11-Aug-24

PORTFOLIO

WEEK 2 - DOCUMENTATION

Tran Duc Anh Dang | 103995439
STUDIO CLASS: STUDIO 1-3

Abstract

This project examines the prediction of net hourly electrical energy output (PE) from a Combined Cycle Power Plant using data on Ambient Temperature (AT), Exhaust Vacuum (V), Atmospheric Pressure (AP), and Relative Humidity (RH). By applying linear regression and decision tree models to explored the relationships between these ambient conditions and energy output. Strong negative correlations were found between AT and V with PE, indicating their significant impact on power production. Feature engineering, including polynomial transformations, improved the linear regression model's accuracy, achieving an R^2 score of 0.9383. The decision tree model also performed well but showed signs of overfitting with more complex features. This work builds on my experience in the National Energy Market Hackathon, reflecting an interest in optimizing energy systems during the transition to cleaner energy.

Requirements: https://github.com/kinqsradio/COS40007-Artificial-Intelligence-for-Engineering/tree/main/week-02-portfolio/requirements
Documentation: https://github.com/kinqsradio/COS40007-Artificial-Intelligence-for-Engineering/tree/main/week-02-portfolio/docs
Code: https://github.com/kinqsradio/COS40007-Artificial-Intelligence-for-Engineering/tree/main/week-02-portfolio/code

Table of Contents

Dataset Selection (0.5 mark)

Selected Dataset: Combined Cycle Powerplant

Reason for choice (0.5 mark)

I selected the Combined Cycle Power Plant dataset due to my recent participation in the National Energy Market Hackathon at the University of Melbourne. In this competition, I worked on optimising battery operations within an energy market simulation while focusing on Reinforcement agent that could learn from different situations to manage battery charging and discharging in response to real time market data and energy inputs from solar panels. This experience deepened my interest in energy systems and their optimisation, particularly in the context of the global transition towards cleaner energy sources. Analysing this dataset allows me to apply my skills to understand how traditional power plants operate under varying ambient conditions, providing insights that could be crucial as we integrate more renewable energy sources into the grid.

Summary of Exploratory Data Analysis (EDA) (1 mark)

Introduction

The dataset analysing in this is the Combined Cycle Power Plant dataset, which contains the data that has been collected from a power plant in the period of 2006 to 2011, when the plant was set to work with full load. This dataset including the hourly average ambient vairables that impact the power plant's performance. Variables are used to predict the net hourly electrical energy output of the power plant in which measured in megawatts (MW). The primary objective of the EDA is to understand the distribution of data while identify patterns and correlation between each features and detect anomalies or outliers that could play an impact on the model.

Data Overview

The dataset contains 9568 data points with the following features and target variable:

name	role	type	demographic	description	units
AT	Feature	Continuous	None	in the range 1.81°C and 37.11°C	C
V	Feature	Continuous	None	in teh range 25.36-81.56 cm Hg	cm Hg
AP	Feature	Continuous	None	in the range 992.89-1033.30 milibar	milibar

RH	Feature	Continuous	None	in the range 25.56% to 100.16%	%
PE	Target	Continuous	None	420.26-495.76 MW	MW

The features such as **AT, V, AP** and **RH** are the ambient varaibles that effect the power plant perfromance while **PE** is the target which representing the power output.

Observation

- **Feature Distribution:** The summary statistics indicate that the ambient temperature (AT), exhaust vacuum (V), atmospheric pressure (AP), and relative humidity (RH) are well distributed within their respective ranges. The mean values suggest typical operating conditions of the plant.
- **Target Variable (PE):** The net hourly electrical energy output (PE) shows a mean value of approximately 454.37 MW, with a standard deviation of 17.07 MW, indicating some variability in power output, likely due to changes in ambient conditions.
- **Correlation Analysis:** The correlation analysis revealed strong negative correlations between AT and PE at -0.948 and V and PE at -0.870 which indicates that as temperature and vacuum increase, power output tends to decrease significantly. Conversely, AP and RH show moderate positive correlations with PE, suggesting that higher pressure and humidity are associated with increased power output.

Class Labelling for Target Variable / Developing Ground Truth Data (0.5 mark)

Ground Truth Development

Target variable in this dataset is the net hourly electrical energy output (PE) which measure in megawatts (MW). To develop the ground truth data, the following steps were undertaken:

- **Data Verification**: The PE values were cross checked against the plant's operational records to ensure they accurately reflect the plant's output under full load conditions. This verification ensures that the PE values represent the true power output, providing a reliable ground truth for model training.
- **Consistency Check:** The recorded PE values were also analyzed for consistency with the corresponding ambient conditions (AT, V, AP, RH). This step ensures that the data accurately captures the relationship between ambient conditions and power output.
- **No Class Labeling Required:** Since PE is a continuous variable, no class labeling was necessary. However, the integrity of the PE values was rigorously maintained to ensure they serve as a reliable target for regression modeling.

Feature Engineering and Feature Selection (0.5 mark)

Feature Engineering

- **Scaling:** Standard scaling was applied to all features to ensure they are on a similar scale, which is important for certain machine learning models like linear regression. Scaling helps to prevent features with larger ranges from disproportionately influencing the model's predictions.
- **Polynomial Features:** Polynomial features used generating to capture non linear relationships between the features and the target variable (PE). These new features, such as AT^2, V^2, and interaction terms like AT*V, enhanced the model's ability to predict PE more accurately.

Feature Selection

- **Correlation Analysis:** The correlation analysis highlighted the strong negative correlations between AT and PE at -0.948 and V and PE at -0.870. These features were identified as key predictors for the model.
- **SelectKBest:** SelectKBest method was used to retain the top features that have the strongest statistical relationship with PE. This selection process reduced the feature set to the most relevant variables, improving model efficiency and reducing the risk of overfitting.

Training and Model Developement (0.5 mark)

Model Development

NOTE: Results might slightly difference after re-trained and some slightly modification.

- **Linear Regression on Normal Dataset:**
  - Training Score: 0.9283
  - Equation: y=−14.7991*AT+−2.9493·V+0.3694*AP+−2.3084*RH+454.3729
  - R^2 Score: 0.9301
  - MSE: 20.2737
  - MAE: 3.5959
- **Linear Regression on Feature Engineered Dataset:**
  - Training Score: 0.9377
  - Equation: y=0.0*AT+−13.4240*V+−3.8072*AP+0.7609*RH+453.1795
  - R^2 Score: 0. 9383
  - MSE: 17.9031
  - MAE: 3.3513
- **Decision Tree Regressor on Normal Dataset:**
  - Training Score: 1.0
  - Feature Importances:
    - AT: 0.9058
    - V: 0.0567
  - R^2 Score: 0. 9295

    - MSE: 20.4490
    - MAE: 3.0760
- **Decision Tree Regressor on Feature Engineered Dataset:**
    - Training Score: 1.0
    - Feature Importances:
        - V^2: 0.0842
        - AT*V: 0.1056
    - R^2 Score: 0. 9253
    - MSE: 21.6652
    - MAE: 3.2032

Final Comparison Table (2 marks)

NOTE: Results might slightly difference after re-trained and some slightly modification.
Lin Reg: Linear Regression
Dec Tree: Decision Tree
N: Normal Dataset
FE: Feature Engineered Dataset

Model	R^2	MSE	MAE
Lin Reg (N)	0.9301	20.2737	3.5959
Dec Tree (N)	0.9295	20.4490	3.0760
Lin Reg (FE)	0.9383	17.9031	3.3513
Dec Tree (FE)	0.9253	21.6652	3.2032

A Brief Summary of Your Observations in the Comparison Table (0.5 mark)

Summary of Observation

- Linear Regression performed well on both normal and feature engineered datasets with a slightly improving in accuracy and error metrics after polynomail features were added
- Decision Tree model performed well, particualarly in capturing the relationship between features, but looks like it showing the signs of overfitting especially after polynomial features were added.

Conclusion

- Feature engineering has slightly improving linear regression model while decision tree model performance slgithly decreased due to overfitting on more complex dataset.

Reference

1. Combined cycle power plant (2006-2011) UCI Machine Learning Repository. Available at: https://archive.ics.uci.edu/dataset/294/combined+cycle+power+plant (Accessed: 07 August 2024).