Report title

Sebastiano Barresi
Politecnico di Torino
Student id: s292519
s292519l@studenti.polito.it

Edoardo Marchetti
Politecnico di Torino
Student id: s303873
s303873@studenti.polito.it

***Abstract*—In this report we propose an approach to the problem of building a Predictive emission monitoring system (PEMS) to predict the amount of carbon monoxide (CO) emitted by a gas turbine in a power plant. This solution of the regression task contains an in-depth analysis of the given dataset and a comparison between several known regression models, which led to achieve more than good performances both in testing the chosen model and in evaluating the results through the sumbission platform.**

I. Problem overview

The aim of this regression task is to create a reliable predictive emission monitoring system to predict the CO emissions of several gas-turbines. PEMSs constitute an important instrument in the validation an the backing up of costly continuous emission monitoring systems used in gas-turbine-based power plants. The gas combustion process is one of the sources of harmful pollutants released in the atmosphere, most common are CO and NO_x. The importance of building a trustworthy PEMS is not only of an economic nature, but also to prevent an increasing level of air pollution.

A. Data Exploration

The given dataset is composed of two parts: *Development dataset* and *Evaluation dataset*.

The Development dataset contains 24448 records, described by 15 features:

- **ID**: the unique ID used to identify the record,
- **YEAR**: the year in which the record has been collected,
- **LOC**: the state where the turbine is located,
- **SN**: the unique code that identifies the turbine's model,
- **AT**: ambient temperature (*Celsius degrees*),
- **AP**: ambient pressure (*mbar*),
- **AH**: ambient humidity (%),
- **AFDP**: air filter difference pressure (*mbar*),
- **GTEP**: gas turbine exhaust pressure (*mbar*),
- **TIT**: turbine inlet temperature (*Celsius degrees*),
- **TAT**: turbine after temperature (*Celsius degrees*),
- **TEY**: turbine energy yield ($MW/hour$),
- **CDP**: compressor discharge pressure (*mbar*),
- **NOX**: nitrogen oxide emitted (mg/m^3),
- **CO**: carbon monoxide (mg/m^3).

The Evaluation dataset contains 12245 records, described by all the features above except for the CO, which is the target variable.

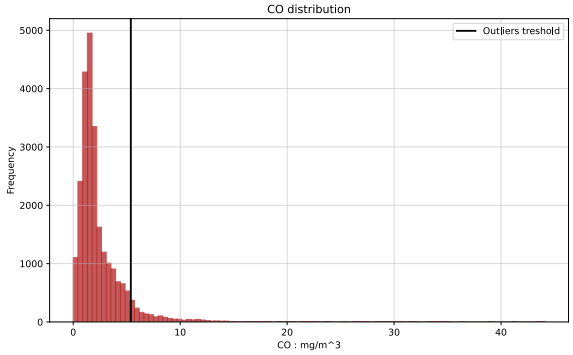

Fig. 1. Distribution of the CO emission of the records in the Development dataset. The black vertical line represents the threshold CO = 5.38 mg/m^3, above which the CO emission is categorized as "extreme".

An initial exploration of the dataset showed that there were no duplicated records and that all columns, with the exception of NO_x, had no missing values. Next, the distribution of turbines with respect to LOC feature was analysed and it was noted that the data only referred to 27 European countries. For each country, an average of 906 records (i.e. 3.7% of the total data) are reported in the development dataset. The dataset covers five years of observations, namely from 2018 to 2022. A similar dataset analyzed in [1] has been found, presenting data from 2011-2015. The records are not reported in temporal order, in fact no correlation between ID and YEAR columns was observed. Finally, the SN column was analysed to see how many turbine models the dataset referred to. It was revealed that all the data belonged to a single model, the 0903XTR. The same considerations were also made for the Evaluation dataset.

B. Feature analysis

A more in-depth univariate analysis was conducted on two groups of predictors:

- **Ambiental predictors**: AT, AP, AH;
- **Processing predictor**: AFDP, GTEP, TIT, TAT, TEY, CDP.

AT and AP show bell shape distributions in the first group, whereas AH has a more skewed distribution. Among the processing predictors, on the other hand, TIT and TAT have a distribution with a high skewness coefficient while TEY,

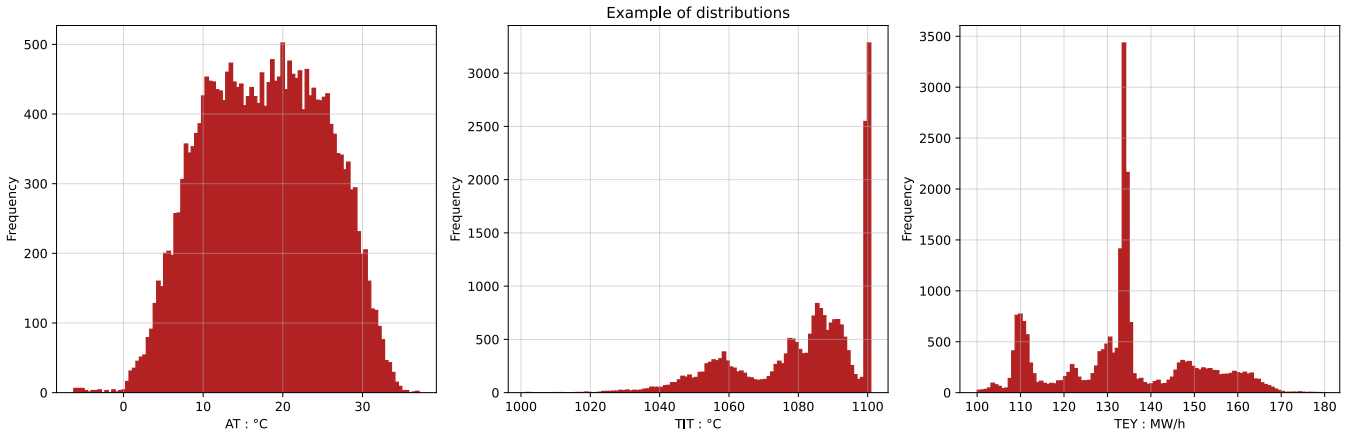

Fig. 2. The figure reports examples of the three different kind of feature's distribution contained in the Development dataset: AT presents a bell-shape distribution, TIT a high skewed distribution and TEY a multimodal distribution.

GTEP and CDP have a multimodal pattern [2]. Example of this distribution diversities can be seen in Fig. 2.

Analyzing the distribution of CO values it was observed that it has a high skewness coefficient (Fig. 1). In fact 93% of the values are enclosed in the interval [0, 5.38], while the remaining 1681 records are detected as outliers. However, as suggested by [3], the latter should not be eliminated from the analysis, but rather considered as "extreme" emissions since the same variations are also observed in the predictors. Subsequently a multivariate analysis between YEAR-CO and LOC-CO was performed. From the first it was observed how the emission values seem to decrease over the 5 years (we can hypothesize that it is due to the use of more efficient technologies, but we do not have enough information to confirm this), while from the second it was understood that the locality does not have a a sensitive impact on the amount of CO released by the turbine.

As a last step for the feature analysis, correlations were studied. From the Fig. 3 it can be easily seen that there are highly related pairs:

- ρ(TEY, CDP) = 0.99,
- ρ(CDP, GTEP) = 0.98,
- ρ(TEY, GTEP) = 0.96,
- ρ(TEY, TIT) = 0.91,
- ρ(TIT, GTEP) = 0.87.

Compared to the target variable, on the other hand, TIT is the feature with the highest Pearson coefficient in absolute terms, ρ(TIT,CO) = -0.70. Moreover, we observed that the processing predictors seem to have a second order correlation with CO.

II. Proposed Approach

In this section we describe how we processed the data before starting model training and how we selected the best regressor. The transformations described are applied both on development and the evaluation datasets.

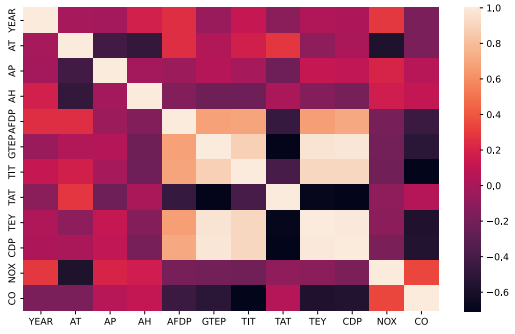

Fig. 3. The figure reports the heatmap showing the Pearson correlation among the features of the Development dataset.

A. Preprocessing

First we removed the SN column as the 0903XTR model is shown in all records and the ID. We then applied one-hot encoding to the LOC column to encode each individual country. To manage the missing values in the NOX column we used a KNNImputer from the python scikit-learn library [4]. In this way, each value is calculated based on the NOX values of the most similar records. Then we mapped the YEAR values from the interval [2018-2022] to [1-5]. The preprocessing phase ended with the application of a StandardScaler to ensure that all predictors had similar domain values. After the preprocessing the dataset consist of 38 predictors: 14 (*initial predictors*) + 27 (*countries*) - 3 (*LOC, SN, ID*).

B. Model and Feature selection

Various regression models available on scikit-learn were trained to predict the CO emissions:

- **LinearRegression, Ridge, Lasso**: models which fit a linear model with coefficients to minimize the residual sum of squares between the observed targets in the dataset

and the targets predicted by the linear approximation. The Ridge and the Lasso models allow to impose a penalty to the size of the coefficients, limiting the complexity of the model.

- **MLPRegressor**: A multi-layer perceptron regressor which optimizes the squared error using LBFGS or stochastic gradient descent.
- **SVR**: The regression version of the Support Vector Classifier. The model is created based only on a subset of training data, because the cost function ignores samples whose prediction is close to their target.
- **RandomForestRegressor (RFR)**: An ensamble model that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

By training the default versions of the models reported the best performances in terms of MSE (mean squared error) have been recorded by RFR and SVR and shown in Table I.

TABLE I
REGRESSORS COMPARISON

Model	MSE
Linear Regressor	1.705
Ridge Regressor	1.705
Lasso Regressor	3.598
MLP Standard	1.784
SVR	**1.361**
RFR	**1.377**

Analyzing the features importance returned by the random forest it was seen that, as expected, TIT (highest correlation coefficient) has the greatest importance, while all countries have practically no impact on the final calculation of the emission value. Surprisingly the second most important feature is TAT, ρ(TAT, CO) = 0.05. Considering instead the pairs of features that have a high correlation value, it has been shown that CDP and TEY are not among the most relevant predictors for the final calculation of CO. For this reason we made a first feature selection by testing SVR and RFR on the development dataset to which the combinations of CDP, TEY and GTEP as well as all the country columns have been removed in turn. The best results were obtained by training the models on the dataset to which the TEY, CDP pair was removed as shown in Table II.

TABLE II
REMOVING FEATURES COMPARISON

Removed feature	CDP	X			X	X	
	TEY		X		X		X
	GTEP			X		X	X
Model	SVR	**1.355**	1.356	1.360	1.357	1.362	1.366
	RFR	1.351	1.392	1.383	**1.349**	1.412	1.371

As reported in I-B some predictors seemed to have a quadratic relation with the target variable. For this reason we tested the default versions of the RFR and of the SVR also on a dataset transformed through the PolynomialFeatures class with degree 2, 3 and 4. Results are reported in Table III.

TABLE III
POLYNOMIAL FEATURES WITH DIFFERENT DEGREE COMPARISON

Model	MSE
poly2 + SVR	1.423
poly3 + SVR	1.473
poly4 + SVR	1.576
poly2 + RFR	1.288
poly3 + RFR	1.197
poly4 + RFR	1.178

The results showed a lowering of the MSE produced by the RFR as the degree increases, while the SVR tends to worsen. Based on these results we have tried to do hyperparameters tuning on the following pipelines:

- PolynomialFeatures(*degree = 2*) + RandomForestRegressor()
- PolynomialFeatures(*degree = 3*) + RandomForestRegressor()

C. Hyperparameters tuning

After a first random search we obtained the set of parameters reported in the Table IV. As can be seen, the model with higher complexity suffers from over-fitting, while the model with *degree = 2* has an even lower MSE in the public score.

TABLE IV
RANDOMSEARCH RESULTING PARAMETERS

PolyFeatures params	RFR params	Training Score	Public Score
degree = 3	# estimators = 300 max depth = 110 min samples leaf = 1 max features = sqrt bootstrap = False	1.165	1.245
degree = 2	# estimators = 300 max features = log2	1.167	1.112

Starting from the hyperparameters obtained for the pipeline with *degree = 2*, we made a further HalvingGridSerch with the parameter ranges reported in Table V.

TABLE V
HALVINGGRIDSEARCH PARAMETER GRID

Parameters	
# estimators	200, 300, 400
max features	0.3, sqrt, log2, None
max depth	50, 100, 200, None
bootstrap	True, False

III. Results

The final model turns out to be the following pipeline:

1) **PolynomialFeatures**(*degree = 2, interaction_only = False, include_bias = True*)
2) **RandomForestRegressor**(*n_estimators=400, bootstrap = False, max_depth = 200, max_features = 'log2'*)

Thanks to this configuration we were able to reach an MSE of 1.1697 mg/m^3 on the Development dataset and a public score of 1.079 mg/m^3.

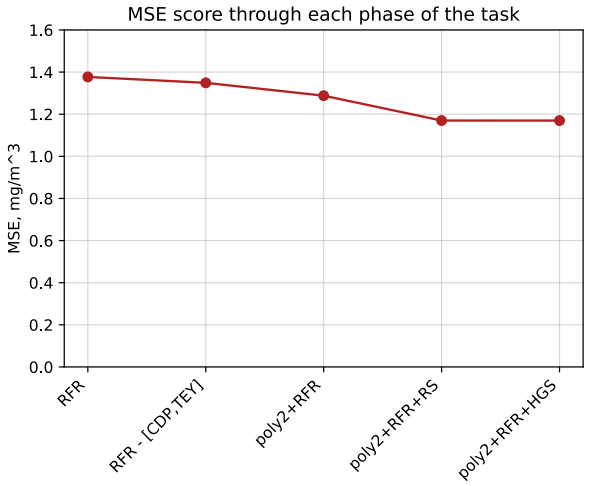

Fig. 4. The figure reports the decresing MSE trend acorss the tuning of the model. From the left: RFR = Default version of RandomForestRegressor; RFR - [CDP,TEY] = Default version of RandomForestRegressor trained on a dataset without CDP and TEY; poly2+RFR = Pipeline with default parameters; poly2+RFR+RS = Pipeline with parameters returned by random search; poly2+RFR+HGS = Pipeline with parameters returned by halving grid search.

IV. Discussion

Analyzing the Fig.5 it can be seen that the major part of the MSE value is due to the CO values that were earlier defined as extreme. So one possible explanation why the MSE in the training phase is worse than the public score is that there are fewer extreme CO values in the evaluation dataset. As a possible improvement one could implement a model capable of classifying records as extreme or standard and then train a regressor for each category, an approach already addressed in [3]. We have implemented a first version of that model reaching a public score of 1.116 mg/m^3.
For any further information about model analysis, the relative jupyter-notebook can be accessed clicking here.

References

[1] H. Kaya, P. Tüfekci, and E. Uzun, "Predicting co and nox emissions from gas turbines: novel data and a benchmark pems," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 27, no. 6, pp. 4783–4796, 2019.

[2] S. S. Chawathe, "Explainable predictions of industrial emissions," in *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, pp. 1–7, 2021.

[3] O. Kochueva and K. Nikolskii, "Data analysis and symbolic regression models for predicting co and nox emissions from gas turbines," *Computation*, vol. 9, no. 12, p. 139, 2021.

[4] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

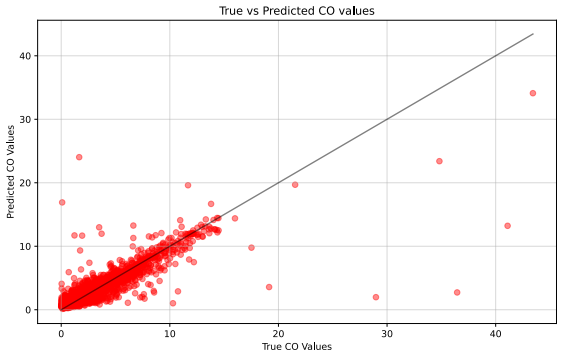

Fig. 5. The figure reports the scatter plot visualization of the true value of CO versus the predicted value. It is possible to see that the few mispredicted values have a great impact on the MSE.

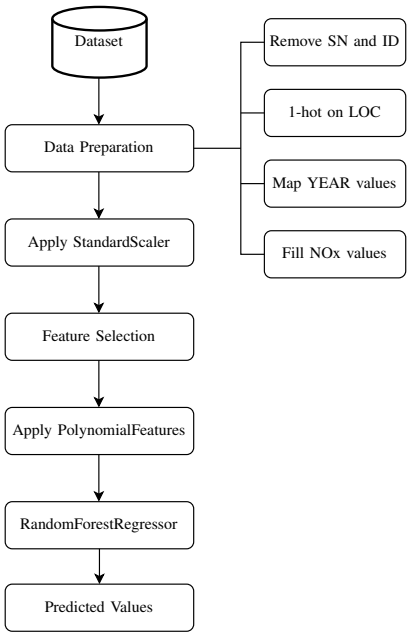

Fig. 6. The figure presents the Regression task workflow. First there is the Data Preparation part, where SN and ID features are removed, LOC feature is 1-hot encoded and NO_x missing values are filled. Then a Standard Scaler is applied to the obtained dataset, and most valuable features are selected. Then PolynomialFeatures is applied to obtain the final feature dataset. At the end the obtained dataset is forwarded to a RandomForestRegressor to obtain the CO predicted values.