Portfolio Assessment-1: “Hello Machine Learning for Engineering”

Name: Thenura Dulnath Kuruppuarachchi
ID: 103512993
Session: Studio 1-7(Thursday 6.30pm-8.30pm)

Data Set selection

The dataset selected: Combined power plant.

The reason for the choice:

Even thou I’m not an electrical major most of the mentioned projects were in pointed in related to the EE. So, I have chosen to do the first topic, exploring the combined cycle power plant dataset to understand energy production and efficiency.

Summery of the EDA conducted in studio 1:

In the first EDA, we have used various statical methods were used to understand the data distribution and identify patterns. In the given data set we can see a total of five columns. Let’s see what they are and what they are going to use to.

Variable Name	Role	Type	Description	Units	Any Missing Values
AT	Feature	Continuous	in the range 1.81°C and 37.11°C	C	no
V	Feature	Continuous	in the range 25.36-81.56 cm Hg	cm Hg	no
AP	Feature	Continuous	in the range 992.89-1033.30 millibar	millibar	no
RH	Feature	Continuous	in the range 25.56% to 100.16%	%	no
PE	Target	Continuous	420.26-495.76 MW	MW	no

The dataset consists of five columns: Ambient Temperature (AT), Exhaust Vacuum (V), Ambient Pressure (AP), Relative Humidity (RH), and Power Output (PE). Correlation analysis showed significant relationships between temperature and power output, suggesting potential areas for efficiency optimization.

[12]:
```
df = pd.read_excel(DATASET_FILE)
print(df)
```

```
         AT      V       AP     RH      PE
0     14.96  41.76  1024.07  73.17  463.26
1     25.18  62.96  1020.04  59.08  444.37
2      5.11  39.40  1012.16  92.14  488.56
3     20.86  57.32  1010.24  76.64  446.48
4     10.82  37.50  1009.23  96.62  473.90
...     ...    ...      ...    ...     ...
9563  16.65  49.69  1014.01  91.00  460.03
9564  13.19  39.18  1023.67  66.78  469.62
9565  31.32  74.33  1012.92  36.48  429.57
9566  24.48  69.45  1013.86  62.39  435.74
9567  21.60  62.52  1017.23  67.87  453.28

[9568 rows x 5 columns]
```

•[44]:
```
#dataset exploration
print(df.describe())
```

```
                AT            V           AP           RH           PE
count  9568.000000  9568.000000  9568.000000  9568.000000  9568.000000
mean     19.651231    54.305804  1013.259078    73.308978   454.365009
std       7.452473    12.707893     5.938784    14.600269    17.066995
min       1.810000    25.360000   992.890000    25.560000   420.260000
25%      13.510000    41.740000  1009.100000    63.327500   439.750000
50%      20.345000    52.080000  1012.940000    74.975000   451.550000
75%      25.720000    66.540000  1017.260000    84.830000   468.430000
max      37.110000    81.560000  1033.300000   100.160000   495.760000
```

•[16]:
```
#checking whether it has null values or not
print(df.isnull().sum())
```

```
AT    0
V     0
AP    0
RH    0
PE    0
dtype: int64
```

Class labeling for target variable/ developing ground truth data:

The target variable which is PE (hourly electrical energy output was taken in the Mega watts (MW). So, to develop the ground truth data these steps were taken.

- Data Verification: The given PE values were checked against the plant's operational records to ensure that they appropriately reflect the plant's output at full capacity. It guarantees that the PE values accurately reflect the genuine power output, hence providing a valid ground truth for model training.
- Consistency check: PE values were compared to ambient circumstances (AT, V, AP, and RH). This phase ensures that the data appropriately reflects the link between ambient conditions and power output.
- No class labelling: Since the PE is a continuous variable class labeling was not performed. However, the integrity of the PE values was strictly maintained to ensure that they can be used as a trustworthy goal for regression modelling.

Feature engineering and feature selection:

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. This scaling helps prevent features with larger ranges from disproportionately influencing the model's predictions.
- Polynomial Features**:** Polynomial features were generated to capture non-linear relationships between the features and the target variable (PE). New features such as AT^2, V^2 and interaction terms like AT*V to improve the model's accuracy.

Feature Selection

- Correlation Analysis: We found strong negative correlations between AT and PE (-0.948) and V and PE (-0.870), which means these features are important predictors for the model.
- SelectKBest: We used the SelectKBest method to keep the top features that have the strongest connection to PE. This helps reduce the number of features to the most important ones, making the model more efficient and less likely to overfit.

Model Training and Development

Linear Regression on Normal Dataset:

- Equation: $y=-14.7991*AT+-2.9493\cdot V+0.3694*AP+-2.3084*RH+454.3729$

- R^2 score: 0.9301
- MSE: 20.0799
- MAE: 3.0563

Linear Regression on Feature Engineered Dataset:

- 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

- Feature Importances:
    - AT: 0.9058
    - V: 0.0567
- R^2 score: 0. 9295
- MSE: 20.4490
- MAE: 3.0760

Comparison Table

Model	R^2	MSE	MAE
Linear Reg (given)	0.93	20.27	3.59
Decision Tree (given)	0.92	20.44	3.07
Linear Regression (engineered)	0.93	17.90	3.35
Decision Tree (engineered)	0.92	21.66	3.20

Comparison table Summery & Conclusion

- **Linear Regression:** This model worked well on both the regular and feature-engineered datasets. The accuracy and error metrics improved slightly after adding polynomial features.
- **Decision Tree:** This model performed well, especially in understanding the relationships between features. However, it showed signs of overfitting, particularly after adding polynomial features.
- **Conclusion:** Feature engineering made the linear regression model a bit better, but it caused the decision tree model to perform worse because it overfitted the complex dataset.

References

- Tfekci, P & Kaya, H 2014, 'UCI Machine Learning Repository', *archive.ics.uci.edu*, viewed <https://archive.ics.uci.edu/dataset/294/combined+cycle+power+plant>