Danylo Redka (14795051) Sergey Romadin (14988585) Oksana Harapyn (14711133)

Exploring and forecasting CO levels through Machine Learning Approach.

Introduction

Gas turbines are an integral part of power generation systems and therefore it is paramount to have a full understanding of their emissions in order to improve overall efficiency and mitigate any harm to the environment. This project zeroed in on a dataset that had detailed 36,733 hourly aggregated measurements that were taken from a gas turbine that was located in Turkey. The main focus revolves around determining the relationship between the parameters of the turbine and the emissions of its waste gases – specifically carbon monoxide (CO).

The dataset covers a period of five years (2011–2015) and comprises 11 very important sensor variables such as ambient temperature (AT), turbine inlet temperature (TIT) and compressor discharge pressure (CDP). Every one of these variables constitutes an important metric of the turbine in which every metric has the capability to influence the level of emission produced and energy produced by the turbine. Employing exploratory data analysis (EDA), feature extraction, and machine learning tools, the study attempts to find variations and conclusions bound to assist in strengthening the emission control technique.

Initial findings from the data exploration phase revealed several important characteristics:

- Feature distributions exhibited multimodal and skewed patterns, suggesting the presence of subpopulations or inherent variability in turbine operations.
- A strong correlation was observed among variables such as TIT, GTEP, and TEY, highlighting their interconnected roles in turbine performance.
- Low-correlation features, like ambient humidity (AH) and air filter differential pressure (AFDP), were noted as potentially valuable due to their unique contributions.

This analysis incorporates advanced visualization techniques, including correlation heatmaps and density plots, to provide a good understanding of the dataset. The insights gained from this study will help in building predictive models and optimizing turbine operations to achieve environmental compliance while maximizing efficiency.

Data Preprocessing

Since the data was presented in the chronological order, before merging the tables we added an additional column with a numeric variable - “Year”. We have also checked our data for the missing values and made sure that each column in the dataset had a proper type. The presence of null values was additionally checked as it might negatively influence some model types, such as Tree models.

Data Evaluation

a) Histogram check

Checking distribution of the features: Most of the graphs seem to be either skewed or multimodal.

Considering the density lines of the features for each year, this effect is not caused by the year difference. Therefore data may have hidden subpopulations.

b) Outliers detection and replacement

We are using IQR method for the outline detection, instead of z-score, since we have a large dataset. The IQR method is also better at handling skewed and multimodal data.

c) Correlation map analysis

A correlation matrix was developed in order to find the relationships between the variables. The following are the key observations derived from it:

i) High Correlation: High values of TEY were strongly related to TIT and CDP, reflecting that they are very important variables that relate to energy output.
ii) Ambient temperature and CO emission have very weak and low correlations, meaning both affect the dependent variables distinctly.
iii) Feature Pairwise Relationships: Scatter plots of important pairs - TIT vs. TEY, AP vs. AH - follow a straight line or a curving form which directs further feature engineering to select the appropriate models for each problem.

Regression models for CO

Dividing dataset. In protocol we are asked to use the first three years of the data as a training set and the rest two for testing. Therefore, we wouldn't be using random test split or Cross Validation.

We have a medium-sized, low-dimensional dataset. Data is continuous, has high variability. Taking this into account, we consider that the it would be reasonable to apply:

- Linear model. Standard scaling and log transformation will be applied to the data, due to numerous outliers. Feature selection applied.
- Forest models. Considering the nature of the data, we expect this model to have the highest accuracy.

Linear Regression models

First we need to scale our data. Considering evaluation of the dataset that was done previously, it is safe to say that standart scaler will be more suitable than mix-max scaling.

Feature selection: We will be using Univariate Statistics and Model-Based Feature Selection for Feature Selection.

- Univariate Statistics. Observing the graphs, 3 selected features seems to be the best option as the increase in feature would neither increase R^score nor decrease the gap between training and test sets results. Both Linear Regression model and Ridge CV can be used.
  Additionally to the R^2 scores, we can provide MSE and MAE for the best selected features: TIT, TEY, CDP. All models are coordinated in their results.
- Model-Based Feature Selection. Unlike with Univariate Statistics, the LinearRegression turned out to have the worst indicators of prediction accuracy. However other models didn't manage to reach better results with the feature selection than Univariate Statistics. On the other hand there are more Selected Features, namely: 'AT', 'AFDP', 'GTEP', 'TIT', 'TAT'. Only "TIT" got chosen twice which suggests the importance of this variable.

Forest Models

We will start with the forest models since they don't require any additional transformations or scaling of our data. Our work with them will be build in the next way:

1) Building the model without any parameter settings applied. This will help us to understand how the model is doing in general and what problems it faces.

2) Parameter tuning, including GridSearchCV in order to find the best parameters.

a) Random Forests. As can be seen from the accuracy of our data, there is the overfit. The ranges for the parameters was chosen in the following way:

- n_estimators: 100 is a good starting point; after playing with the parameters it becomes obvious that n_estimator greater than 300 results only in more memory and time consumption. Therefore range(100, 300) is taken.
- max_features: following the good rule of thumb for regression max_features=log2(n_features). Therefore 3 is taken.
- max_depth: additional feature that will decrease code running time and resolve the problem of overfitting. It was observed that the max_depth of the original tree without any restrictions was around 30. Taking max_depth <10 is pointless, since a gradient boosted regression model will have better performance with swallow trees. Therefore range(10,25) is taken.

The best accuracy of the test set that we were able to achieve is 54.4%. Therefore, for this model on average we are expected to achieve R^2 scores in range(0.5,0.6) which means that it explains some of the variance within the data for CO emissions. However, the accuracy of the training set being around 0.9 on average. Considering the gap between training and test sets R^2 scores, we may conclude that our model is still overfitting , capturing noise -> more poor generalization on the test set.

b) Gradient Boosting Regressor model. The model failed to be qualified as workable - the gap between scores is huge and the test accuracy for the test set is in range (0.1, 0.2). This was somehow expected as GBS is highly dependent on the parameter tuning. As GridSeachCV also failed to find the set of parameters' values that would make our model useful, we tried manual tuning. Lowering the learning rate seems to have the greatest success. Our model achieved the same R^2 score range as the Random Forest model, making it actually useful. But what is more important, the GBS model seems to handle the overfitting problem much better than the Random Forests model: the gap between R^2 scores of training and test sets are relatively small (around 0.1), meaning that it generalizes data for CO emissions better.

Conclusion on regression

All in all, the best prediction model turned out to be Gradient Boosting Regressor. Their indicators of accuracy tend to be higher than for linear models, providing us with a better explanation of the variance in the data, without overfitting the data as the Random Forests model does.

Clustering

Elbow Method for Optimal K

The first figure below provides the Elbow Method in order to find the optimum number of clusters K, using K-Means clustering. The plot compares the inertia-the sum of squared distances of points to their nearest cluster centroids-against the number of clusters K.

Key Observations: "elbow" is found at K=3 as, after this point, a notable drop in inertia occurs with flattening out. This infers that 3 could be an appropriate number for segmentation of the dataset.

K-Means Clustering and PCA

Left Subplot: The scatter plot represents the application of K-Means clustering with K=3. Every cluster is portrayed in a different color (for example, orange, blue). For clarity, the feature axes are also plotted-Feature 0 and Feature 1-and the clear separation of the clusters.

Explained Variance-PCA: The following graphs are followed by the PCA results, which give the explained variance ratios for the top 4 components. These values represent the amount of variance each principal component captures in the scaled dataset, with the first component contributing 47.98% and the second contributing 19.99%. After that we scatter the plot of Cumulative Explained Variance which highlights the optimal K-value (K=3) as it covers 90% of variance.

These visualizations help analyze the structure of the dataset and confirm that clustering into 3 distinct groups provides meaningful segmentation, while dimensionality reduction via PCA captures significant variability.