Data processing:

1. Missing values: only independent variable ‘parentspecies’ contains missing value, maybe fillna by its mode

2. For ambiguous ‘parentspecies’, do we drop them?

3. abnormal data: NumofConf’s max value seems surprisingly high.

4. Check multicollinearity: how to deal with the categorial independent variable? I used one-hot encoding, is there better approach?

strong correlation between NumOfO and MW, NumOfC and NumOfAtoms,whose correlation equal to 0.88 and 0.84 respectively.

maybe consider drop one of them if it improves model performance.

Non-linear

Regression

Decane\_apin

Regression models:

Lasso

Random forest

Number of variables very large values.

Check r2 score

Inputing

Model training:

**Model1: Ridge-Regularized Polynomial Regression Model (degree 1)**

* Consider Drop features due to multicolinearlity:

|  |  |  |
| --- | --- | --- |
| Feature dropped | Train r2 | Test r2 |
| Original model | 0.7024 | 0.7158 |
| Drop NumOfO & NumOfAtoms | 0.7019 | 0.7153 |
| Drop NumOfC | 0.7053 | 0.7145 |
| Drop MW | 0.7024 | 0.7158 |

Consider drop MW to simplify the model

* Apply Forward selection:

|  |  |  |
| --- | --- | --- |
| Selected features:  ['NumHBondDonors', 'NumOfC', 'hydroxyl (alkyl)', 'carbonylperoxynitrate',  'carboxylic acid', 'NumOfConf', 'parentspecies\_toluene', 'C=C (non-aromatic)',  'peroxide', 'ether (alicyclic)', 'ketone', 'parentspecies\_decane',  'aldehyde', 'NumOfO', 'NumOfAtoms'] | 0.7023750736260639 | 0.7165 |
|  |  |  |

* Further apply Ridge regularization

|  |  |  |
| --- | --- | --- |
|  | 0.7028869771573394 | 0.7165 |
|  |  |  |

**Model 2: linearGAM**

|  |  |  |
| --- | --- | --- |
| Generalized gam | 0.7139419423429578 | 0.73 |
| GAM with Piecewise ReLU | 0.7034966131068829 | 0.7173 |
| GAM with b-splines=4 | 0.7173 | 0.7302 |
| GAM with b-Splines=5 | 0.7175 | 0.7303 |
| 6 | 0.7176 | 0.7303 |
| 7 | 0.7176 | 0.7303 |
| 8 | 0.7177 | 0.7303 |
| 9 | 0.7180 | 0.7305 |

Based on the Testing R², select n\_splines = 9, as it provides the best generalization performance.