|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Feature score threshold (%) | Number of features | Train MAE | Test MAE | Train MSE | Test MSE | Train R^2 | Test R^2 |
| 0.00 | 18 | 0.5424 | 0.7468 | 0.5683 | 1.4984 | 0.4834 | 0.4557 |
| 0.05 | 13 | 0.5432 | 0.7512 | 0.5705 | 1.5180 | 0.4813 | 0.4485 |
| 0.12 | 12 | 0.5427 | 0.7498 | 0.5711 | 1.5123 | 0.4808 | 0.4507 |
| 0.80 | 11 | 0.5430 | 0.7507 | 0.5714 | 1.5143 | 0.4805 | 0.4499 |
| 1.61 | 10 | 0.5441 | 0.7547 | 0.5725 | 1.5231 | 0.4795 | 0.4467 |
| 1.66 | 9 | 0.5441 | 0.7547 | 0.5726 | 1.5234 | 0.4795 | 0.4466 |
| 1.86 | 8 | 0.5440 | 0.7570 | 0.5733 | 1.5341 | 0.4788 | 0.4427 |
| 2.90 | 7 | 0.5450 | 0.7575 | 0.5768 | 1.5361 | 0.4756 | 0.4420 |
| 3.00 | 6 | 0.5452 | 0.7573 | 0.5774 | 1.5351 | 0.4750 | 0.4423 |
| 4.50 | 5 | 0.5491 | 0.7678 | 0.5848 | 1.5706 | 0.4683 | 0.4295 |

Polynomial degree = 1

Polynomial degree = 2

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Feature score threshold (%) | Number of features | Train MAE | Test MAE | Train MSE | Test MSE | Train R^2 | Test R^2 |
| 0.00 | 18 | 0.3385 | 0.5246 | 0.2207 | 0.9246 | 0.7994 | 0.6641 |
| 0.05 | 13 | 0.3897 | 0.5702 | 0.2802 | 0.9550 | 0.7452 | 0.6531 |
| 0.12 | 12 | 0.4377 | 0.6348 | 0.3520 | 1.1173 | 0.6800 | 0.5941 |
| 0.80 | 11 | 0.4408 | 0.6349 | 0.3566 | 1.1140 | 0.6758 | 0.5953 |
| 1.61 | 10 | 0.4510 | 0.6523 | 0.3727 | 1.1981 | 0.6612 | 0.5648 |
| 1.66 | 9 | 0.4526 | 0.6569 | 0.3759 | 1.2151 | 0.6583 | 0.5586 |
| 1.86 | 8 | 0.4613 | 0.6652 | 0.3880 | 1.2341 | 0.6472 | 0.5517 |
| 2.90 | 7 | 0.4644 | 0.6694 | 0.3952 | 1.2495 | 0.6407 | 0.5461 |
| 3.00 | 6 | 0.4656 | 0.6687 | 0.3981 | 1.2458 | 0.6381 | 0.5475 |
| 4.50 | 5 | 0.4737 | 0.6918 | 0.4125 | 1.3386 | 0.6250 | 0.5138 |

Polynomial degree = 3

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Feature score threshold (%) | Number of features | Train MAE | Test MAE | Train MSE | Test MSE | Train R^2 | Test R^2 |
| 0.00 | 18 | 0.1818 | 0.3030 | 0.0718 | 0.3757 | 0.9348 | 0.8635 |
| 0.05 | 13 | 0.2289 | 0.3534 | 0.1077 | 0.4516 | 0.9021 | 0.8360 |
| 0.12 | 12 | 0.3290 | 0.4903 | 0.2217 | 0.6719 | 0.7984 | 0.7559 |
| 0.80 | 11 | 0.3334 | 0.4932 | 0.2260 | 0.6725 | 0.7946 | 0.7557 |
| 1.61 | 10 | 0.3577 | 0.5329 | 0.2585 | 0.8120 | 0.7650 | 0.7050 |
| 1.66 | 9 | 0.3558 | 0.5306 | 0.2574 | 0.8050 | 0.7660 | 0.7076 |
| 1.86 | 8 | 0.3583 | 0.5345 | 0.2615 | 0.8176 | 0.7623 | 0.7030 |
| 2.90 | 7 | 0.3654 | 0.5436 | 0.2737 | 0.8481 | 0.7511 | 0.6919 |
| 3.00 | 6 | 0.3697 | 0.5471 | 0.2829 | 0.8683 | 0.7428 | 0.6846 |
| 4.50 | 5 | 0.3906 | 0.5949 | 0.3166 | 1.0267 | 0.7121 | 0.6270 |

Polynomial degree = 4

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Feature score threshold (%) | Number of features | Train MAE | Test MAE | Train MSE | Test MSE | Train R^2 | Test R^2 |
| 0.00 | 18 | 0.1122 | 0.2331 | 0.0304 | 0.2805 | 0.9724 | 0.8981 |
| 0.05 | 13 | 0.1509 | 0.2558 | 0.0506 | 0.2763 | 0.9540 | 0.8996 |
| 0.12 | 12 | 0.2870 | 0.4468 | 0.1807 | 0.5614 | 0.8357 | 0.7961 |
| 0.80 | 11 | 0.2878 | 0.4450 | 0.1818 | 0.5672 | 0.8347 | 0.7939 |
| 1.61 | 10 | 0.3167 | 0.4992 | 0.2161 | 0.7672 | 0.8035 | 0.7213 |
| 1.66 | 9 | 0.3163 | 0.4990 | 0.2170 | 0.7703 | 0.8027 | 0.7202 |
| 1.86 | 8 | 0.3202 | 0.5003 | 0.2219 | 0.7636 | 0.7983 | 0.7226 |
| 2.90 | 7 | 0.3311 | 0.5156 | 0.2370 | 0.8114 | 0.7845 | 0.7053 |
| 3.00 | 6 | 0.3356 | 0.5173 | 0.2425 | 0.8087 | 0.7795 | 0.7062 |
| 4.50 | 5 | 0.3568 | 0.5709 | 0.2771 | 1.0055 | 0.7481 | 0.6347 |

**Pipeline**

1. Creating new compound features in the training dataset.
2. Eliminating the observations that present outliers of the normalized target (y\_norm).
3. Scoring features with Lasso regression (grid search for alpha parameter).
4. Selecting features according to a relative score threshold (percentage of total score).
5. Creating polynomial features, with polynomial degree n.
6. Standardizing the features’ vector.
7. Fitting a linear regression with Scikit-Learn’s function LinearRegression().
8. Applying the same steps to the features from the testing set, except for the elimination of outliers.
9. Predicting the test target feature (y\_norm).
10. Determining the mean absolute error (MAE) and coefficient of determination (R^2) for the training and testing sets.
11. Repeating the process for other values of the polynomial degree n and relative feature score threshold.

**Performance visualization**

**Зображення, що містить ряд, Барвистість, знімок екрана

Автоматично згенерований опис**

**Зображення, що містить ряд, Графік, Барвистість, знімок екрана

Автоматично згенерований опис**

**Зображення, що містить ряд, Графік, знімок екрана

Автоматично згенерований опис**

**Зображення, що містить знімок екрана, Барвистість, ряд, Графік

Автоматично згенерований опис**

**Polynomial degrees greater than or equal to 5:**

The regression type is no longer linear, since there would be a lot overfitting. We have chosen to use a Lasso regression without feature selection beforehand, since it has been proven that linear regression has best performance when considering all features.

Polynomial degree = 5

* 1/20th of the whole samples due to memory issues

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Alpha | Train MAE | Test MAE | Train MSE | Test MSE | Train R^2 | Test R^2 |
| 0.000001 | 0.0401 | 0.3242 | 0.0038 | 0.3634 | 0.9968 | 0.8753 |
| 0.000010 | 0.0434 | 0.3029 | 0.0044 | 0.3283 | 0.9962 | 0.8874 |
| 0.000100 | 0.0709 | 0.2493 | 0.0120 | 0.2546 | 0.9898 | 0.9127 |
| 0.000500 | 0.1144 | 0.2653 | 0.0319 | 0.2735 | 0.9731 | 0.9062 |
| 0.001000 | 0.1437 | 0.2859 | 0.0483 | 0.3272 | 0.9592 | 0.8877 |
| 0.005000 | 0.2277 | 0.3599 | 0.1107 | 0.4401 | 0.9066 | 0.8490 |
| 0.010000 | 0.2727 | 0.4032 | 0.1500 | 0.5308 | 0.8735 | 0.8179 |
| 0.050000 | 0.4452 | 0.6258 | 0.3600 | 1.1719 | 0.6963 | 0.5979 |
| 0.100000 | 0.5237 | 0.7285 | 0.4857 | 1.5038 | 0.5901 | 0.4841 |

Polynomial degree = 6

* 1/30th of the whole samples due to memory issues

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Alpha | Train MAE | Test MAE | Train MSE | Test MSE | Train R^2 | Test R^2 |
| 0.000001 | 0.0087 | 0.5716 | 0.0004 | 1.2236 | 0.9997 | 0.6400 |
| 0.000010 | 0.0153 | 0.4755 | 0.0008 | 0.8757 | 0.9993 | 0.7424 |
| 0.000100 | 0.0434 | 0.3396 | 0.0050 | 0.4904 | 0.9955 | 0.8557 |
| 0.000500 | 0.0866 | 0.3173 | 0.0181 | 0.4190 | 0.9839 | 0.8767 |
| 0.001000 | 0.1119 | 0.3427 | 0.0294 | 0.4807 | 0.9739 | 0.8585 |
| 0.005000 | 0.2015 | 0.4229 | 0.0857 | 0.7510 | 0.9237 | 0.7790 |
| 0.010000 | 0.2533 | 0.4738 | 0.1285 | 0.9118 | 0.8857 | 0.7317 |
| 0.050000 | 0.4217 | 0.6936 | 0.3207 | 1.5557 | 0.7152 | 0.5423 |
| 0.100000 | 0.4999 | 0.7901 | 0.4418 | 1.8815 | 0.6077 | 0.4464 |

We have selected this model (green) because it is the one that seems to present a better balance between training and testing performance, being the least overfitted model.

Зображення, що містить ряд, знімок екрана, Графік, Барвистість

Автоматично згенерований опис

Зображення, що містить ряд, Графік, знімок екрана, Паралель

Автоматично згенерований опис