Math 656: Turbine Analysis

Luke Botti

Data Prep

The first step towards preparing the data was getting a general idea of the difference in the features. Generating summary statistics for the dataset shows wild differences in center and spread for the data.

	AT	AP	AH	AFDP	GTEP	TIT	TAT	TEY	CDP	CO	NOX
mean	17.225259	1014.50911	68.647464	3.598909	26.130149	1078.974689	546.642484	133.993380	12.097025	3.129986	59.890509
std	8.095783	6.89543	13.541116	0.610226	4.473737	19.762449	5.489066	16.179208	1.136601	2.234962	11.132464
min	-6.234800	989.40000	24.085000	2.368800	17.698000	1016.000000	516.040000	100.020000	9.870800	0.212800	25.905000
25%	11.073250	1009.67500	59.447250	3.117300	23.147000	1070.500000	544.747500	126.255000	11.465750	1.808175	52.399000
50%	17.456500	1014.00000	70.952000	3.538500	25.331000	1080.300000	549.720000	131.600000	11.933000	2.533400	56.838500
75%	23.684750	1018.30000	79.653750	4.194825	30.018250	1099.900000	550.030000	147.160000	13.148000	3.702550	65.093250
max	37.103000	1036.60000	96.666000	5.239500	40.716000	1100.400000	550.590000	179.500000	15.159000	41.097000	119.680000

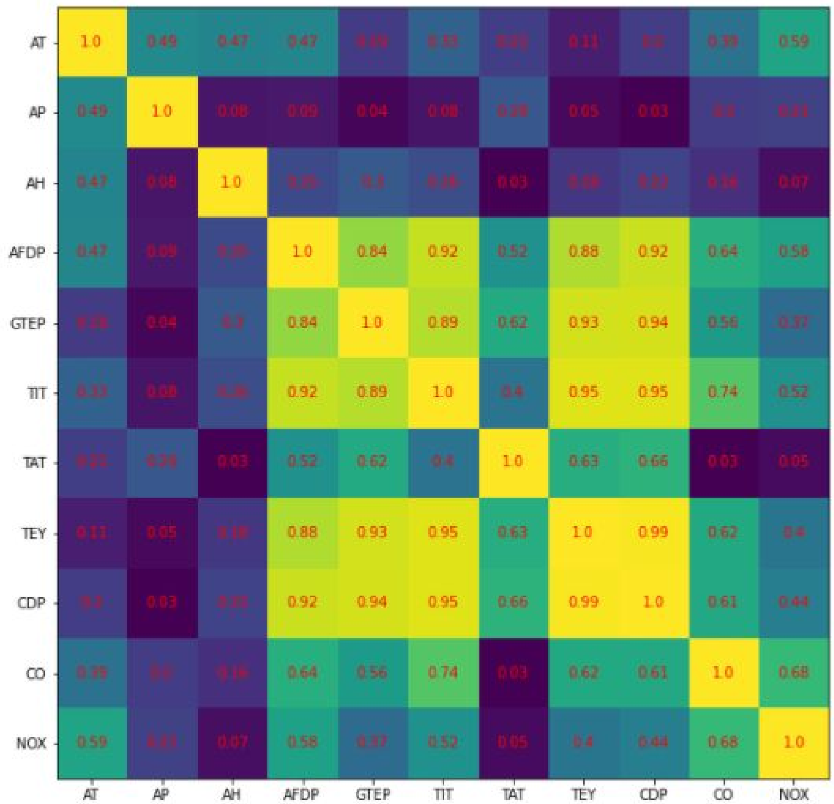
For example the column TIT (Turbine inlet temperature) has a mean of 1078 and standard deviation of 19.7, while the column CO (Carbon Monoxide) has a mean of 3.1 and standard deviation of 2.2. Comparing the 75th percentile for CO to the max also reveals the presence of outliers in the data, so I needed to be mindful of those when normalizing the data.

With these things in mind I used Sklearn's RobustScaler to normalize the data. This performs linear scaling and centering. However, it is mostly robust to outliers because it uses the interquartile range instead of the standard range. The resulting summary statistics are much more similar while still maintaining the presence of outliers.

	AT	AP	AH	AFDP	GTEP	TIT	TAT	TEY	CDP	CO	NOX
mean	-1.833570e-02	0.059027	-1.140492e-01	0.056063	1.163032e-01	-0.045079	-0.582587	0.114488	0.097503	0.314925	2.404246e-01
std	6.419366e-01	0.799470	6.701367e-01	0.566322	6.510805e-01	0.672192	1.039104	0.773940	0.675643	1.179789	8.769690e-01
min	-1.878547e+00	-2.852174	-2.319402e+00	-1.085543	-1.110860e+00	-2.187075	-6.375769	-1.510643	-1.225858	-1.224995	-2.436812e+00
25%	-5.061452e-01	-0.501449	-5.693589e-01	-0.390896	-3.178461e-01	-0.333333	-0.941316	-0.255680	-0.277753	-0.382831	-3.497253e-01
50%	1.408514e-16	0.000000	3.516406e-16	0.000000	2.585145e-16	0.000000	0.000000	0.000000	0.000000	0.000000	2.798665e-16
75%	4.938548e-01	0.498551	4.306411e-01	0.609104	6.821539e-01	0.666667	0.058684	0.744320	0.722247	0.617169	6.502747e-01
max	1.557824e+00	2.620290	1.272561e+00	1.578618	2.239039e+00	0.683673	0.164695	2.291318	1.917670	20.356899	4.950391e+00

Feature Relationships and Selection

The first step I took in finding relationships between the variables is examining the correlation. Below is a heatmap of the absolute correlations.



The heatmap clearly shows a group of highly correlated variables. I investigated them further by attempting to build linear regression models. The three variables that can be predicted with a high degree of reliability (Adjusted R-squared >=0.95) off of the rest are: Compressor discharge pressure (CDP), Turbine inlet temperature (TIT), and Turbine energy yield (TEY).

Compressor Discharge Pressure

```
                            OLS Regression Results
=======================================================================================
Dep. Variable:                    CDP   R-squared (uncentered):                   0.969
Model:                            OLS   Adj. R-squared (uncentered):              0.969
Method:                 Least Squares   F-statistic:                          2.849e+04
Date:                Sun, 13 Dec 2020   Prob (F-statistic):                        0.00
Time:                        01:32:38   Log-Likelihood:                          5126.1
No. Observations:                7384   AIC:                                 -1.024e+04
Df Residuals:                    7376   BIC:                                 -1.018e+04
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
AT            -0.1916      0.005    -39.294      0.000      -0.201      -0.182
AP             0.0098      0.002      4.329      0.000       0.005       0.014
AH            -0.0373      0.003    -12.608      0.000      -0.043      -0.032
AFDP           0.7051      0.007     97.637      0.000       0.691       0.719
GTEP           0.3702      0.005     68.442      0.000       0.360       0.381
TAT           -0.0514      0.002    -23.410      0.000      -0.056      -0.047
CO            -0.0565      0.002    -25.683      0.000      -0.061      -0.052
NOX           -0.0043      0.003     -1.567      0.117      -0.010       0.001
==============================================================================
Omnibus:                      654.769   Durbin-Watson:                   0.317
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1830.781
Skew:                          -0.490   Prob(JB):                         0.00
Kurtosis:                       5.234   Cond. No.                         10.1
==============================================================================
```

The above regression on the normalized data shows a statistically significant relationship between the remaining features and CDP, with the exception of Nitrogen oxides (NOX). Most notably ambient temperature (AT), air filter difference pressure (AFDP), and gas turbine exhaust pressure (GTEP) have the most influence on CDP. An increase in AT decreases CDP, while increases in AFDP and GTEP result in increases in CDP.

Turbine Inlet Temperature

```
                              OLS Regression Results
=========================================================================================
Dep. Variable:                    TIT   R-squared (uncentered):                   0.955
Model:                            OLS   Adj. R-squared (uncentered):              0.955
Method:                 Least Squares   F-statistic:                          1.951e+04
Date:                Sun, 13 Dec 2020   Prob (F-statistic):                        0.00
Time:                        01:38:58   Log-Likelihood:                          3878.8
No. Observations:                7384   AIC:                                     -7742.
Df Residuals:                    7376   BIC:                                     -7686.
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
AT            -0.2385      0.006    -41.307      0.000      -0.250      -0.227
AP             0.0123      0.003      4.577      0.000       0.007       0.017
AH            -0.0468      0.004    -13.342      0.000      -0.054      -0.040
AFDP           0.8550      0.009     99.999      0.000       0.838       0.872
GTEP           0.4492      0.006     70.139      0.000       0.437       0.462
TAT            0.2040      0.003     78.413      0.000       0.199       0.209
CO            -0.0694      0.003    -26.637      0.000      -0.075      -0.064
NOX           -0.0110      0.003     -3.379      0.001      -0.017      -0.005
==============================================================================
Omnibus:                      707.060   Durbin-Watson:                   0.271
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2435.747
Skew:                          -0.464   Prob(JB):                         0.00
Kurtosis:                       5.656   Cond. No.                         10.1
==============================================================================
```

The above regression once again shows a statistically significant relationship between all the remaining variables and TIT. Once again, AT, AFDP, GTEP have significant coefficients, but this time turbine after temperature (TAT) significantly influences the model as well. Like before an increase in ambient temperature decreases the inlet temperature. The other three variables are all correlated to an increase in inlet temperature.

Turbine Energy Yield

```
                            OLS Regression Results
=======================================================================================
Dep. Variable:                    TEY   R-squared (uncentered):                   0.960
Model:                            OLS   Adj. R-squared (uncentered):              0.960
Method:                 Least Squares   F-statistic:                          2.228e+04
Date:                Sun, 13 Dec 2020   Prob (F-statistic):                        0.00
Time:                        01:45:33   Log-Likelihood:                          3243.8
No. Observations:                7384   AIC:                                     -6472.
Df Residuals:                    7376   BIC:                                     -6416.
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
AT            -0.4014      0.006    -63.779      0.000      -0.414      -0.389
AP            -0.0119      0.003     -4.066      0.000      -0.018      -0.006
AH            -0.0634      0.004    -16.599      0.000      -0.071      -0.056
AFDP           0.8395      0.009     90.094      0.000       0.821       0.858
GTEP           0.4659      0.007     66.749      0.000       0.452       0.480
TAT           -0.0104      0.003     -3.663      0.000      -0.016      -0.005
CO            -0.0723      0.003    -25.466      0.000      -0.078      -0.067
NOX           -0.0088      0.004     -2.470      0.014      -0.016      -0.002
==============================================================================
Omnibus:                      539.041   Durbin-Watson:                   0.311
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1584.304
Skew:                          -0.382   Prob(JB):                         0.00
Kurtosis:                       5.137   Cond. No.                         10.1
==============================================================================
```

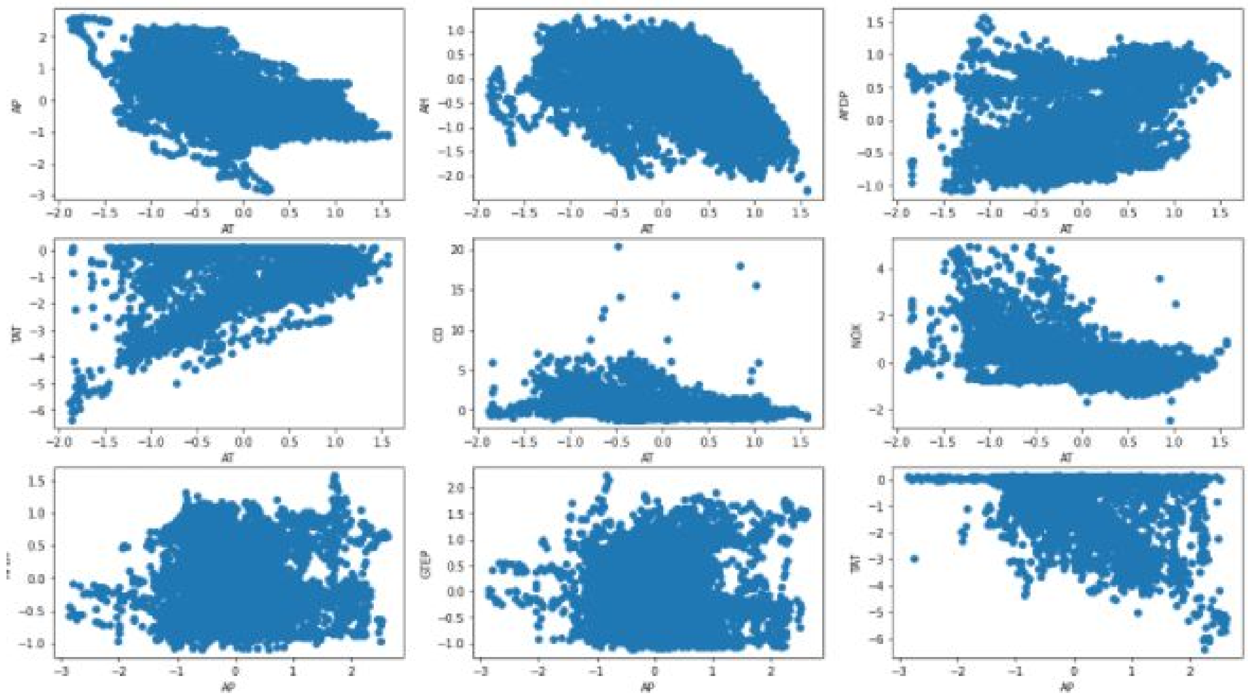
For the third time, the above regression shows a statistically significant relationship between all the remaining variables and turbine energy yield. Like with compressor discharge pressure, the three most correlated variables are AT, AFDP, and GTEP. Additionally they have the same directional relationships. An increase in AT decreases energy yield, and increases in AFDP and GTEP increase energy yield.

Given the similarities between how the most significant predictor features affect our target features, we can additionally conclude that our target features are all positively correlated together. This is supported by the correlation matrix (not included). We conclude the feature selection by dropping these three features from the dataset before doing any clustering; as all the information contained in them is almost perfectly captured by the rest of the data.

Clustering

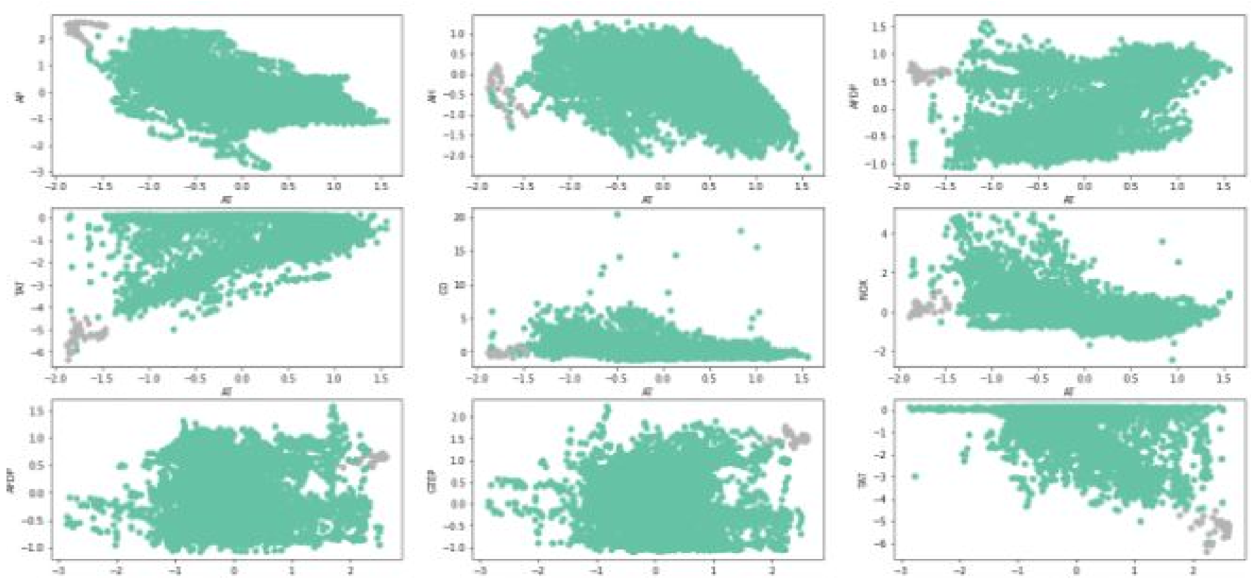
Visually examining the different combinations of three and two dimensional cross sections of the data do not lead to very much hope for finding many interesting clusters.

This is highlighted in a sample of 2D cross sections below:



Generally we can see a large main body and a potential small outcropping of points in most of the cross sections. If the algorithm identifies those outcroppings as the same cluster, then it's likely significant.

Running DBSCAN with a maximum distance between neighboring points of 0.85 gives 11 clusters, but visually inspecting the results identifies two main clusters. The larger cluster (7348 points) is in green and the smaller one (36 points) is in grey.



In most of the cross sections the outcroppings are all part of the same cluster.

Cluster Analysis

To determine what separates the cluster from the rest of the data we train a decision tree with the clusters as the class. Fitting the decision tree with balanced class weights, and a max depth of two produces the following rules for identifying the small cluster: (99.9% accuracy, no cross validation)

1. Turbine after temperature <= -4.48 (526 degrees celsius)
2. Ambient temperature <= -1.33 (0.683 degrees celsius)

Essentially there is a small but significant cluster with low ambient and turbine after temperatures. Based on the regression models above this cluster will have low compressor discharge pressure and energy yields as well.

Bibliography

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.