**DATA PREPROCESSING**

2018-2020 vs 2021:

**New columns**: ['PREOP\_COVID', 'POSTOP\_COVID', 'IMMUNO\_CAT', 'OXYGEN\_SUPPORT', 'HEMO', 'CASETYPE', 'HOMESUP', 'HXFALL', 'HXDEMENTIA', 'DELIRIUM', 'DISHOMESVC', 'DISFXNSTAT']

**Removed columns:** ['ELECTSURG', 'DYSPNEA', 'RENAFAIL', 'WNDINF', 'WTLOSS', 'DPRPT', 'PRPT', 'EMERGNCY', 'WNDCLAS', 'NRENAINSF', 'RENAINSF', 'DRENAINSF', 'PODIAG', 'PODIAGTX', 'REOPOR1ICD91', 'REOPOR2ICD91', 'READMRELICD91', 'READMUNRELICD91', 'READMRELICD92', 'READMUNRELICD92', 'READMRELICD93', 'READMUNRELICD93', 'READMRELICD94', 'READMUNRELICD94', 'READMRELICD95', 'READMUNRELICD95', 'WOUND\_CLOSURE', 'PODIAG\_OTHER']

**Same name different values:** ['RACE\_NEW', 'TRANST', 'DISCHDEST', 'ANESTHES', 'SURGSPEC', 'PRSEPIS', 'OTHERPROC1', 'OTHERCPT1', 'OTHERPROC2', 'OTHERPROC3', 'OTHERPROC4', 'OTHERPROC5', 'OTHERPROC6', 'OTHERPROC7', 'CONCURR1', 'CONCURR2', 'CONCURR3', 'CONCURR4', 'CONCURR5', 'ASACLAS', 'SSSIPATOS', 'PODIAG10', 'PODIAGTX10', 'RETORRELATED', 'REOPOR1ICD101', 'REOPERATION2', 'RETOR2RELATED', 'REOPOR2ICD101', 'REOPERATION3', 'UNPLANNEDREADMISSION1', 'READMRELATED1', 'READMSUSPREASON1', 'READMRELICD101', 'READMISSION2', 'UNPLANNEDREADMISSION2', 'READMRELATED2', 'READMSUSPREASON2', 'READMRELICD102', 'READMISSION3', 'UNPLANNEDREADMISSION3', 'READMRELATED3', 'READMSUSPREASON3', 'READMRELICD103', 'PODIAG\_OTHER10', 'ANESTHES\_OTHER']

**Preselect cols removed:** ['DYSPNEA', 'RENAFAIL', 'WNDINF', 'WTLOSS', 'PRPT', 'EMERGNCY']

**Preselect cols with different values:** ['RACE\_NEW', 'ANESTHES', 'ASACLAS']

**Cols value comparison:**

RACE\_NEW

df1: ['American Indian or Alaska Native', 'Asian', 'Black or African American', 'Native Hawaiian or Other Pacific Islander', 'Native Hawaiian or Pacific Islander', 'Some Other Race', 'Unknown/Not Reported', 'White']

df2: ['American Indian or Alaska Native', 'Asian', 'Black or African American', 'Native Hawaiian or Other Pacific Islander', 'Some Other Race', 'Unknown/Not Reported', 'White']

ANESTHES

df1: ['General', 'MAC/IV Sedation', 'Other']

df2: ['General']

ASACLAS

df1: ['1-No Disturb', '2-Mild Disturb', '3-Severe Disturb', '4-Life Threat', '5-Moribund', 'None assigned']

df2: ['2-Mild Disturb', '3-Severe Disturb', '4-Life Threat', '5-Moribund', 'None assigned']

**Combined data size:** (6655, 288)

2018-2021 vs 2022:

**Preprocessing notes:**

* Capitalize variable names
* Rename ‘BLEEDDIS’ back to ‘BLEEDIS’
* Reconcile different values with same variable name: 'ETHNICITY\_HISPANIC', ‘OTHBLEED’

**New columns:** ['PREOP\_CREAT\_MSINCR', 'POSTOP\_CREAT\_MSINCR', 'OP\_APPROACH', 'ROBOT\_USED', 'UNPLANNED\_CONV\_OPEN', 'HAND\_OPEN\_ASSIST']

**Removed columns:** ['ELECTSURG', 'DYSPNEA', 'WNDINF', 'WTLOSS', 'DPRPT', 'PRPT', 'EMERGNCY', 'WNDCLAS', 'PODIAG', 'PODIAGTX', 'REOPOR1ICD91', 'REOPOR2ICD91', 'READMRELICD91', 'READMUNRELICD91', 'READMRELICD92', 'READMUNRELICD92', 'READMRELICD93', 'READMUNRELICD93', 'READMRELICD94', 'READMUNRELICD94', 'READMRELICD95', 'READMUNRELICD95', 'WOUND\_CLOSURE', 'PODIAG\_OTHER']

**Same name different values:** ['RACE\_NEW', 'ETHNICITY\_HISPANIC', 'TRANST', 'DISCHDEST', 'ANESTHES', 'PRSEPIS', 'OTHERPROC1', 'OTHERCPT1', 'OTHERPROC2', 'OTHERPROC3', 'OTHERPROC4', 'OTHERPROC5', 'OTHERPROC6', 'OTHERPROC7', 'CONCURR1', 'CONCURR2', 'CONCURR3', 'CONCURR4', 'CONCURR5', 'RENAINSF', 'OPRENAFL', 'OTHBLEED', 'OTHDVT', 'PODIAG10', 'PODIAGTX10', 'RETORRELATED', 'REOPOR1ICD101', 'REOPERATION2', 'RETOR2RELATED', 'REOPOR2ICD101', 'REOPERATION3', 'UNPLANNEDREADMISSION1', 'READMRELATED1', 'READMSUSPREASON1', 'READMRELICD101', 'READMISSION2', 'UNPLANNEDREADMISSION2', 'READMRELATED2', 'READMSUSPREASON2', 'READMRELICD102', 'READMISSION3', 'UNPLANNEDREADMISSION3', 'READMRELATED3', 'READMSUSPREASON3', 'READMRELICD103', 'PODIAG\_OTHER10', 'ANESTHES\_OTHER', 'POSTOP\_COVID', 'IMMUNO\_CAT', 'HOMESUP', 'HXFALL', 'HXDEMENTIA', 'DELIRIUM', 'DISHOMESVC', 'DISFXNSTAT']

**Preselect cols removed:** ['DYSPNEA', 'WNDINF', 'WTLOSS', 'PRPT', 'EMERGNCY']

**Preselect cols with different values:** ['RACE\_NEW', 'ETHNICITY\_HISPANIC', 'ANESTHES', 'OTHBLEED']

**Cols value comparison:**

RACE\_NEW

df1: ['American Indian or Alaska Native', 'Asian', 'Black or African American', 'Native Hawaiian or Other Pacific Islander', 'Native Hawaiian or Pacific Islander', 'Some Other Race', 'Unknown/Not Reported', 'White']

df2: ['American Indian or Alaska Native', 'Asian', 'Black or African American', 'Native Hawaiian or Other Pacific Islander', 'Some Other Race', 'Unknown/Not Reported', 'White']

ETHNICITY\_HISPANIC

df1: ['N', 'U', 'Y']

df2: ['No', 'Unknown', 'Yes']

ANESTHES

df1: ['General', 'MAC/IV Sedation', 'Other']

df2: ['General']

OTHBLEED

df1: ['No Complication', 'Transfusions/Intraop/Postop']

df2: ['Blood Transfusion', 'No Complication']

**Combined dataset size:** (8587, 294)

Baseline data preprocessing notes:

* Recode -99 as Nan
* Remove features with over 50% missing: N of features 294 🡪129
* Fix ‘DOTHBLEED’: replace nan with -1
* Fix ‘AGE’: convert to number after fixing 90+
* Calculate ‘BMI’ using ‘HEIGHT’ and ‘WEIGHT’
* Drop ‘CASEID’
* Drop ‘DOTHBLEED’ (r=-.99) & ‘NOTHBLEED’(r=-.85)
* Encode all variables into numeric values

Preselect 43 feature notes:

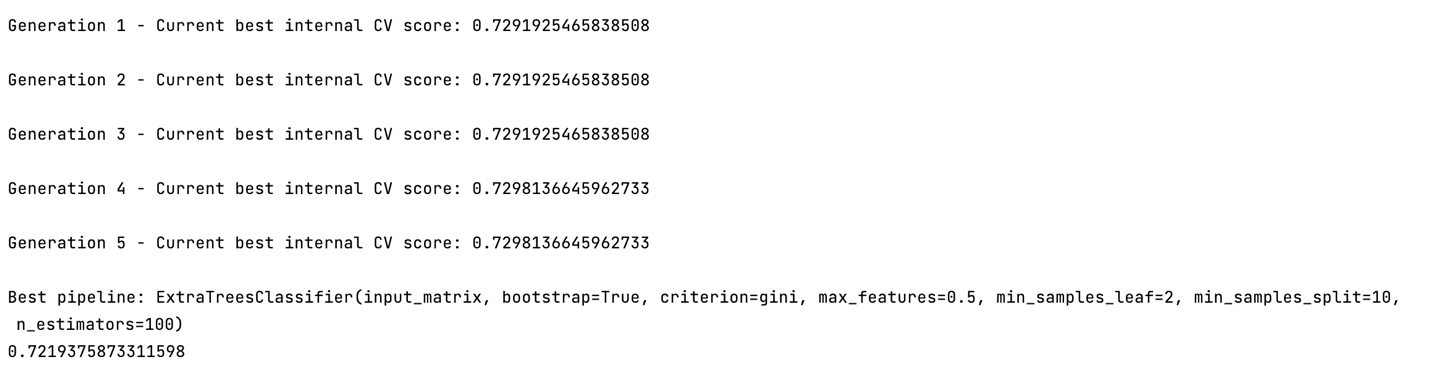
* ‘PRPT’ is removed from the preselect list: all nans
* Drop highly correlated variables: 'HEIGHT','WEIGHT','ETHNICITY\_HISPANIC'
* Final feature size: 40 features + target

Preselect 20 features notes:

* Subset of 40-feature dataset

Tpot - 40 features:

Accuracy:

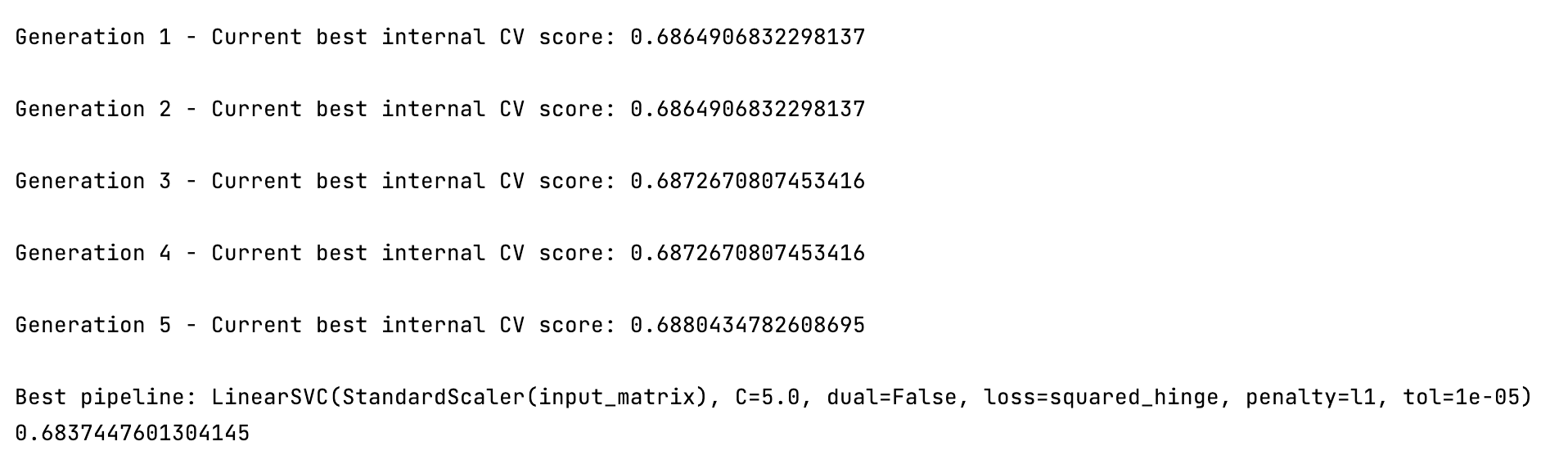


F1 score:

A screenshot of a computer code

Description automatically generated

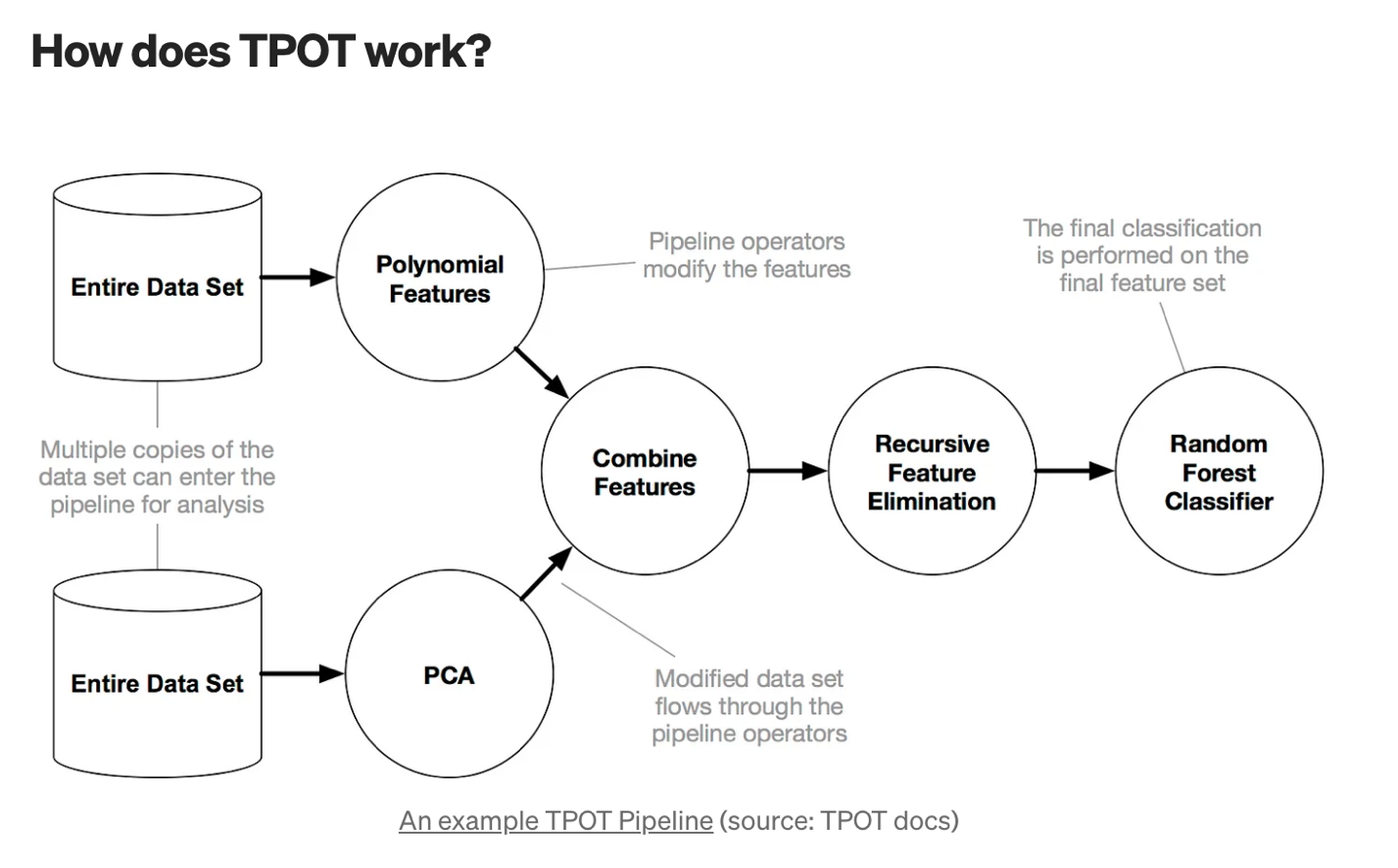
Tpot – autofeat top 10 features:



Tpot – autofeat top 5 features:

A number and numbers on a white background

Description automatically generated



A screenshot of a computer code

Description automatically generated

**March 6, 2024 UPDATES**

Data Preprocessing

* Solved the mystery of perfect modeling score with baseline data:
  + Looked into [correlation with target](../../code/Jenny/processed_data/2018_2022/CABG_5yr_baseline_review.xlsm)
  + Found derivatives of target: DOTHBLEED (r=-.85), NOTHBLEED (r = -.99)
  + After removing these two variables:
    - Random Forest: Accuracy = .745 , F1 = .727
* 2nd round of variable review by Dr. Gupta:
  + Confirmed that the 80+ variables are post-op or irrelevant except ‘OTHERCPT1’
    - Do we still need to create the ‘baseline dataset’? Or we should just use the 40+ feature dataset as baseline?
  + Can add ‘OTHERCPT1’ 🡪 to-dos for next week
  + Sent in more datasets: 2015, 2016, 2017 🡪 to-dos for next week

Autofeat datasets

* Using top 20/10/5 important features from Random Forest to generate new datasets
* Results (Random Forest):

|  |  |  |
| --- | --- | --- |
| **Autofeat Data** | **Accuracy** | **F1-score** |
| Top 20 features | .732 | .706 |
| Top 10 features | .706 | .686 |
| Top 5 | .708 | .689 |

Synthetic data generation

* [Bayesian Network](https://colab.research.google.com/drive/1Nnl6xSXORK-UIcPaTH9fQ2LZz-Flkc-R?authuser=3#scrollTo=PRdJ-hlhUmXV) (DataSynthesizer) - good
* [GANs](https://colab.research.google.com/drive/1Nvfm2PZFevi0gOmNn55wg9J0Gp5CuJJD?authuser=3#scrollTo=urxbgNVRSwRD) – codes were working but produced weird data 🡪 which activation function should we use in generator? Softmax?
* [Realtabformer](https://colab.research.google.com/drive/1SHUWKw6YzxlviW3_zf_o9Rro4aAmBe1z?authuser=3#scrollTo=16xNsgtFPFhl) – Jenny had runtime errors

Questions for Prof. Amir:

* Why do we need synthetic data for this project? (for NN training?)
* What are the expectations? (prefer using advanced techniques (e.g., NN) to generate synthetic data? How much more?)
* Can we use 40-feature dataset as baseline data to generate synthetic data?