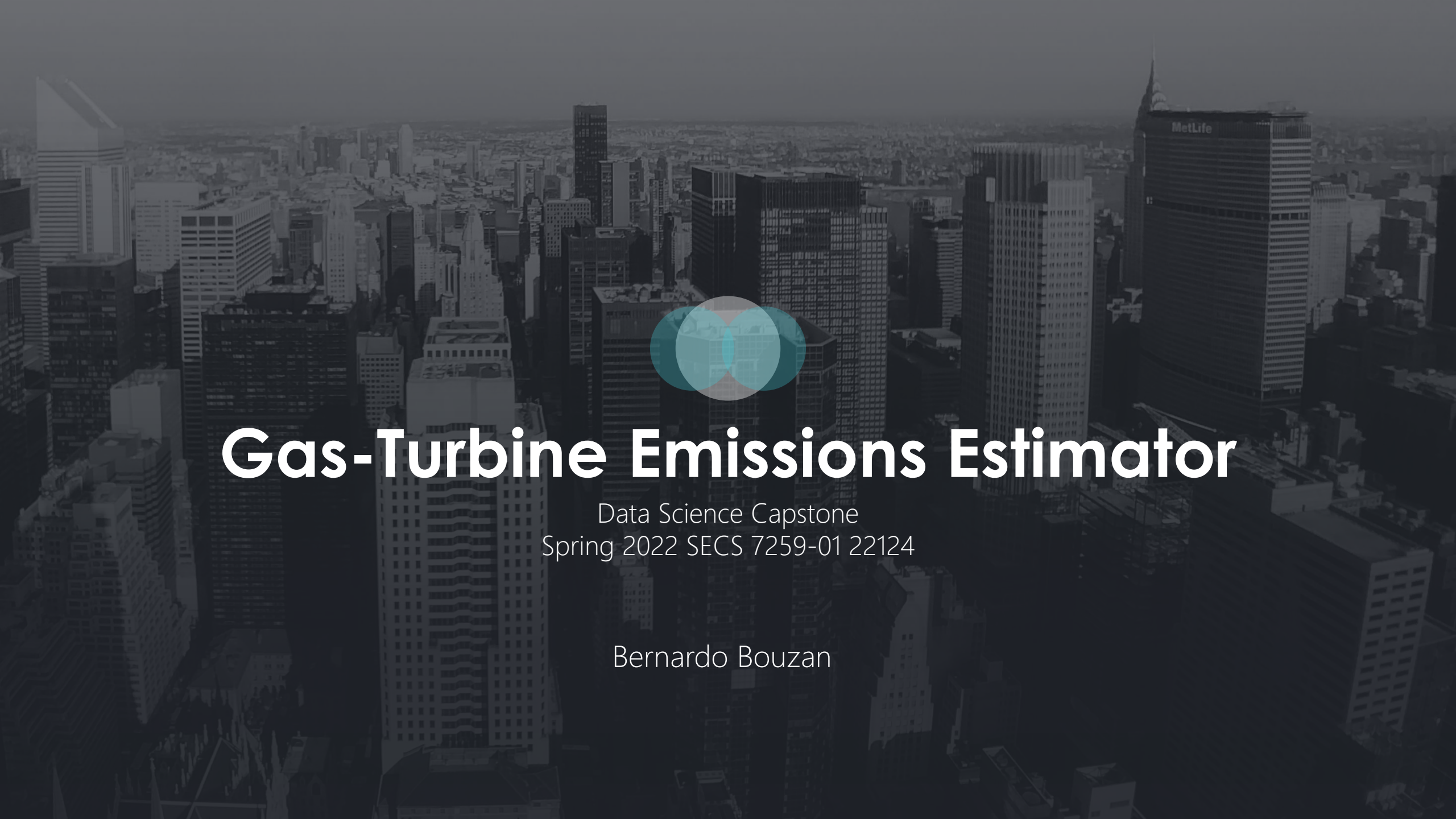


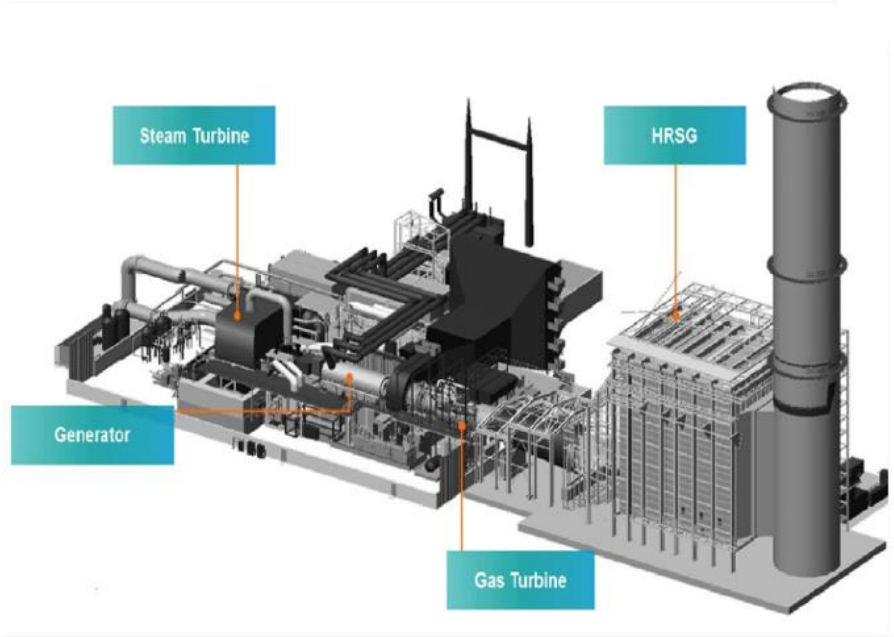
Gas-Turbine Emissions Estimator

Data Science Capstone
Spring 2022 SECS 7259-01 22124

Bernardo Bouzan

GAS-TURBINES OVERVIEW

- Key technology for global emission reduction
- Add-on to renewable energy sources
- Accepts a variety of fuel types
- Emission of CO and NOx



IEEE Spectrum FOR THE TECHNOLOGY INSIDER

Type to search

OPINION | ENERGY

Gas Turbines Have Become by Far the Best Choice for Add-on Generating Power › They offer instant-on power that's compact, mobile, quiet, economical, durable, and matchlessly efficient

BY VACLAV SMIL | 21 NOV 2019 | 2 MIN READ

DATASET

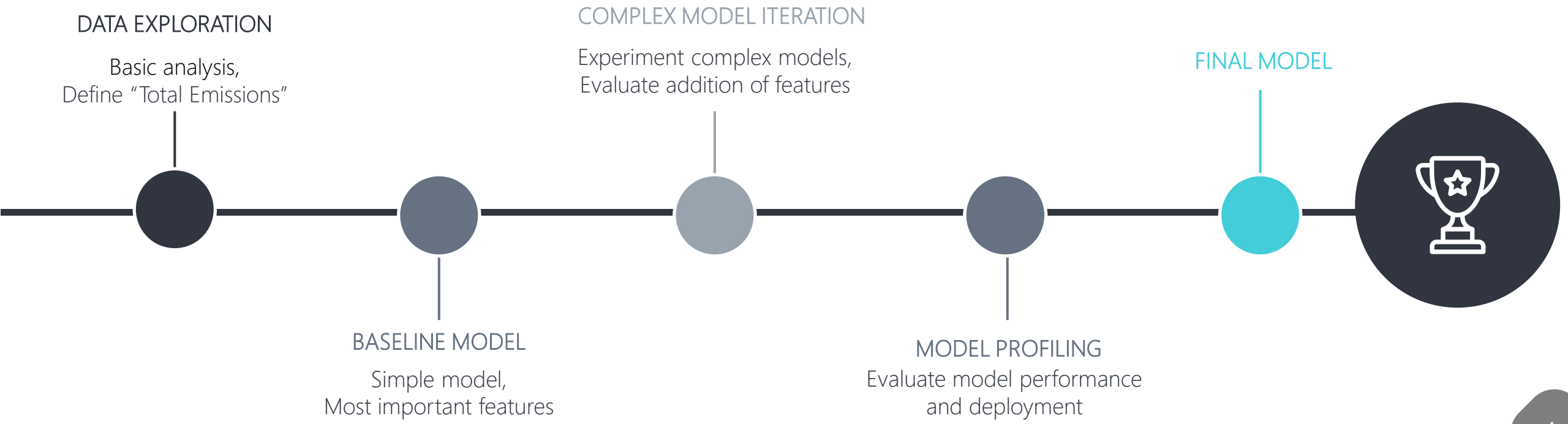
- Data collected from a power plant gas-turbine operating in Turkey
  - Gas-Turbine CO and NOx Emission Data | Kaggle
- Values are hourly averages collected from 2011 to 2015
- 9 Feature variables
  - 3 Ambient data
  - 6 Gas-turbine data
- 2 Emission variables
  - CO
  - NOx
- 36700 samples

Variables	Unit
Ambient temperature (AT)	ºC
Ambient pressure (AP)	mbar
Ambient humidity (AH)	%
Air filter difference pressure (AFDP)	mbar
Gas turbine exhaust pressure (GTEP)	mbar
Turbine inlet temperature (TIT)	ºC
Turbine after temperature (TAT)	ºC
Compressor discharge pressure (CDP)	mbar
Turbine energy yield (TEY)	MWh
Carbon monoxide (CO)	mg/m3
Nitrogen oxides (NOx)	mg/m3

PROJECT PLAN

- Main goal
  - Model to Estimate Total Gas-Turbine Emissions

- Plan



DATA EXPLORATION

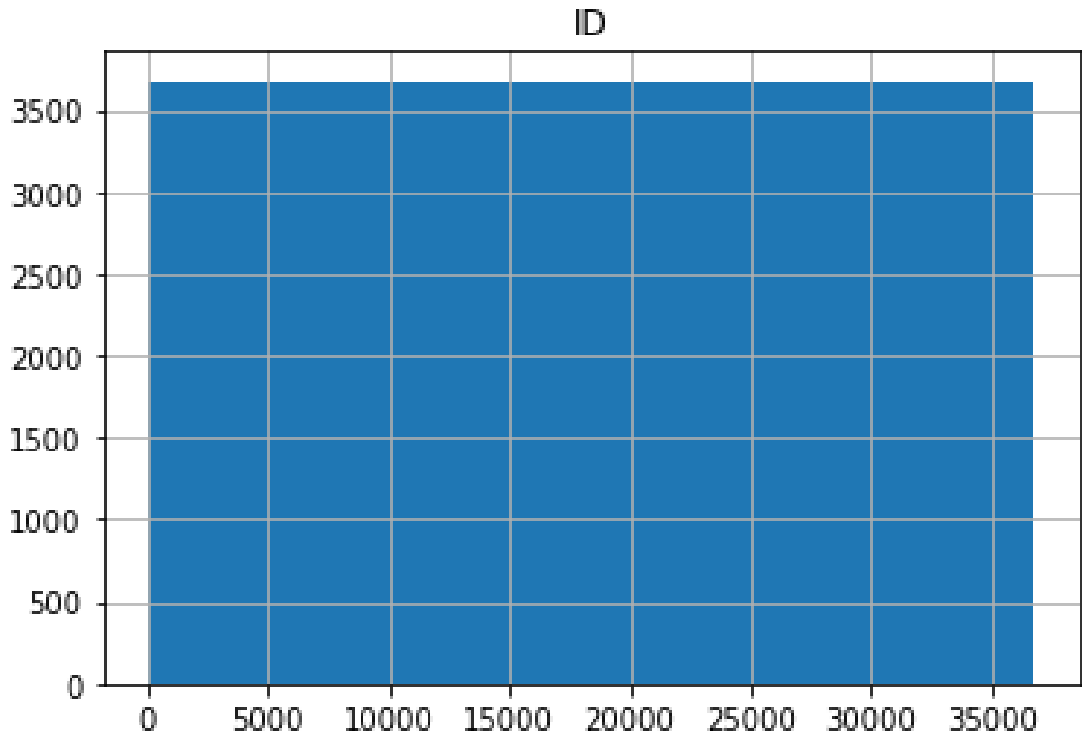
- Descriptive Statistics

	AT	AP	AH	AFDP	GTEP	TIT	TAT	TEY	CDP	CO	NOX
count	36733	36733	36733	36733	36733	36733	36733	36733	36733	36733	36733
mean	17.7127	1013.07	77.867	3.92552	25.5638	1081.43	546.159	133.506	12.0605	2.37247	65.2931
std	7.44745	6.46335	14.4614	0.77394	4.19596	17.5364	6.84236	15.6186	1.0888	2.26267	11.6784
min	-6.2348	985.85	24.085	2.0874	17.698	1000.8	511.04	100.02	9.8518	0.00039	25.905
25%	11.781	1008.8	68.188	3.3556	23.129	1071.8	544.72	124.45	11.435	1.1824	57.162
50%	17.801	1012.6	80.47	3.9377	25.104	1085.9	549.88	133.73	11.965	1.7135	63.849
75%	23.665	1017	89.376	4.3769	29.061	1097	550.04	144.08	12.855	2.8429	71.548
max	37.103	1036.6	100.2	7.6106	40.716	1100.9	550.61	179.5	15.159	44.103	119.91

DATA EXPLORATION

- Index of data point (ID)

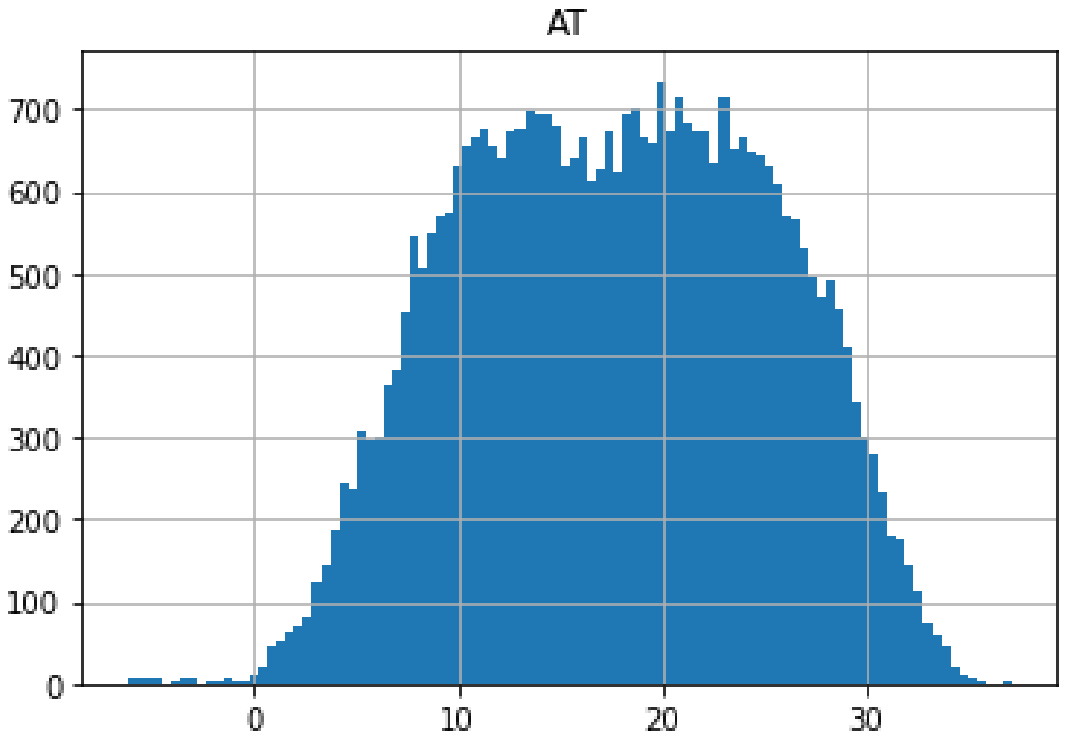
- Just a counter of the sample/row
- Even distribution is expected



DATA EXPLORATION

- Ambient temperature (AT)

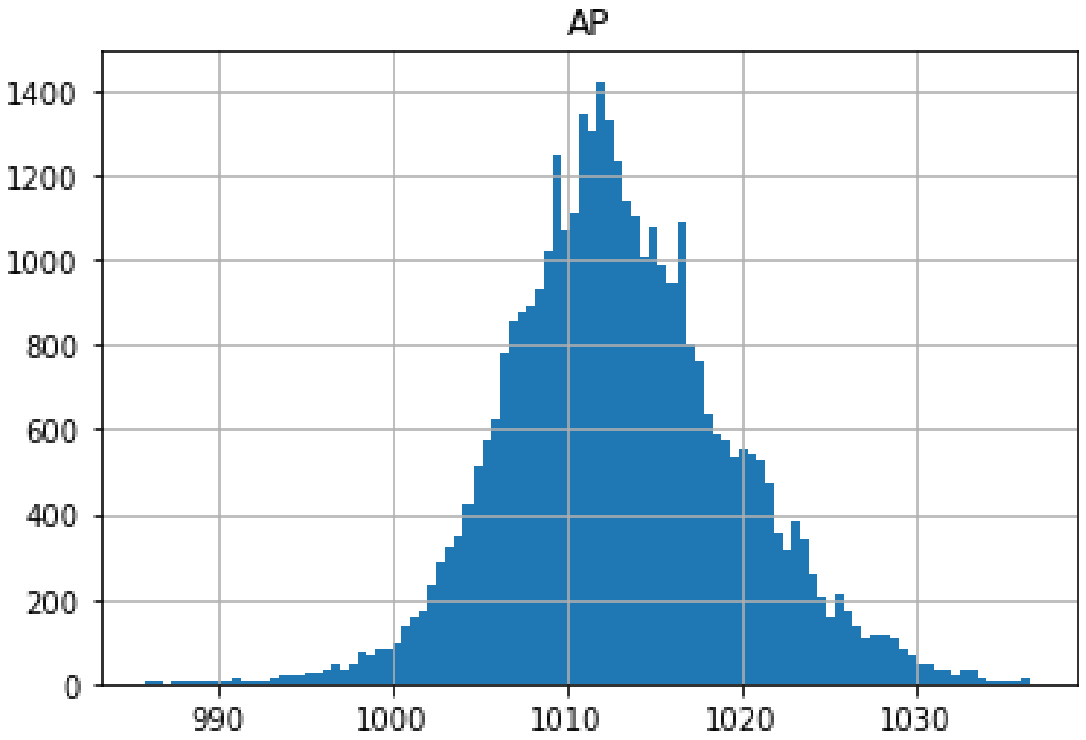
- Weather temperature seems to be fair, mostly around 10 and 25 degree Celsius
- Data seems mostly evenly distributed
- Few datapoints on negative temperature range look like outliers



DATA EXPLORATION

- Ambient pressure (AP)

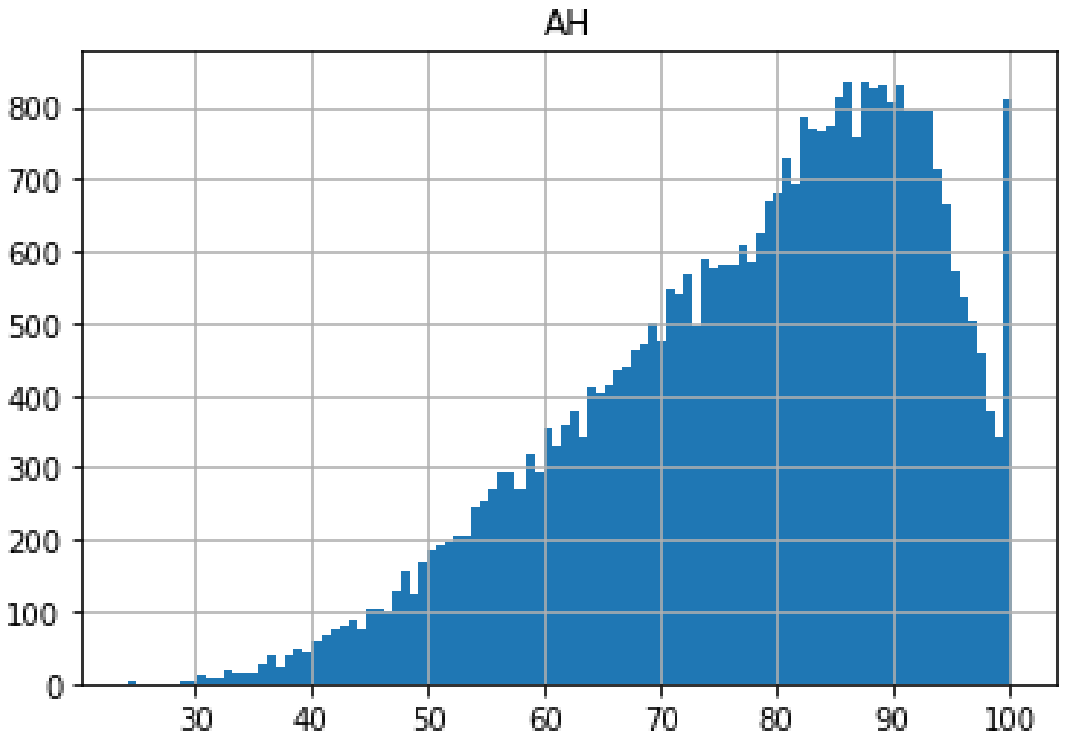
- Fairly evenly distributed on the dataset
- Few outliers on the lower end



DATA EXPLORATION

- Ambient humidity (AH)

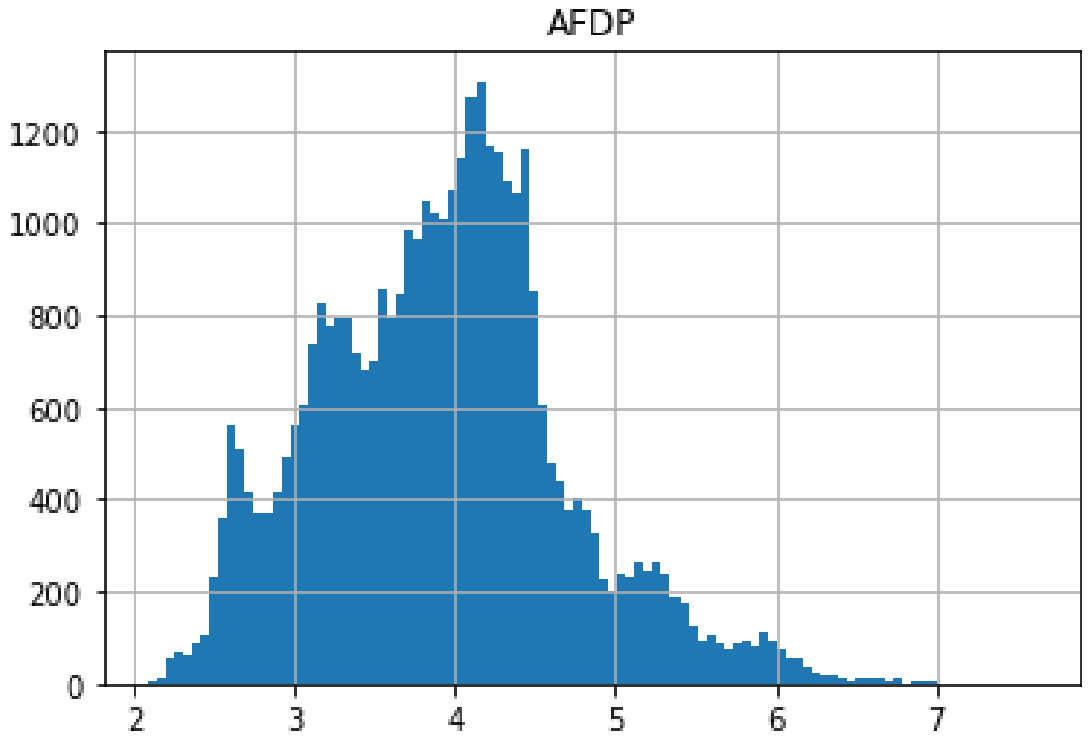
- Power plant location is fairly humid
- Range of data looks great, with data from 24% to 100% humidity
- Few outliers on the lower end



DATA EXPLORATION

- Air filter difference pressure (AFDP)

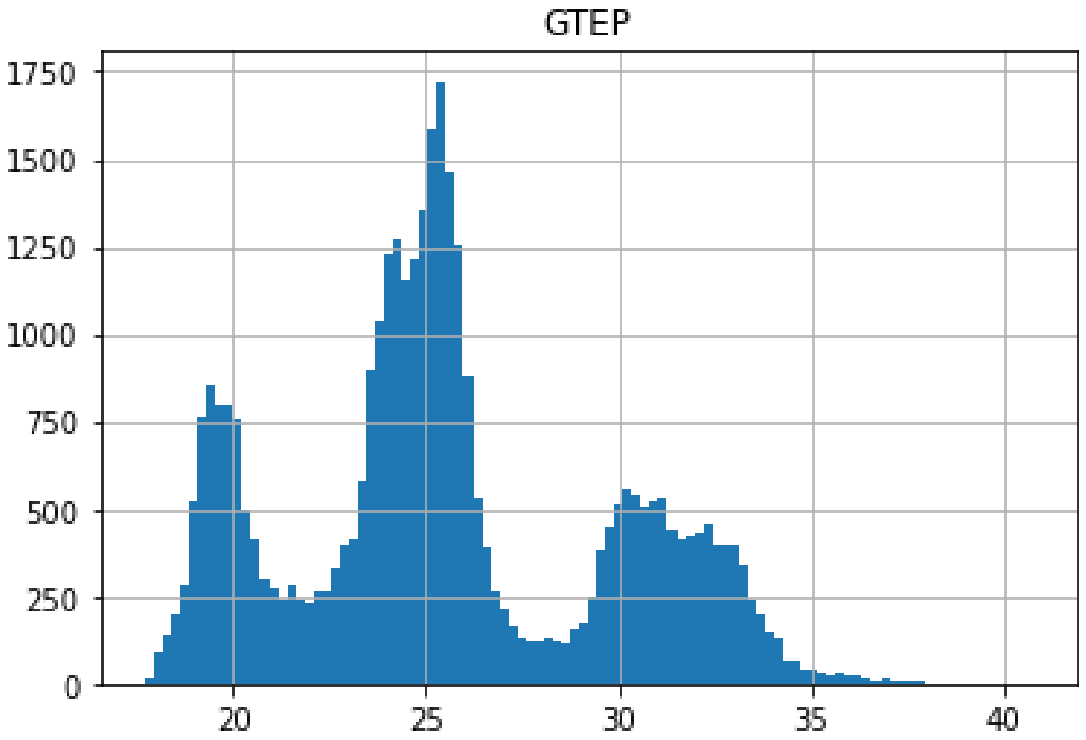
- Process variable, not straight-forward to interpret
- Range seems well-defined
- Few outliers on the upper end



DATA EXPLORATION

- Gas turbine exhaust pressure (GTEP)

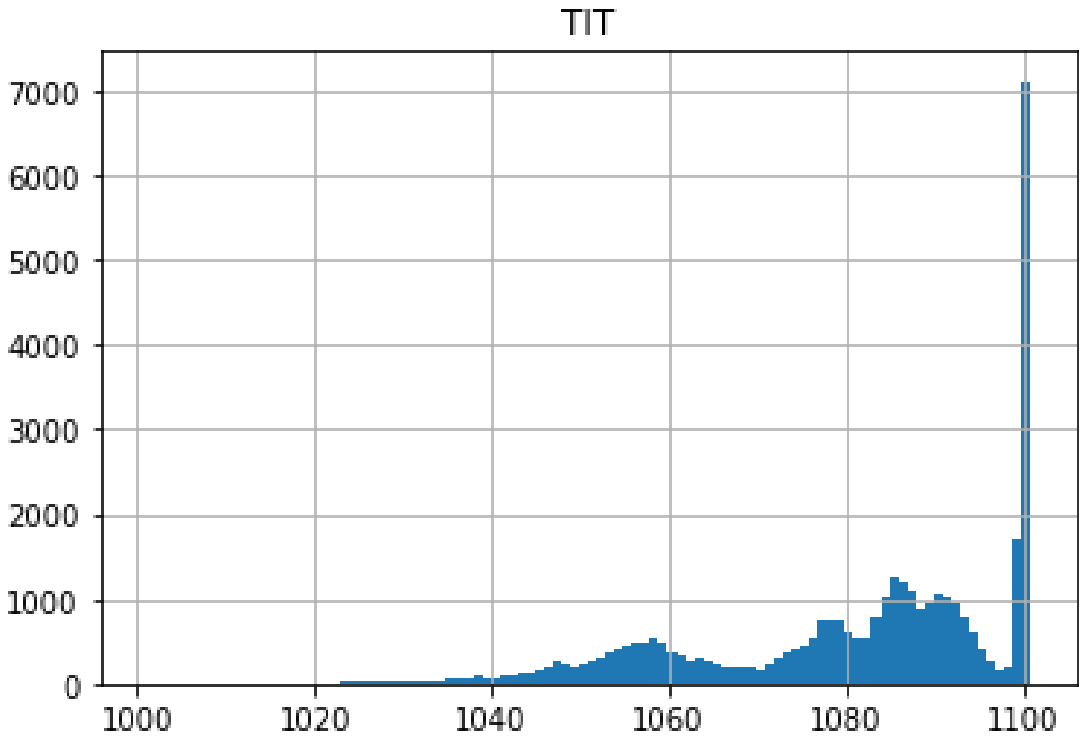
- Three main concentration points around 19, 25 and 32 mbar
- Possible candidate to be broken into labels (low, medium, high)
- May suggest the turbine has been operated in three main speeds
- Few outliers on the upper end



DATA EXPLORATION

- Turbine inlet temperature (TIT)

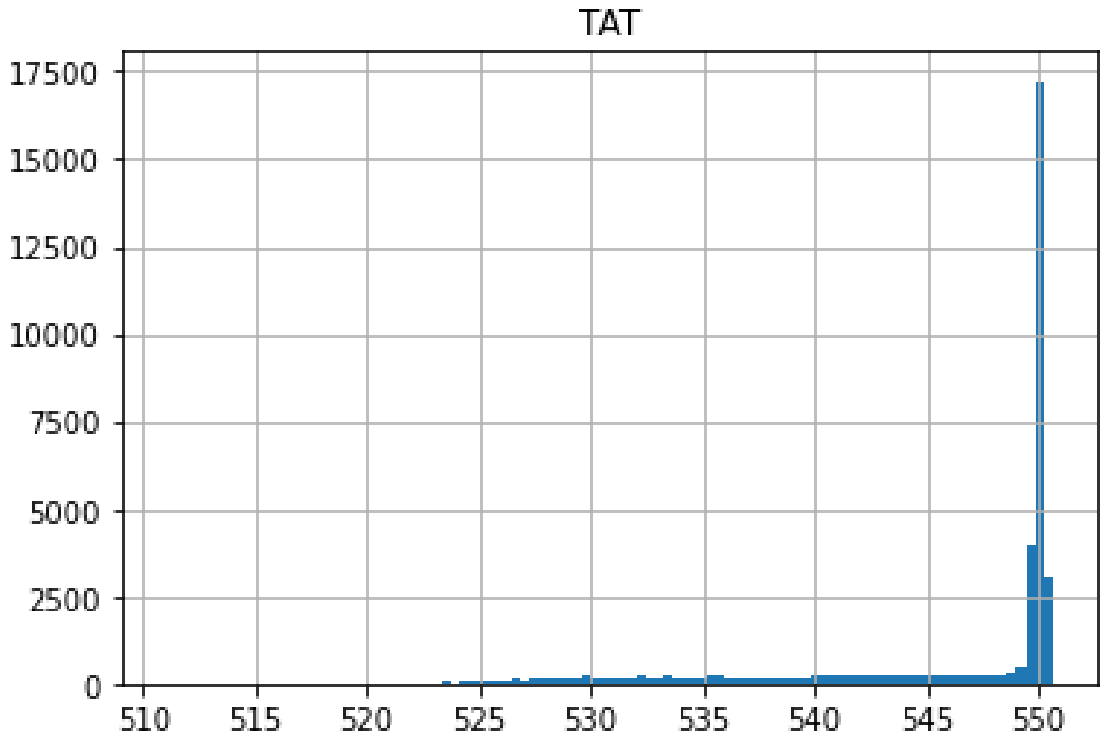
- Data seems highly concentrated on the value 1100 degC
- Suggests a main operation steady-state
- Outliers on the lower end



DATA EXPLORATION

- Turbine after temperature (TAT)

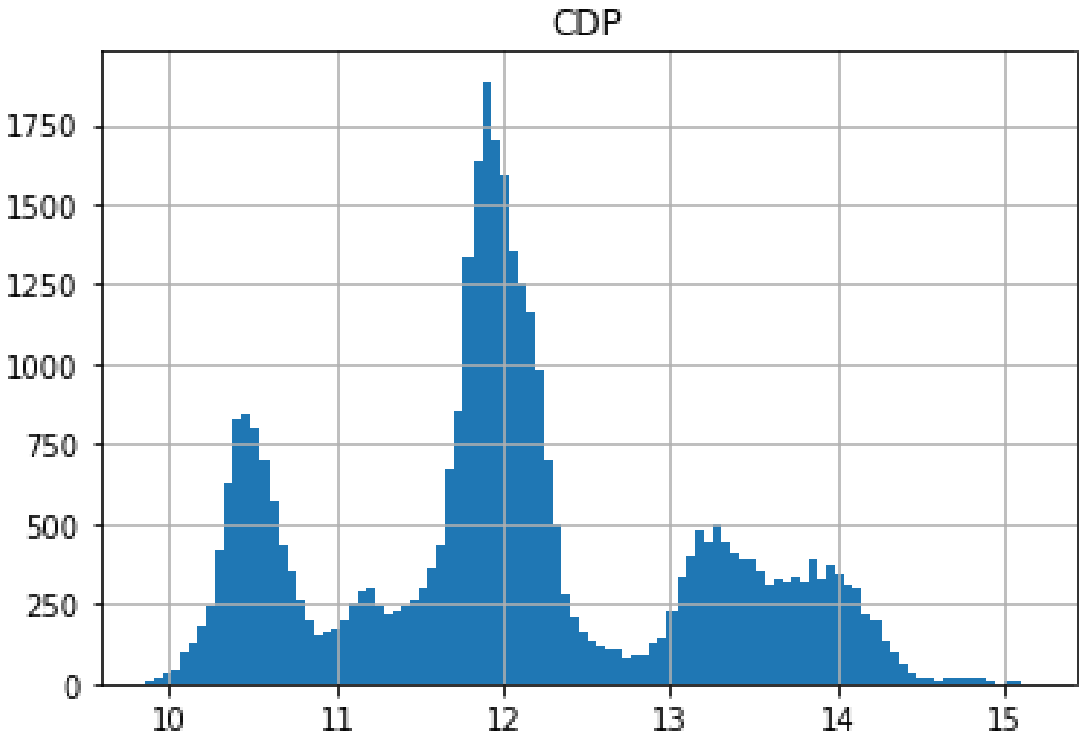
- Very concentrated at 550 degC
- May not be ideal to be used as a feature due to low variability
- Outliers on the lower end



DATA EXPLORATION

- Compressor discharge pressure (CDP)

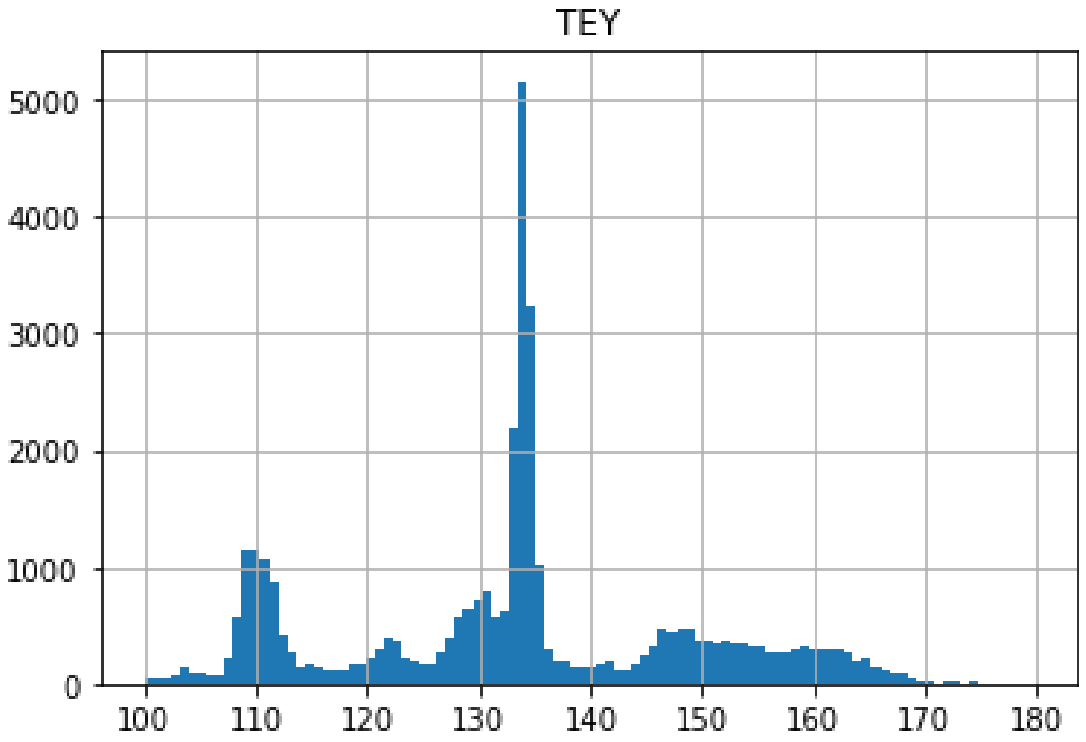
- Visually is very correlated to GTEP, which makes sense once it's also a pressure reading located after the burner
- Strong candidate for follow-up correlation analysis with GTEP
- Three main concentration points around 10.5, 11.8 and 13.4 mbar
- Few outliers on the upper end



DATA EXPLORATION

- Turbine energy yield (TEY)

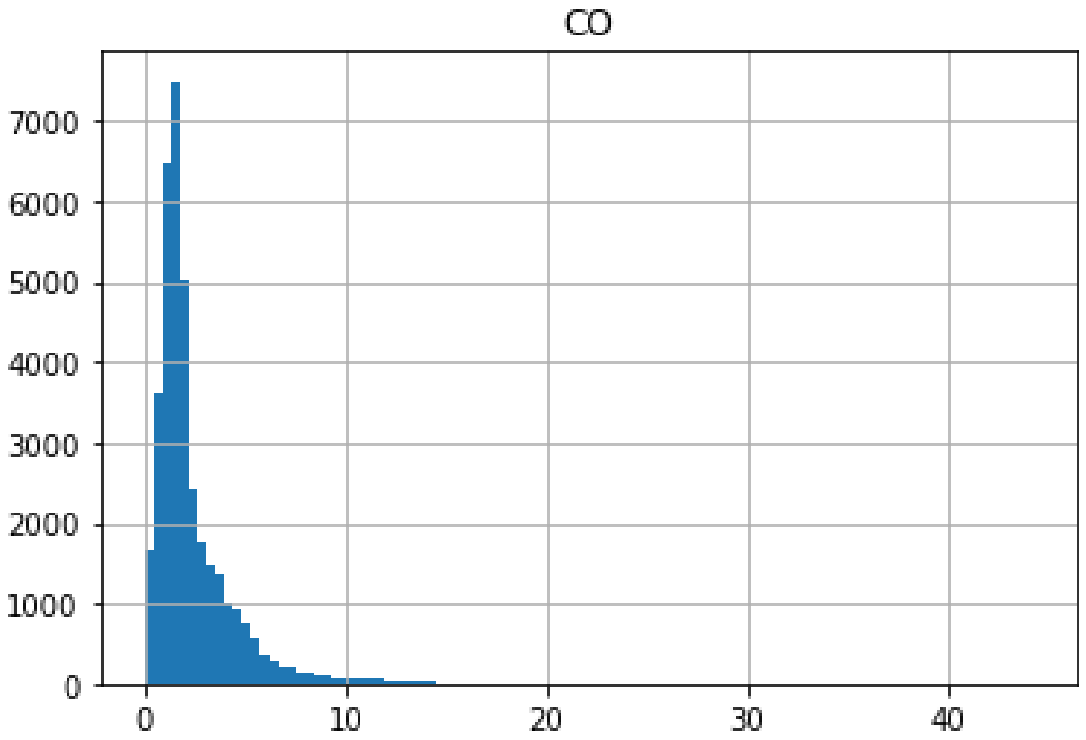
- It's clear the turbine operates most of the time producing 135 MWh
- It suggests that such operational steady-state is what generates concentrated distributions for TIT, TAT
- Central peak with two humps on each side is also similar profile as GTEP and CDP
- Few outliers on the upper end



DATA EXPLORATION

- Carbon monoxide (CO)

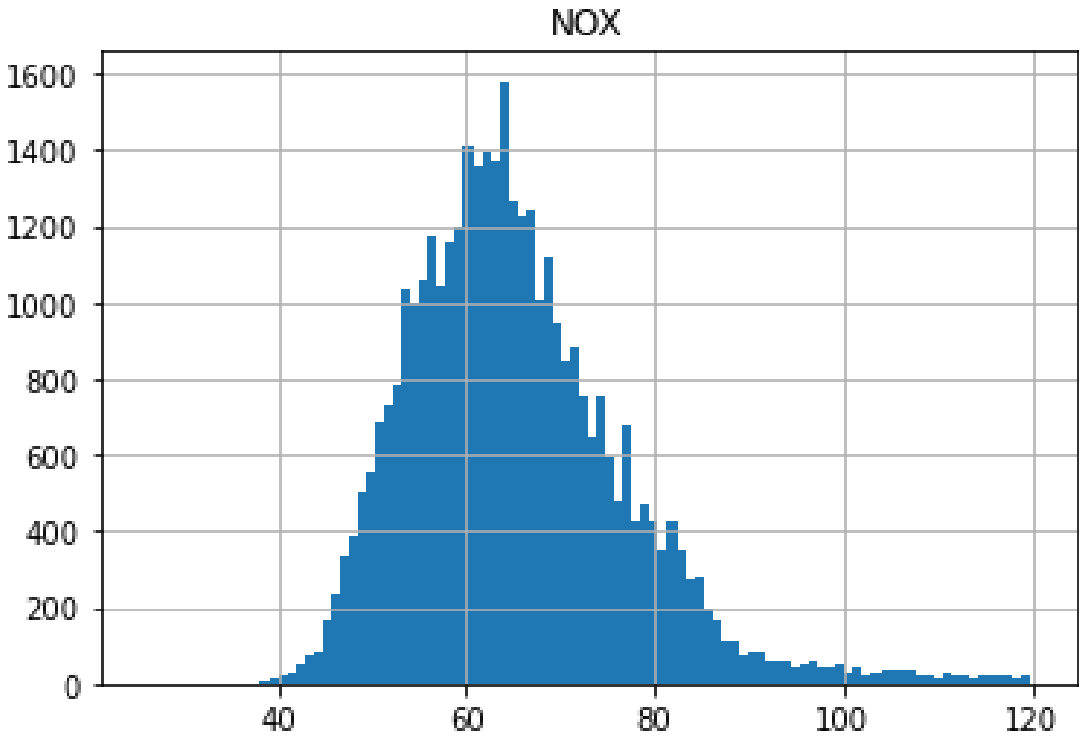
- CO concentration is skewed, most of the time below 5 mg/m3
- Outliers on the upper end



DATA EXPLORATION

- Nitrogen oxides (NOX)

- NOX concentration is way more distributed than CO
- Few outliers on the lower end



DATA EXPLORATION - TARGET

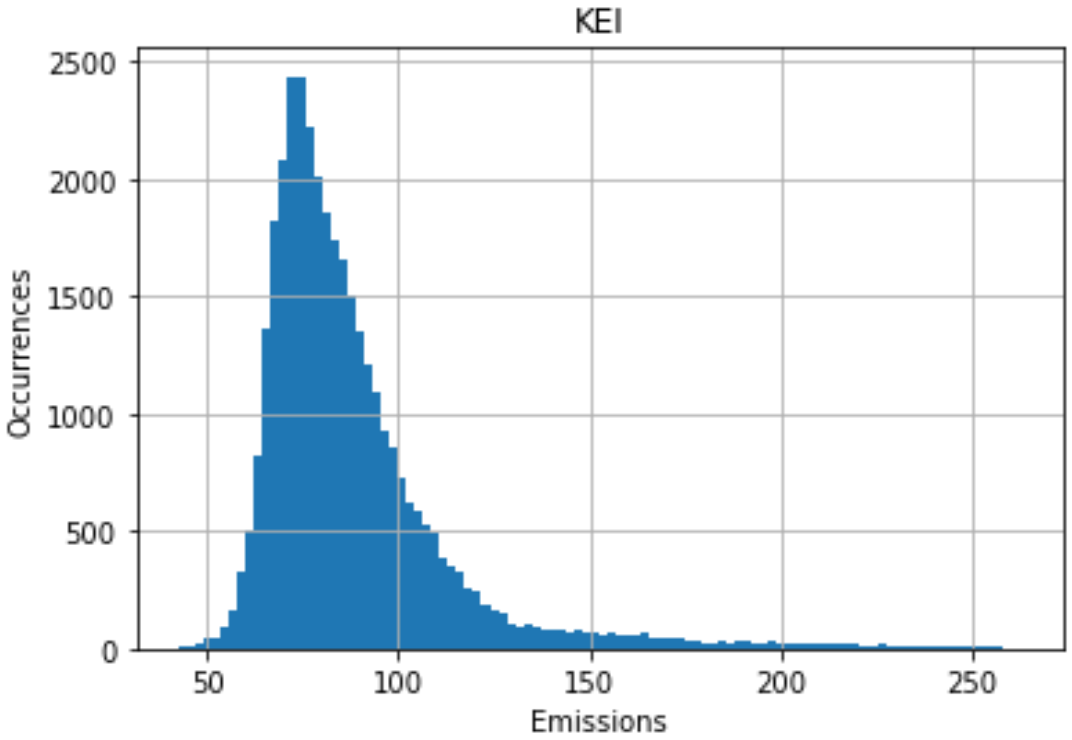
- Definition of Target Variable
  Key Emissions Indicator
  KEI = 10 * CO + NOX

- KEI clearly contains the skewness of CO distribution, along with the wider range of NOX distribution

- Given the goal of KEI it seems well balanced to be used as target of the ML algorithm

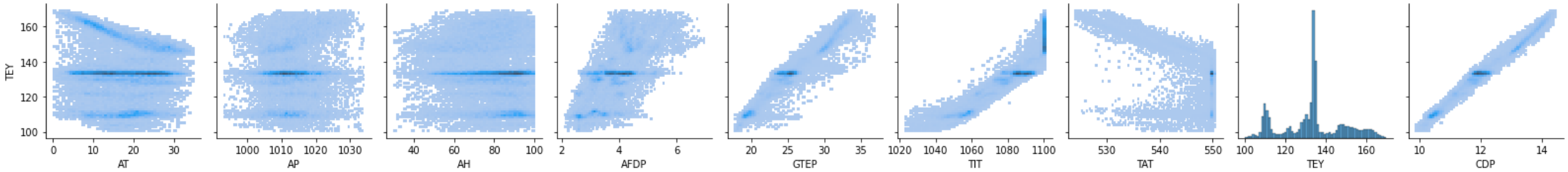
- Being a dependent variable, the outliers were already removed from CO and NOX

- Range of 219.92
  - Percentage of error over this range (%E) will be used as a metric to evaluate performance of models



DATA EXPLORATION - RESULTS

- TEY has a strong linear correlation with GTEP, TIT and CDP

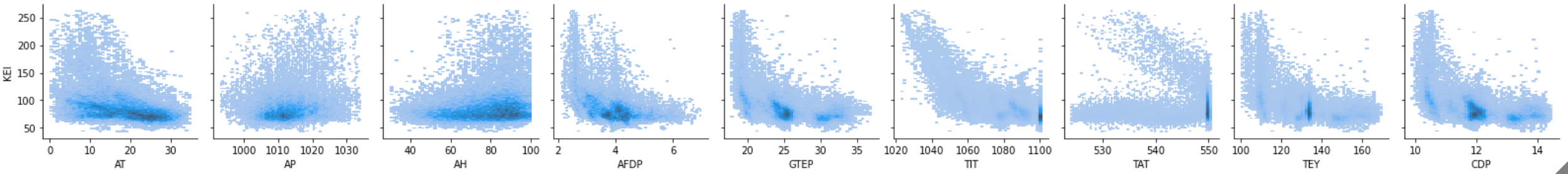


Feature	Correlation to TEY
AT	-0.07161
AP	0.100609
AH	-0.13777
GTEP	**0.963983**
TIT	**0.914626**
TAT	-0.68725
AFDP	0.665742
CDP	**0.988653**

DATA EXPLORATION - RESULTS

- Ambient features have lower correlation to KEI when compared to process variables
- AT seems to be the most important ambient feature
- As predicted on the distribution analysis section, TAT has almost no variation, thus ends up being a weak feature to use in this dataset. Maybe a more sensitive sensor to collect a wider range of data would help to increase the relevancy of this feature
- All the other process variables seem to have a somewhat similar correlation with KEI. They range on the 40% to 65% range and the shape is between a linear and a quadratic expression

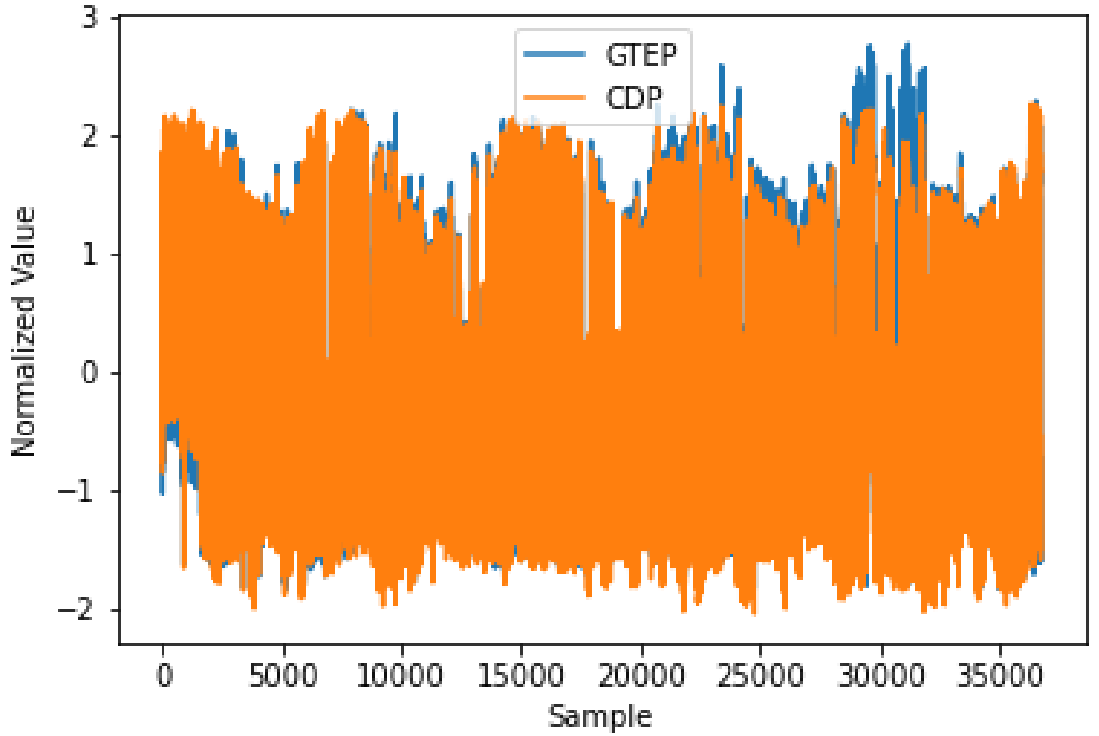
Feature	Correlation to KEI
TIT	-0.665
CDP	-0.5414
GTEP	-0.52915
TEY	-0.52616
AFDP	-0.46942
AT	-0.40402
TAT	0.048608
AP	0.152253
AH	0.164995



DATA EXPLORATION - RESULTS

- As suggested by the distribution analysis, being both pressure readings from the turbine exhausts, GTEP and CDP profiles match very closely

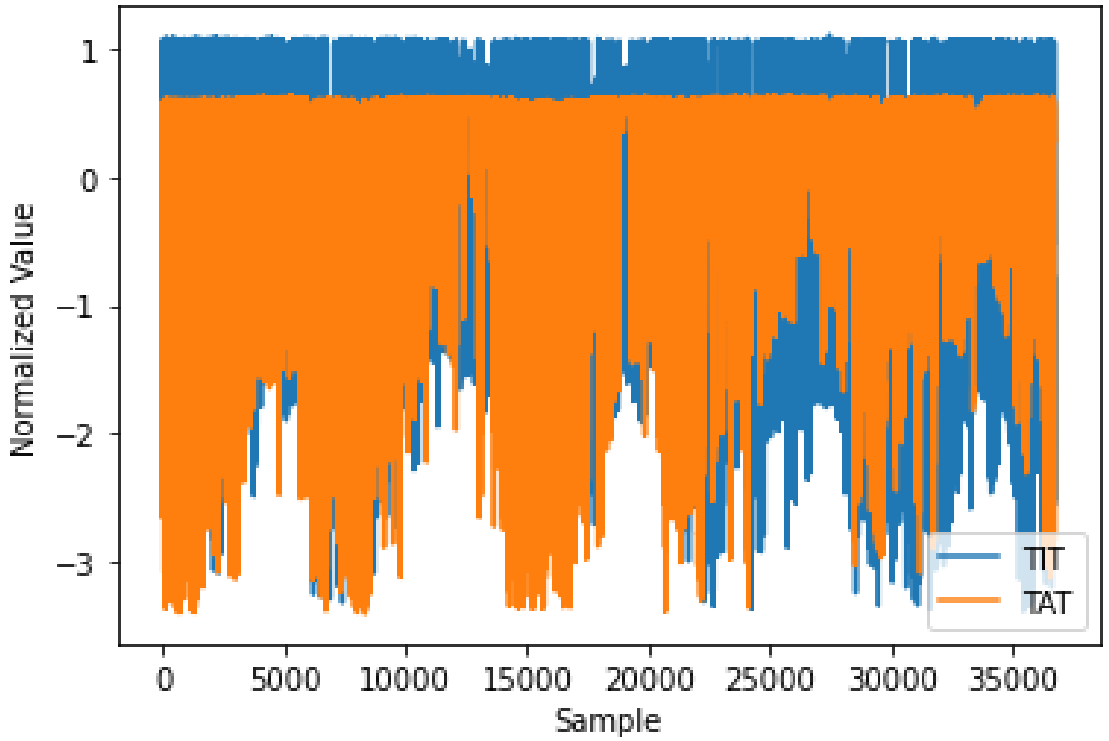
- Indicates that only one of them should be enough to achieve a satisfactory result on modelling KEI



DATA EXPLORATION - RESULTS

- The distribution analysis, and the fact that both measurements are temperatures in different stages of the turbine, had suggested that TIT and TAT could have the very same temporal profile

- But checking the comparison graphs above indicates that even though a similarity exists, the match of the profile is not as tight as imagined.



DATA EXPLORATION - RESULTS

- Baseline Model

- The baseline model to be used for this dataset shall be a linear model consisting of the main independent features identified in the dataset

- Given the close relationship between TEY and TIT, only TIT will be used in the baseline, once it has the best correlation with KEI

- Given the profile similarity between GTEP and CDP, only CDP will be used in the baseline, once it has the best correlation with KEI

- Process variables have way better correlation than ambient weather data

- Features for the baseline model will be:
  - TIT
  - CDP
  - AFDP

DATA EXPLORATION - RESULTS

- First Model

- As a first model, given the visual indication that the relationship may not be linear, a quadratic model seems to be a good fit to offer a noticeable improvement over the baseline

- Regarding the model features, it also seems to be worth experimenting with adding features that were not used on the baseline, but that still possess good correlation:
  - GTEP
  - TEY
  - AT

BASELINE MODEL

- Linear Regression
  - Train: R2=0.466 | MAE=13.5 | MSE=351.2 | RMSE=18.7 | %E=8.5
  - Test: R2=0.484 | MAE=13.4 | MSE=341.9 | RMSE=18.5 | %E=8.4
  - Coefficients = [-25.98451945 11.19343522 -1.70719158]
  - As expected from the EDA, TIT has the highest coefficient, followed by CDP, and then AFDP
  - Regressor doesn't seem to overfit, once Test and Train scores are similar

- Ridge Regression + Cross Validation
  - Train: R2=0.466 | MAE=13.5 | MSE=351.2 | RMSE=18.7 | %E=8.5
  - Test: R2=0.484 | MAE=13.4 | MSE=341.9 | RMSE=18.5 | %E=8.4
  - Coefficients = [-25.97562879 11.18500457 -1.70739805], Best Alpha = 1
  - Ridge regression arrived at same results as basic linear regression, which is understandable given only 3 features

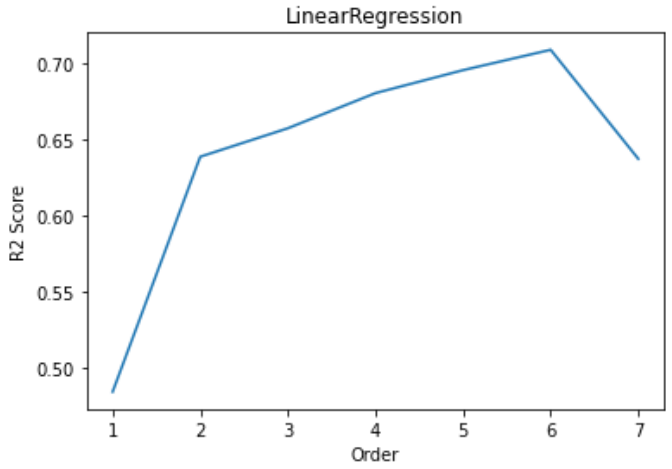
BASELINE MODEL

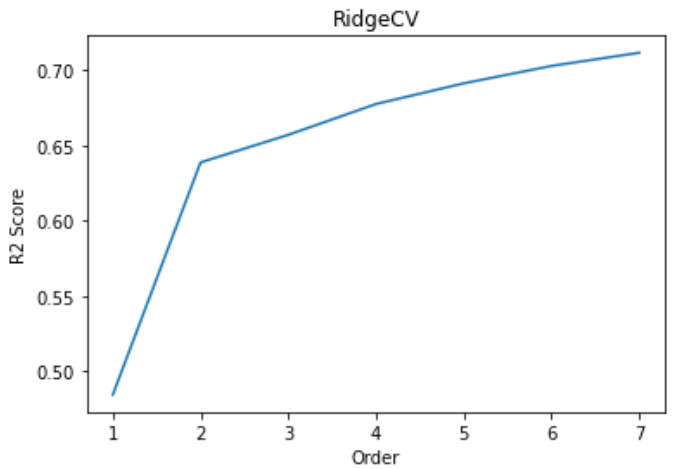
- Lasso Regression + Cross Validation
  - Train: R2=0.435 | MAE=13.3 | MSE=371.8 | RMSE=19.3 | %E=8.8
  - Test: **R2=0.453** | MAE=13.2 | MSE=362.3 | RMSE=19.0 | **%E=8.7**
  - Coefficients = [-15.94357744 0. -0.], Best Alpha = 1
  - Arrived at interesting result, where TIT is the only feature selected by it

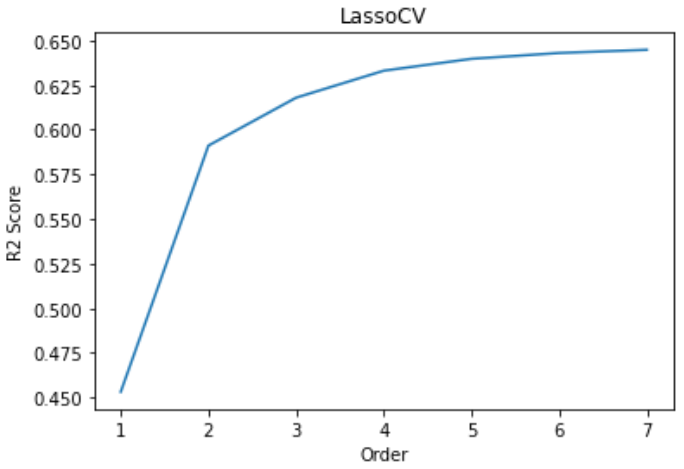
- Given the small difference from using all 3 features and using just TIT with Lasso, and given the simplicity of the single feature, it can be said that the Baseline is the result from Lasso regression

POLYNOMIAL MODEL

- Evaluation on Order 1 to 7 using same features as baseline
  - R2 Score used for comparison given the linearity observed with other metrics
- The R2 Score improves dramatically when using higher order polynomials
- On all tested cases a polynomial of order 2, and then order 3, bring significant gains
- The best score was achieved using the Ridge regularization method, with a polynomial of order 7
- It certainly is a high order value but given that that turbines work at high-speed it is not totally unexpected to need higher order coefficients to properly model the dynamic behaviors inside it

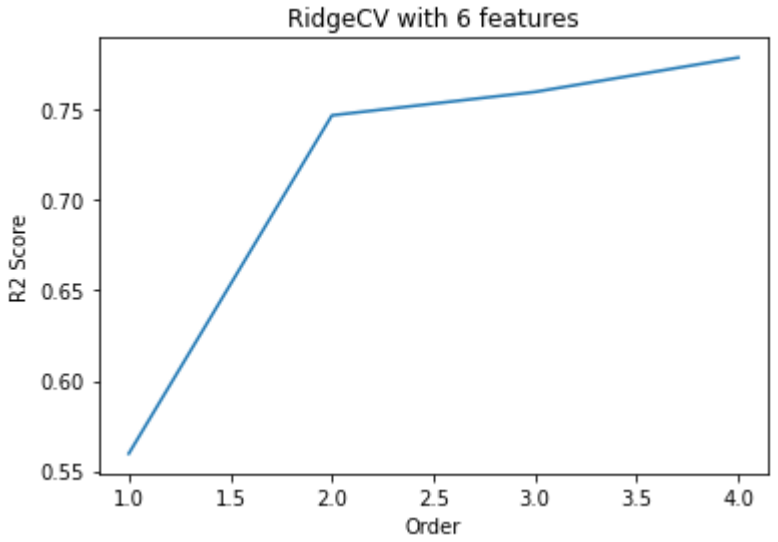


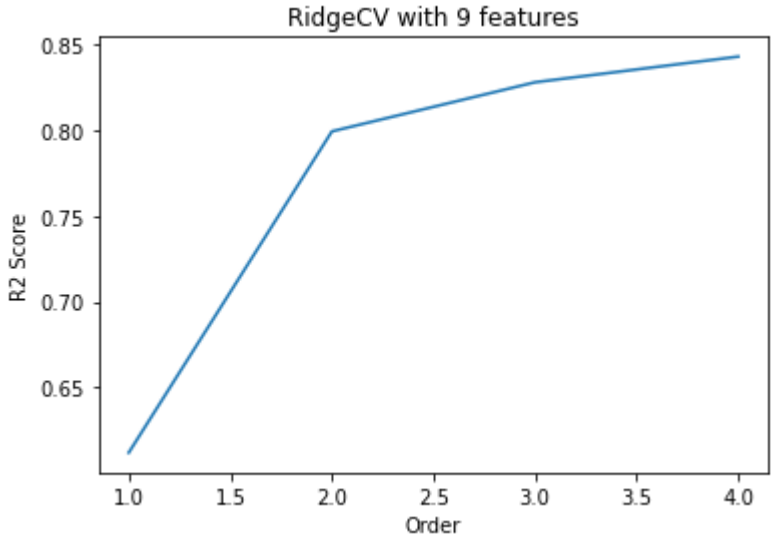




POLYNOMIAL MODEL WITH ADDITIONAL FEATURES

- Evaluation set of main 6 features identified on EDA
  - Lower order models achieve results similar to the ones seen with just the 3 main features
  - For example, a quadratic model is already able to achieve a score of **0.747**, which is better than the best case for the baseline features case
- Evaluation using all 9 features
  - Improves the overall score by almost 10%
  - Up to a model of order 3 the test dataset continuously improves and performs better than the train one
  - After that (order 4) the train data set outperforms the test one, which is an early indication of overfitting





FINAL POLYNOMIAL MODEL

- **Order 3** Polynomial using all available features

- Results:
  - **Train**: R2=0.824 | MAE=7.4 | MSE=116.1 | RMSE=10.8 | %E=4.9
  - **Test**: **R2=0.828** | MAE=7.3 | MSE=113.9 | RMSE=10.7 | **%E=4.9**
  - Best Alpha = 1

- It clearly outperforms the **Baseline** model, which had a score of **0.453** and percentage error over the target range of **8.7**

- Given that all the features from this dataset are commonly required for turbine operations anyway, it's deemed acceptable to use the value of all sensors to feed the emissions estimation model

- A polynomial of order 3 is feasible to be implemented and cyclically executed on an embedded real-time system hardware

COMPLEX MODEL

- Experiment with modern models in use by the industry

- Gradient Boosting
  - Ensemble of weak prediction models
    - Weak models are typically decision trees
  - Relies on the intuition that the best possible next model, when combined with previous models, minimizes the overall prediction error
  - Each new model takes a step in the direction that minimizes prediction error
  - Usually outperforms random forest
- Two main library publicly available
  - XGBoost
  - LightGBM

COMPLEX MODEL

- XGBoost
  - Train: R2=0.970 | MAE=3.0 | MSE=19.7 | RMSE=4.4 | %E=2.0
  - Test: R2=0.859 | MAE=6.5 | MSE=93.7 | RMSE=9.7 | %E=4.4
  - Best Model Parameters
    - 'max_depth': 6
    - 'n_estimators': 500
    - 'reg_lambda': 10
  - Great accuracy out of the box
  - Tends to overfit
    - Hyperparameter tunning required to avoid it

COMPLEX MODEL

- LightGBM
  - Train: R2=0.915 | MAE=5.1 | MSE=56.2 | RMSE=7.5 | %E=3.4
  - Test: R2=0.864 | MAE=6.4 | MSE=90.1 | RMSE=9.5 | %E=4.3
  - Best Model Parameters
    - 'max_depth': 6
    - 'n_estimators': 500
    - 'reg_lambda': 10
  - Way faster to execute than XGBoost, which allows for a lot more experimentation
  - Arrives at better results than XGBoost based on same hyperparameter range
  - Less overfitting than XGBoost
    - More test cases will be performed using it

FINAL COMPLEX MODEL

- LightGBM selected as preferred model
- Extensive experimentation with hyperparameter tuning to prevent overfitting
- Best compromise between accuracy and overfitting
  - **Train**: R2=0.866 | MAE=6.3 | MSE=87.8 | RMSE=9.4 | %E=4.3
  - **Test**: **R2=0.850** | MAE=6.8 | MSE=99.3 | RMSE=10.0 | **%E=4.5**
  - Best Model Parameters
    - 'bagging_fraction': 0.9
    - 'bagging_freq': 15
    - 'feature_fraction': 1
    - 'max_depth': 6
    - 'min_data_in_leaf': 300
    - 'min_sum_hessian_in_leaf': 0.001
    - 'n_estimators': 480
    - 'num_leaves': 20
    - 'reg_lambda': 1

FINAL MODEL

- LightGBM selected as final model
  - Performed better over all metrics analyzed

- Train

Model	R2	MAE	MSE	RMSE	%E
Baseline	0.435	13.3	371.8	19.3	8.8
Polynomial	0.824	7.4	116.1	10.8	4.9
LightGBM	0.866	6.3	87.8	9.4	4.3

- Test

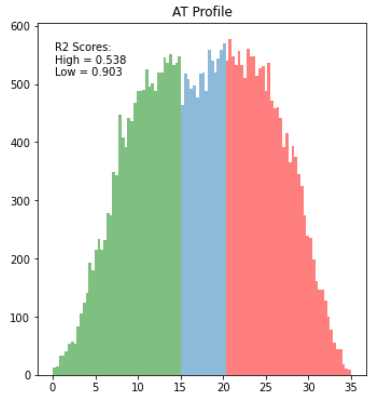
Model	R2	MAE	MSE	RMSE	%E
Baseline	0.453	13.2	362.3	19.0	8.7
Polynomial	0.828	7.3	113.9	10.7	4.9
LightGBM	0.850	6.8	99.3	10.0	4.5

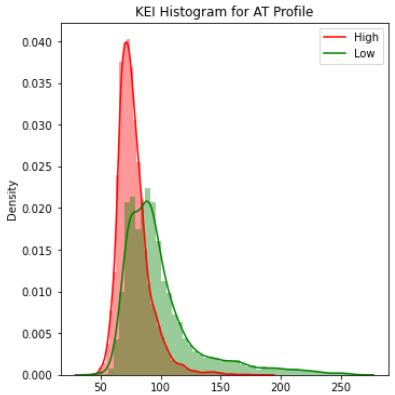
- Industrial sensors have accuracy on the 5% to 20% range
  - 4.5% error is acceptable

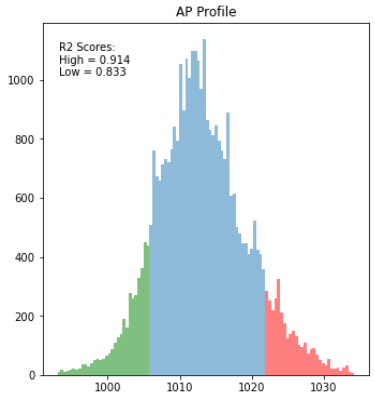
MODEL PROFILING

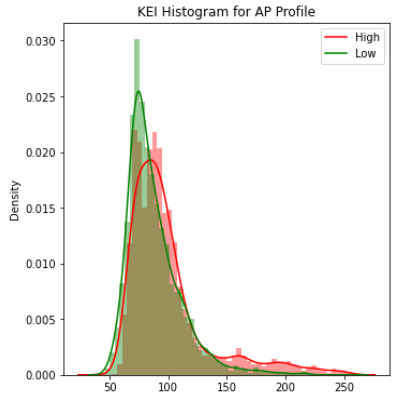
- High x Low Range of Feature Space
  - 2 x Std. Dev. wide

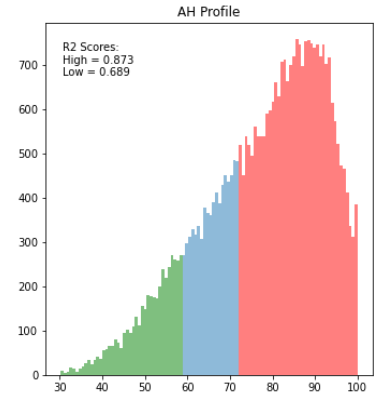
- Range distribution over the feature space
- Range distribution over the target space

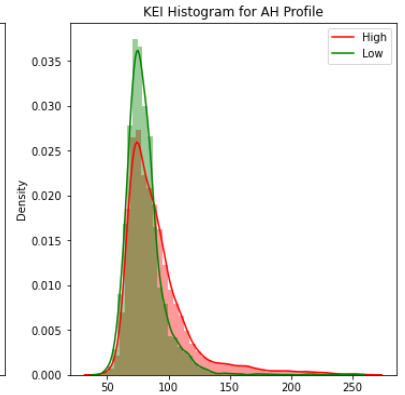


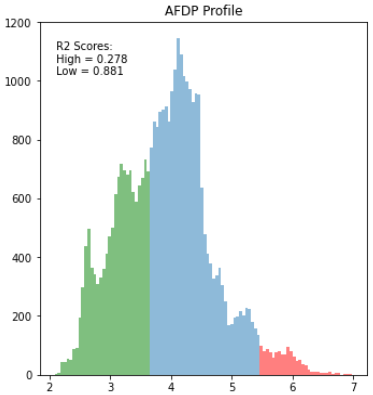


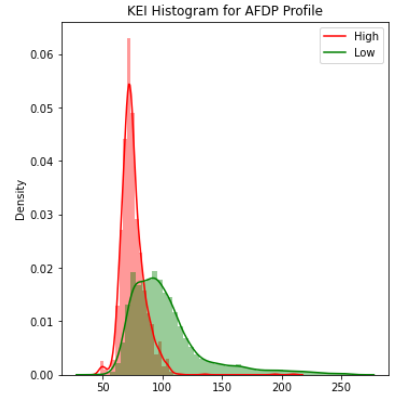


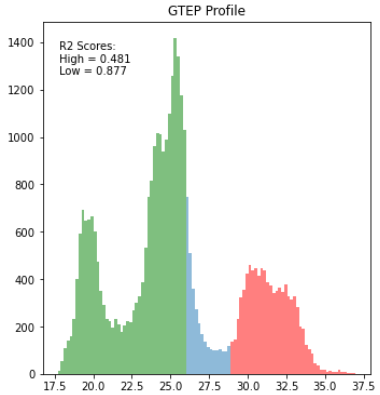


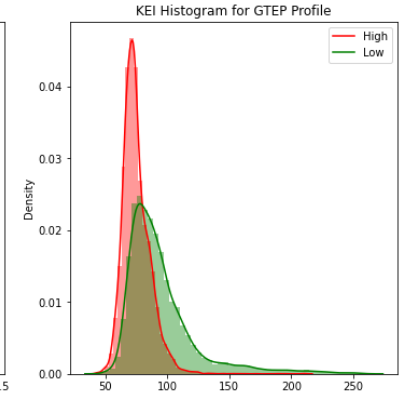


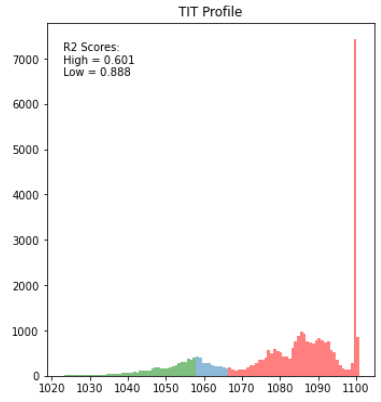


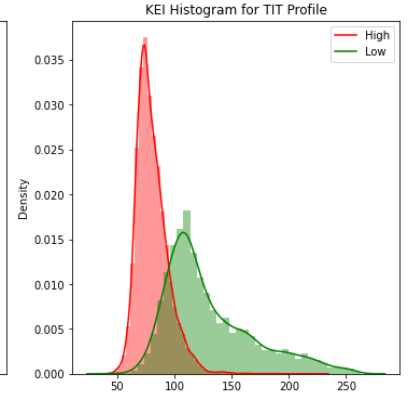






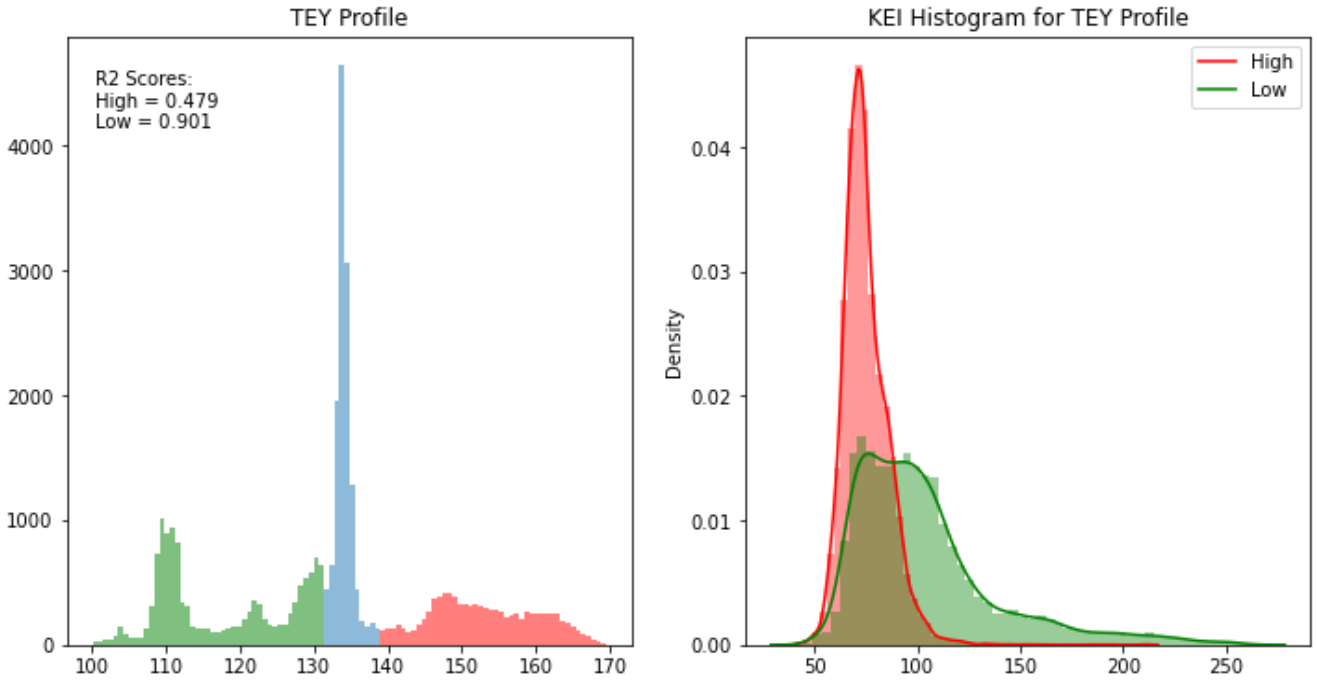


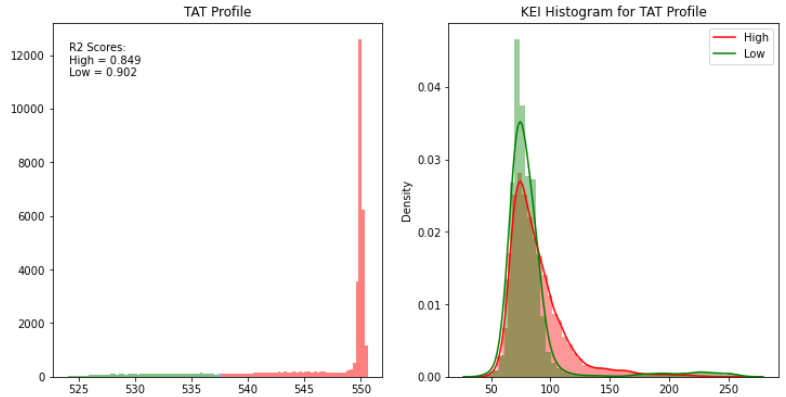


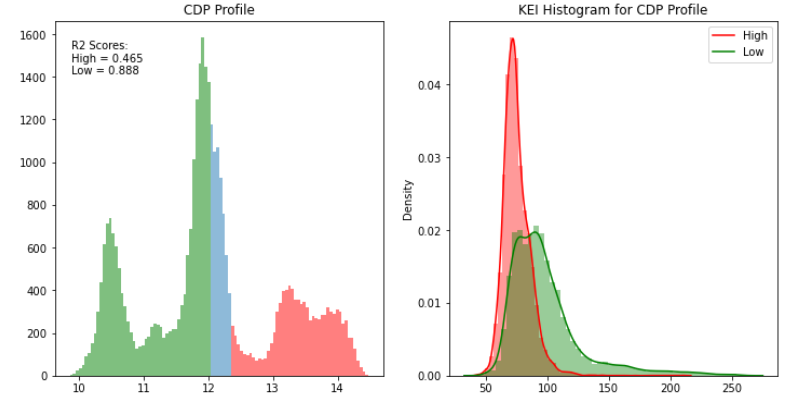


MODEL PROFILING

- One range side always has the R2 score better than the model average of 0.850
- Worst performing side of each feature caused by narrow (non-gaussian) distribution of target variable







MODEL PROFILING

- Model Improvement Suggestions
- Transforming the target variable to have a better (Gaussian) distribution
- ‘Oversampling’ the worst performing ranges on the training dataset
- ‘Undersampling’ the best performing ranges on the training dataset

FIELD DEPLOYMENT

- Leverage existing LightGBM library
  - Ported to a variety of programming languages
- Power plant with same characteristics of train dataset
  - Turbine Model
  - Fuel Type/Grade
- On-Premises Deployment
  - Access to live sensor data (or historian)
- Validation against calibrated gas concentration sensor
- Periodic maintenance/assessment performance

CONCLUSIONS

- ML is a feasible approach
- Next steps
  - Wider data range
    - More serial numbers
    - Wider operational window
    - More fuel types/grades
  - Iterative model training/testing based on the wider dataset
    - Profiling to assess the overall robustness
  - Live trial run of the model on an instrumented turbine
    - Evaluate quality of the estimation


THANK YOU