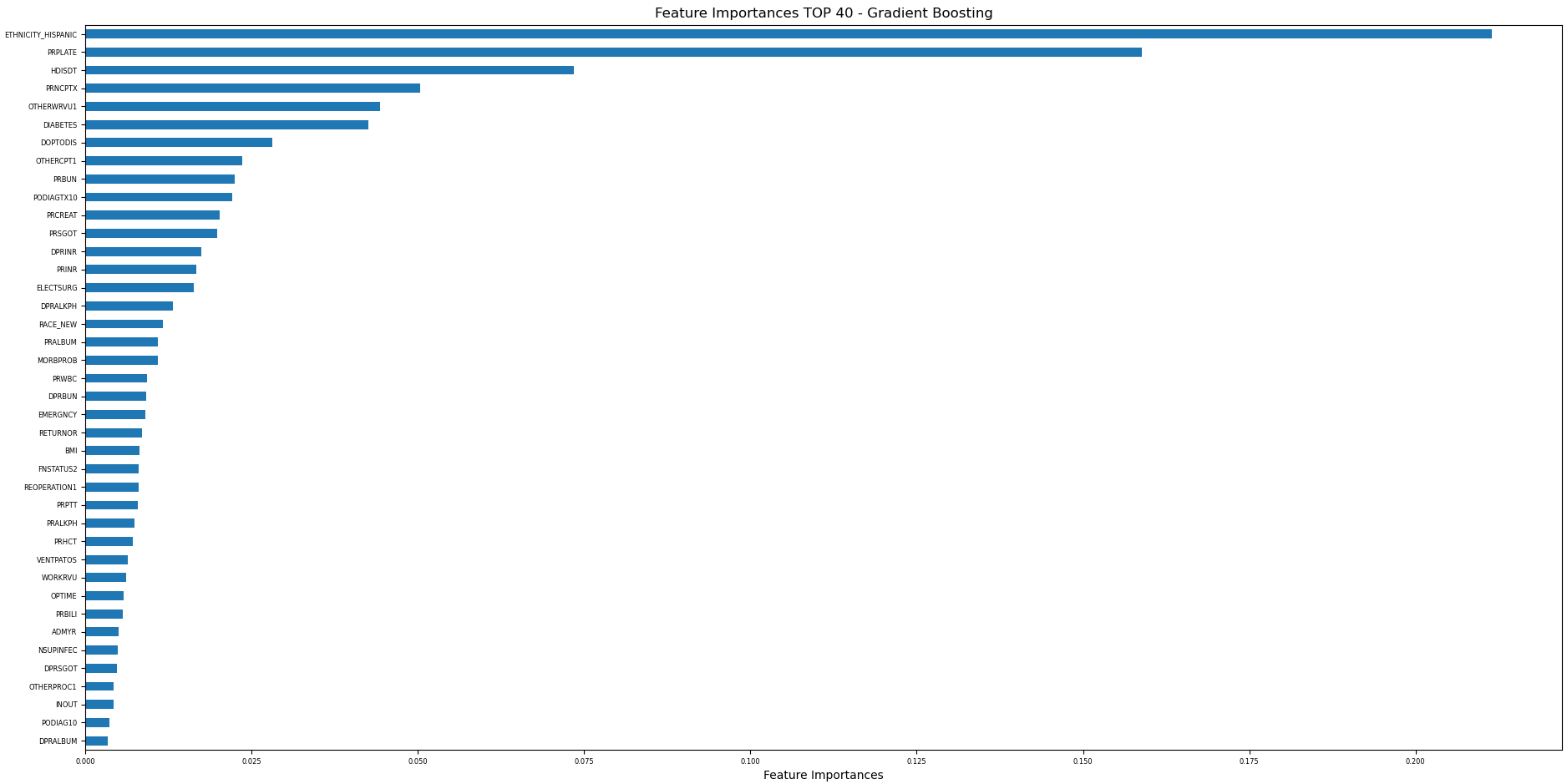
Step 3: Correlation Coefficients between features

Features Correlation Pairs Ranking

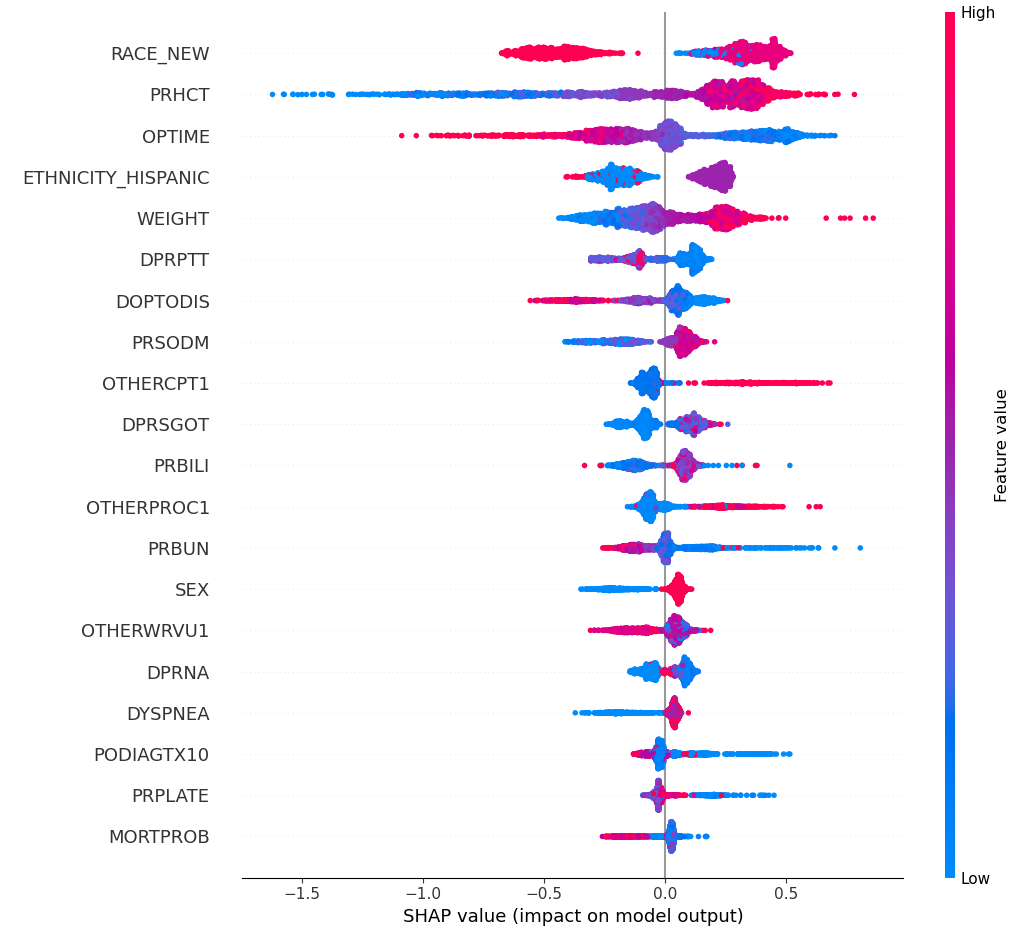
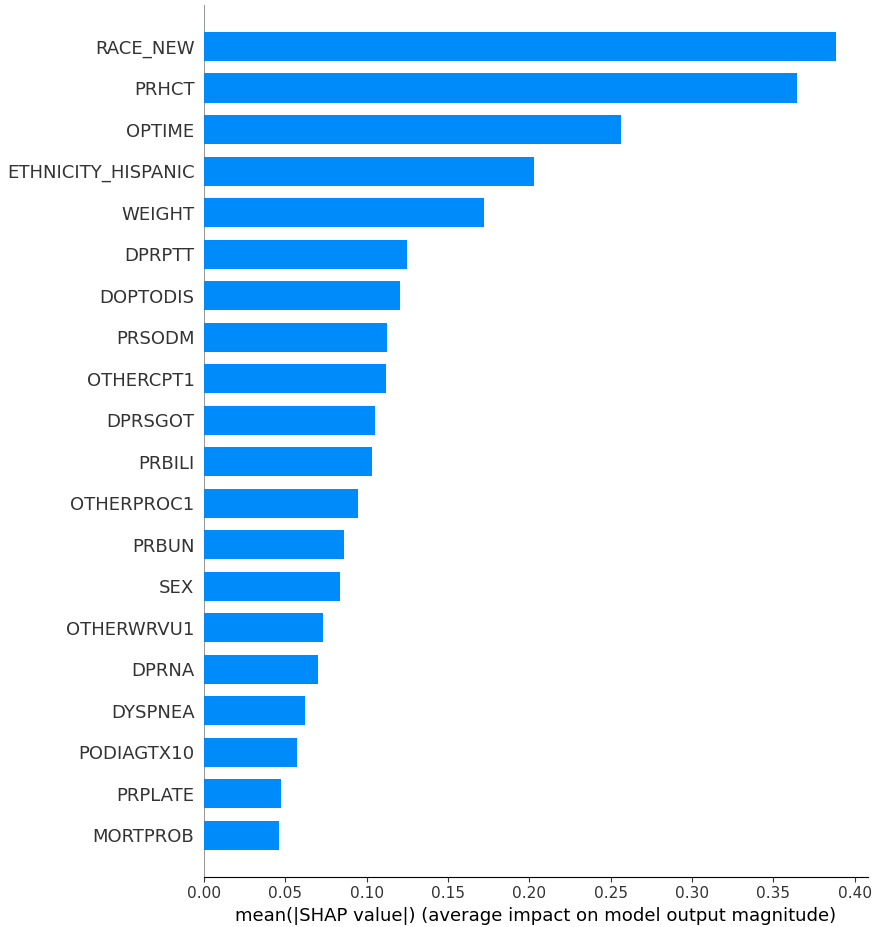
 

**Step 2: Features Importances (from Gradient Boosting model)**



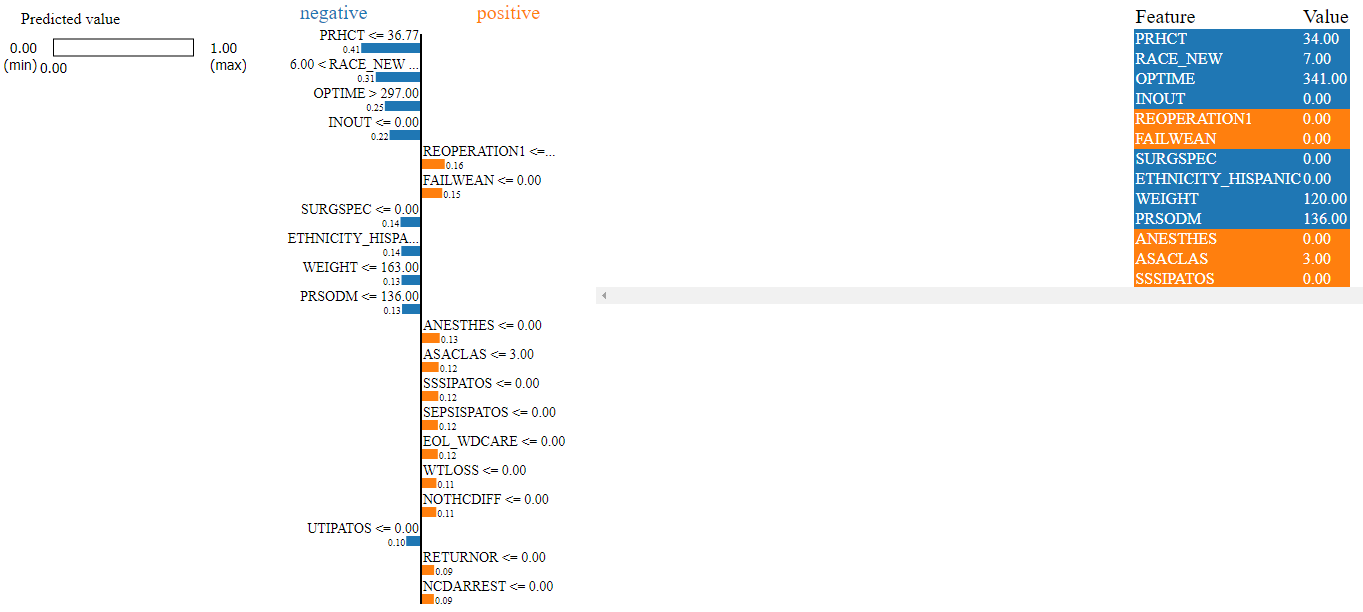


Step 3: SHAP

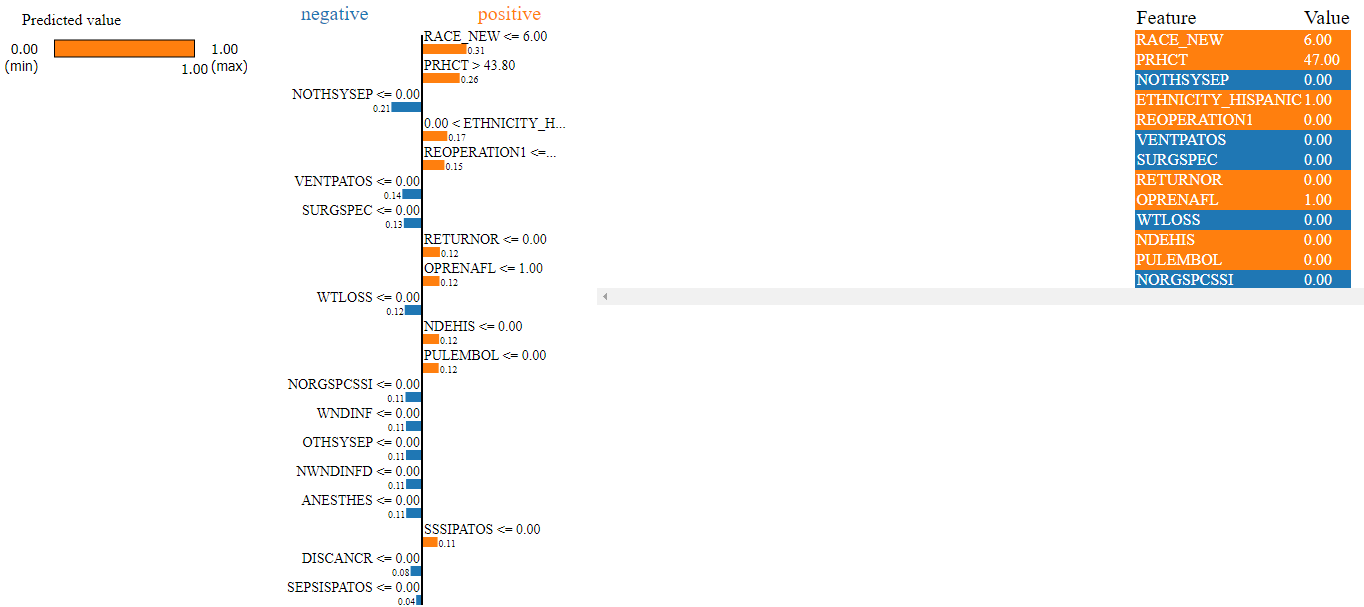


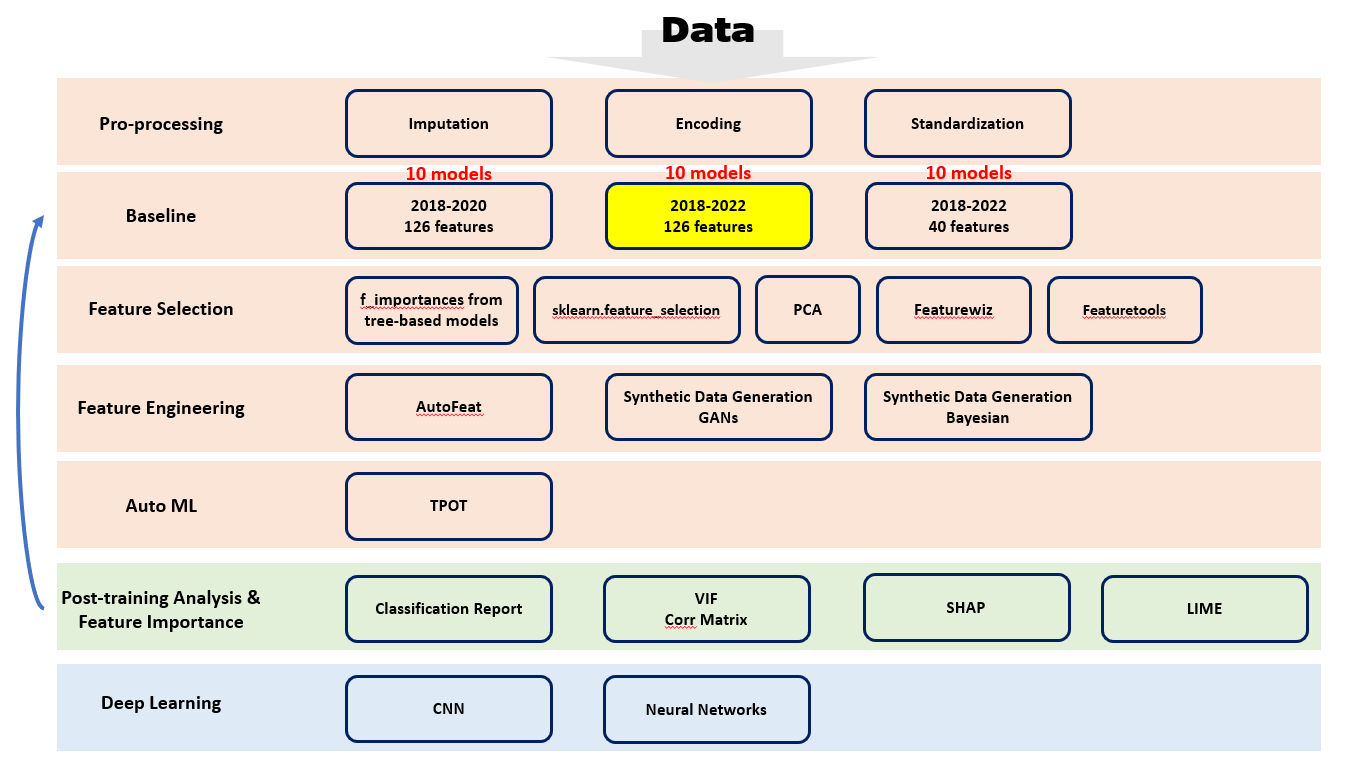
Step 4: LIME

***Observation #5 in X\_test***

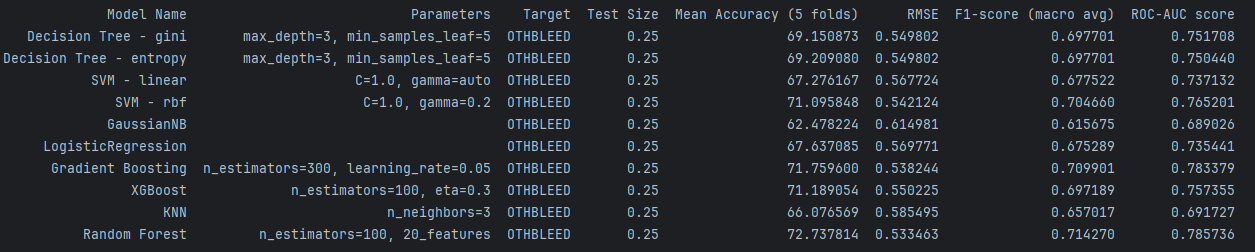


***Observation #150 in X\_test***



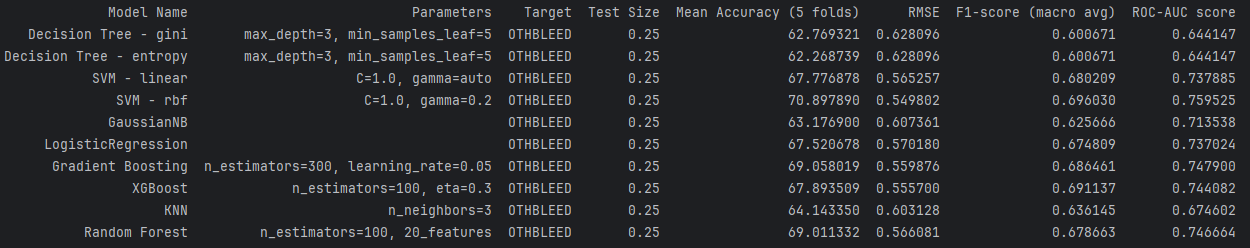


* Year 2018-2022
* features: 20, observations: 8587 (shape: 8587x21)
  + Dr. Gupta selected 43 features
  + Picked top 20 from feature importances (random forest)



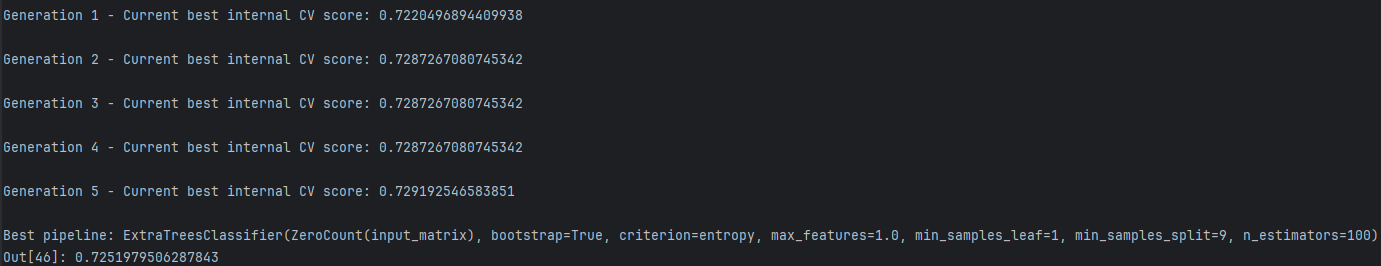
# Iteration 5: Feature Selection - PCA

* Run PCA on 20 features from iteration #4
* 19 features included (feature dimension reduced 1)



# Iteration 6: Feature Engineering - TPOT

* Run TPOT on 40 features dataset – the best model so far
* 0.25 on test size, generations=5, population\_size=20, verbosity=2



# Iteration 7: Feature Engineering - AutoFeat

=================

# Feature Selection

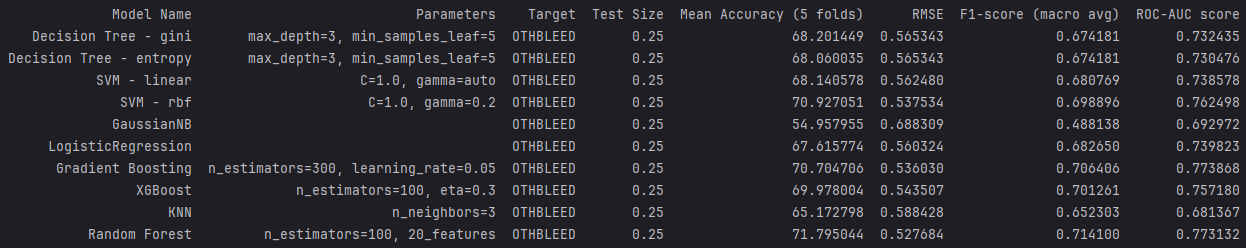
* sklearn.feature\_selection
* featuretools
* featurewiz

# Post-Training Analysis

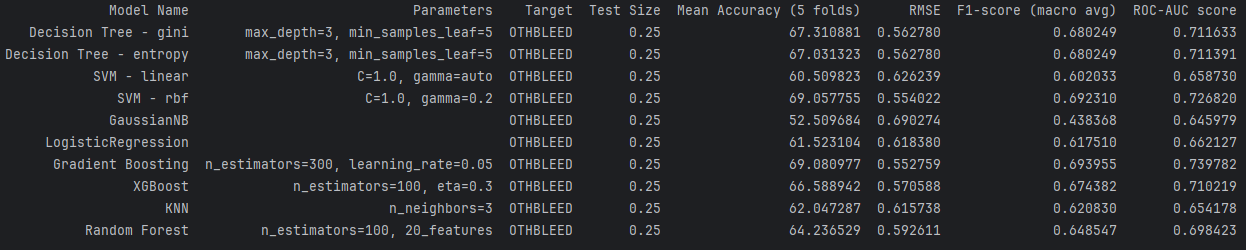
* performance by class; check 0 and 1 balance
* Feature importance analysis - FI estimates the contribution of each feature (independent parameter) to the model output
* Tree-based models like Random **Forest, Gradient Boosting, Log** provide feature importances as a built-in feature.
* SHAP (SHapley Additive exPlanations)
  + shap library
* LIME (Local Interpretable Model-agnostic Explanations)
  + lime library
* Partial Dependence Plots (PDP)
  + Partial dependence plots show how the predicted outcome changes with variations in a particular feature while accounting for the average effects of all other features.
  + **pdpbox** and **sklearn.inspection** library in scikit-learn
* Feature Correlation Analysis
  + Corr matrix
  + Heatmap

BACK-UP Notes

43 features



Jenny’s 20 features after random forest



**Step 2: Features Importances (from Random Forest model)**

