Combined cycle power plant performance prediction

Presented by Artem Ramus

Background

Abstract:
The dataset contains 9568 data points collected from a Combined Cycle Power Plant over 6 years (2006-2011), when the plant was set to work with full load.

Source:
Pınar Tüfekci, Çorlu Faculty of Engineering, Namık Kemal University, TR-59860 Çorlu, Tekirdağ, Turkey Email: ptufekci '@' nku.edu.tr

Heysem Kaya, Department of Computer Engineering, Boğaziçi University, TR-34342, Beşiktaş, İstanbul, Turkey Email: heysem '@' boun.edu.tr

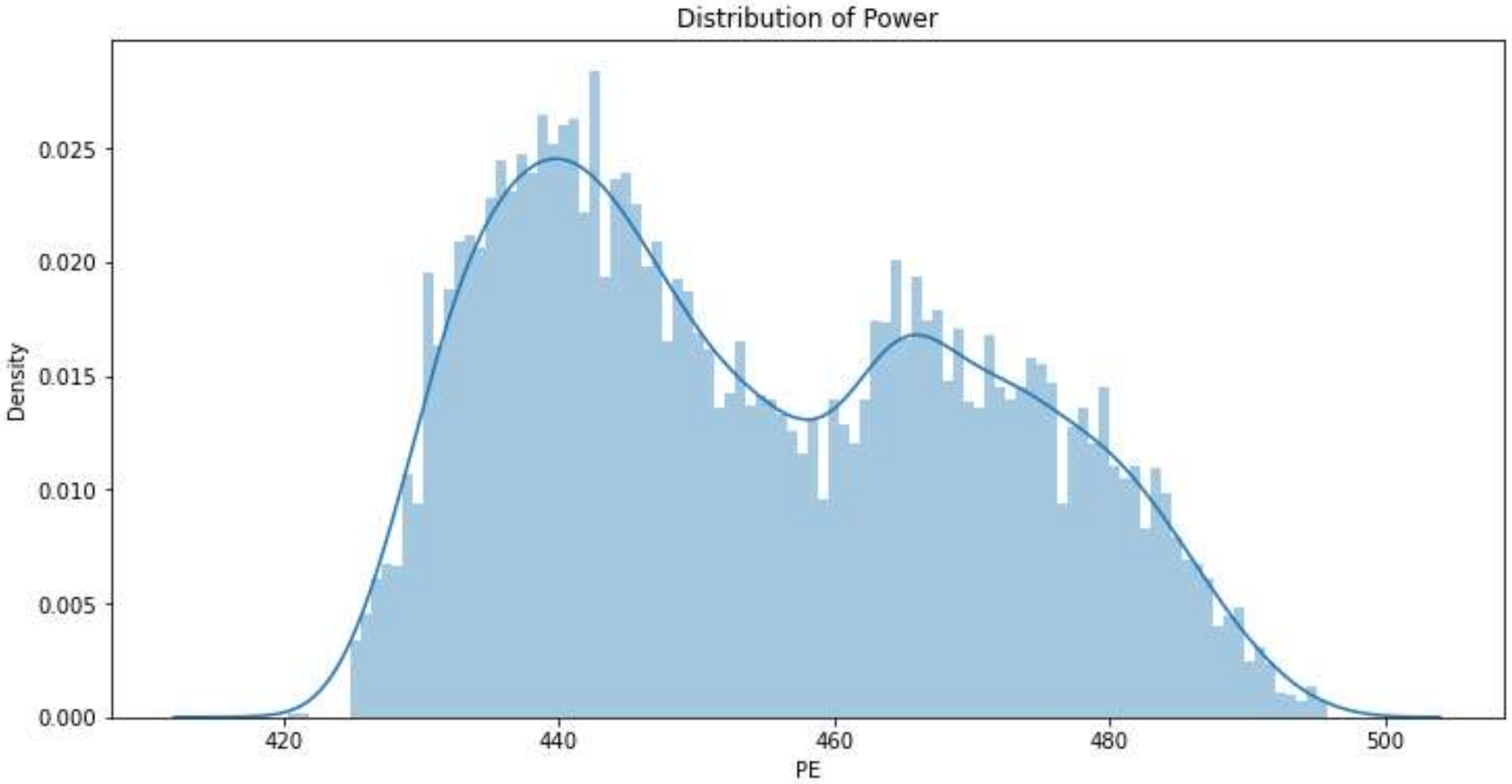
https://archive.ics.uci.edu/ml/datasets/Combined+Cycle+Power+Plant

Introduction

The data set contains 9568 data points collected from a Combined Cycle Power Plant over 6 years (2006-2011), when the power plant was set to work with full load. Features consist of hourly average ambient variables Temperature (T), Ambient Pressure (AP), Relative Humidity (RH) and Exhaust Vacuum (V) to predict the net hourly electrical energy output (EP) of the plant. A combined cycle power plant (CCPP) is composed of gas turbines (GT), steam turbines (ST) and heat recovery steam generators. In a CCPP, the electricity is generated by gas and steam turbines, which are combined in one cycle, and is transferred from one turbine to another. While the Vacuum is collected from and has effect on the Steam Turbine, he other three of the ambient variables effect the GT performance.

Data Analysis

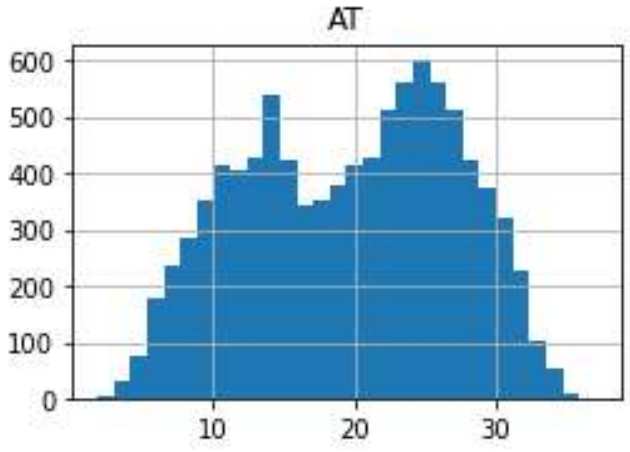
The target distribution seems to be combined from two distributions

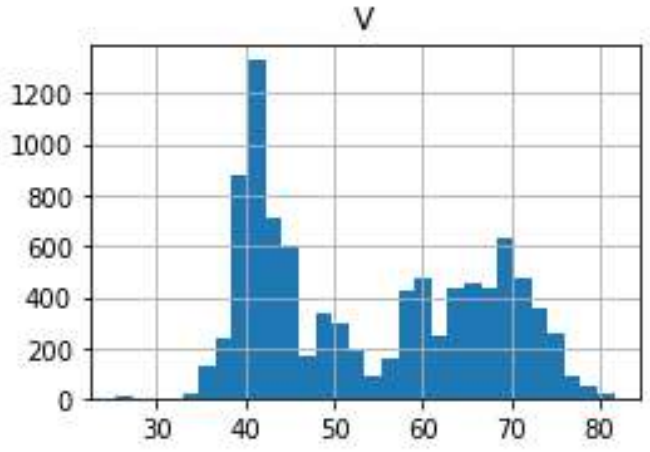


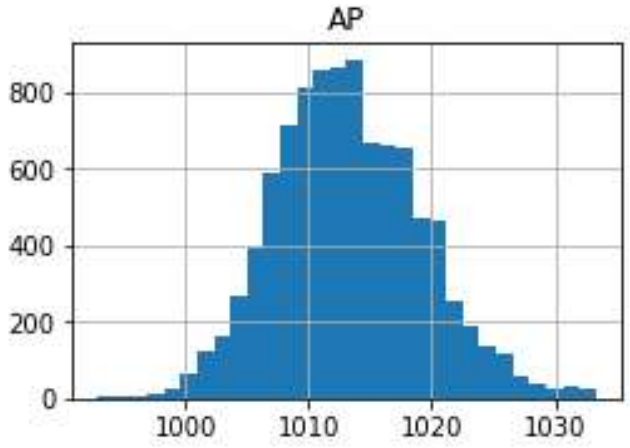
Data Analysis

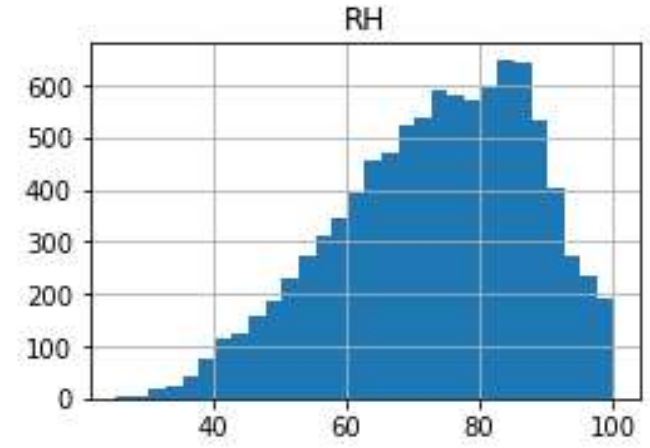
	AT	V	AP	RH	PE
count	9568.0	9568.0	9568.0	9568.0	9568.0
mean	19.7	54.3	1013.3	73.3	454.4
std	7.5	12.7	5.9	14.6	17.1
min	1.8	25.4	992.9	25.6	420.3
25%	13.5	41.7	1009.1	63.3	439.8
50%	20.3	52.1	1012.9	75.0	451.6
75%	25.7	66.5	1017.3	84.8	468.4
max	37.1	81.6	1033.3	100.2	495.8

Skewness is between -0.5 and 0.5 for all the variables, no missing values



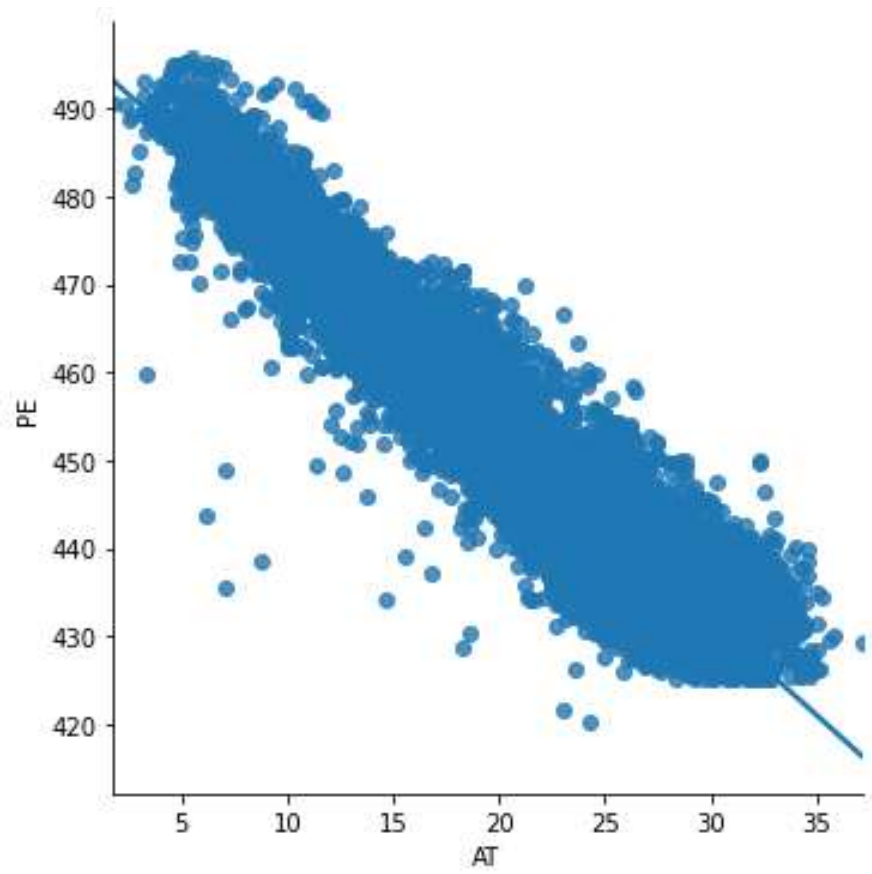


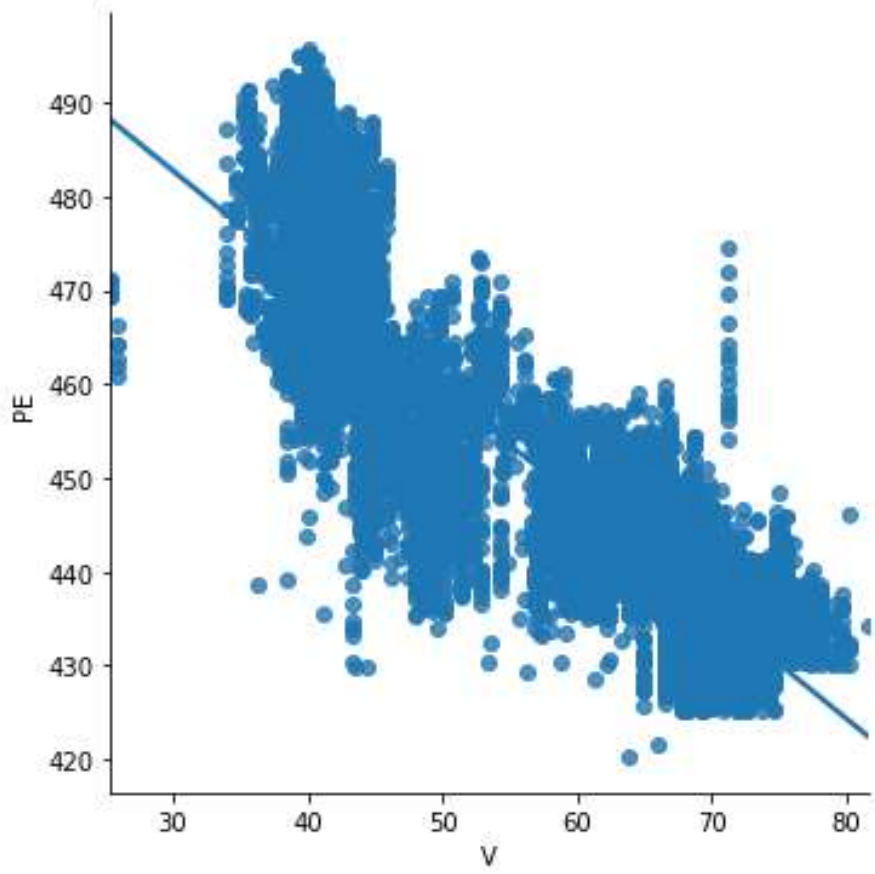




Data Analysis

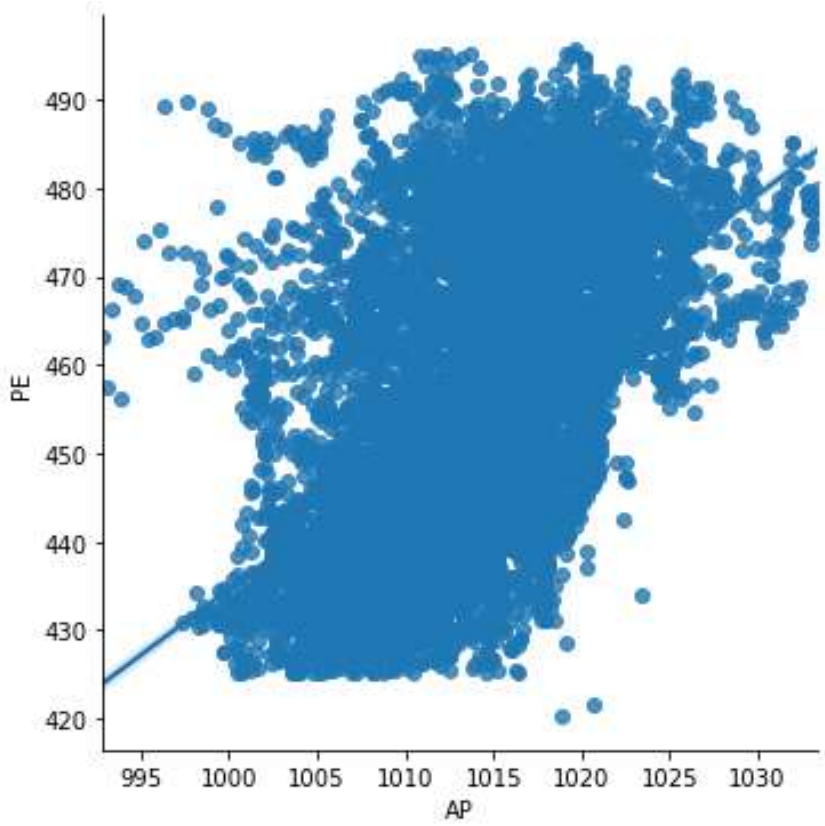
Correlations with the target:

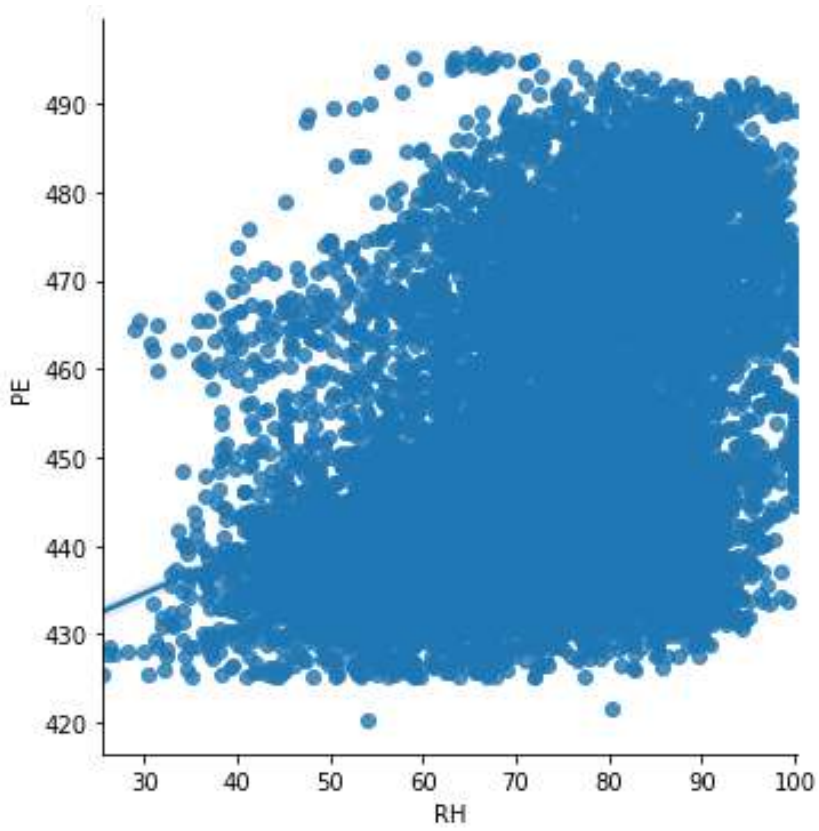




Data Analysis

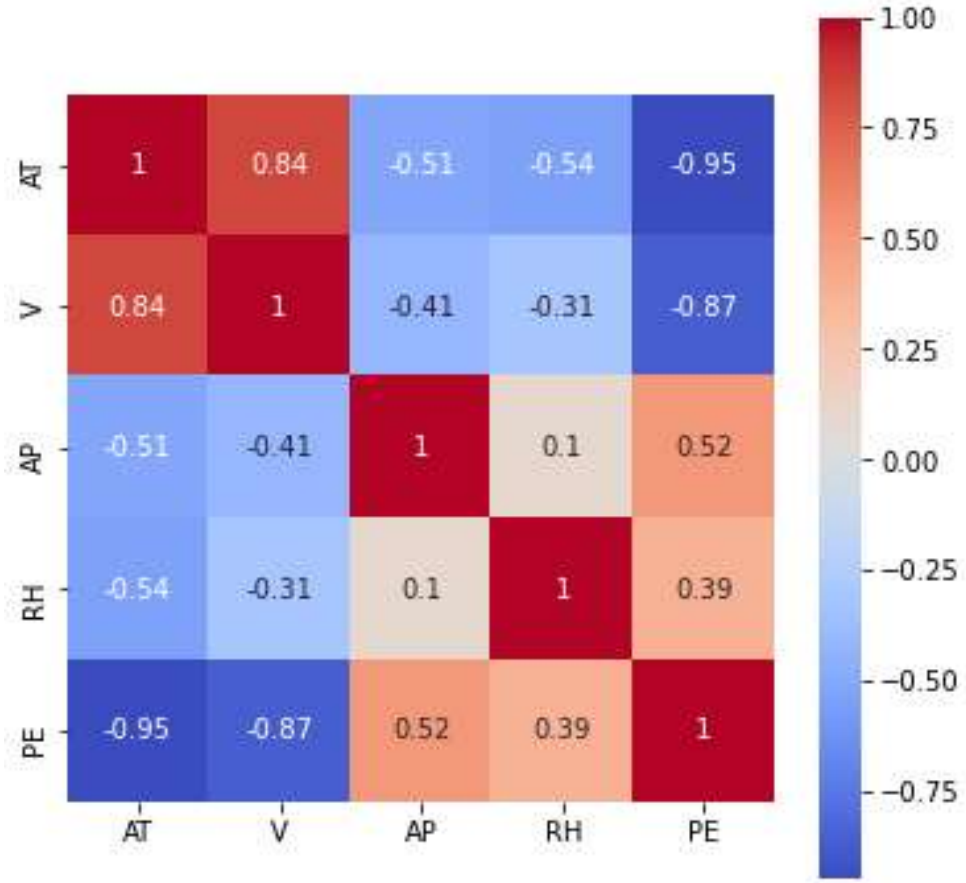
Correlations with the target:





Data Analysis

Correlation matrix:



Performance of the Model

- Performance of the following models was estimated based on Pearson coefficient:
  - Linear regression
  - Decision tree
  - Random forest
  - XGBoost
  - Artificial neural network

Summary and Conclusions

XGBoost regression gave best R square score of 96%, random forest - 95.7%.

Methodology

- The data set is checked for duplicates, null values and homogeneous features.
- Data distribution is checked with histogram, distribution and box plots for skewness and outliers.
- Correlation between features was examined with scatter plots and the correlation matrix.

The end

Thank you for your attention!