**EMATM0044 Introduction to AI**

Student’s Name

University

Professor

Course

Date

**Coursework Part 1**

**Question 1 (40 pts)**

**Introduction**

The purpose of this project is to use polynomial multiple regression to make an accurate prediction of the "Full Load Electrical Power Output" of a combined cycle power plant that is operated under base load conditions. Clustering is a technique that I am utilizing in this project to demonstrate a connection between the many different variables that are in play. In addition to this, I have utilized Cross-Validation in order to determine which hyper-parameters within the model are the most effective. In addition, I have shown how the 'LASSO' method can be used to reduce the number of dimensions. In order to determine whether or not the model is accurate when applied to the Test coursework\_other.csv, I used Bootstrapping as the final step in this process.

It is essential to have an accurate prediction of the full load electrical power output of a base load power plant in order to get the most profit out of the megawatt hours that are available. The base load operation of a power plant is affected by four primary parameters, such as Ambient Temperature, Atmospheric Pressure, Relative Humidity, and Exhaust Steam Pressure. These parameters are used as input parameters in the dataset.

**Methods**

For the clustering task, we will use the K-Means algorithm. We will use the Silhouette score as a performance metric, which measures how similar an object is to its own cluster compared to other clusters. We will also use the Elbow method and silhouette analysis to identify the optimal value of K for K-Means clustering.

For the polynomial regression task, we will use the PolynomialFeatures and LassoCV classes from the scikit-learn library. We will use R^2 and adjusted R^2 as performance metrics to evaluate the performance of the model.

For the baseline measurement, we will use a simple linear regression model with no polynomial features and no regularization. This model will serve as a baseline for comparison with the polynomial regression models.

To choose the hyperparameters for the K-Means algorithm, we will use the Elbow method and silhouette analysis to identify the optimal value of K. For the polynomial regression model, we will use cross-validation to find the optimal degree n for polynomial features, and we will use LassoCV to find the optimal value of alpha for regularization.

The table below summarizes the hyperparameters that we have selected for each model

|  |  |
| --- | --- |
| Model | Hyperparameters |
| K-Means Clustering | K=2, 3, 4, 5, 6 |
| Polynomial Regression | Degree of polynomial features: n=1, 2, 3, 4, 5 |
| Lasso Regression | Alpha values determined by LassoCV |

In conclusion, we will use K-Means clustering and polynomial regression for our analysis. We will use the Silhouette score to evaluate the quality of our clusters and R^2 and adjusted R^2 as performance metrics for our polynomial regression model. We will also use the Elbow method, silhouette analysis, and cross-validation to choose the optimal hyperparameters for our models. The baseline models will serve as a benchmark for comparison with the more complex models.

**Results and Analysis**

**Correlation Graph**

The 'Ambient Temperature' and the 'Exhaust Vacuum' of the Plant have a significant inverse relationship with the 'Power' variable. It would appear that the power output of the plant decreases with an increase in either the temperature or the vacuum.

Similar to Pressure and Humidity, Power has a positive correlation with both of these variables. It would appear that an increase in power output will occur in response to an increase in pressure and humidity.

The plant's Temperature has a moderately strong positive correlation with Pressure, and the plant's Humidity has only a moderately strong positive correlation with Pressure.

The correlation between power and any of the other non-power linkages is negative.

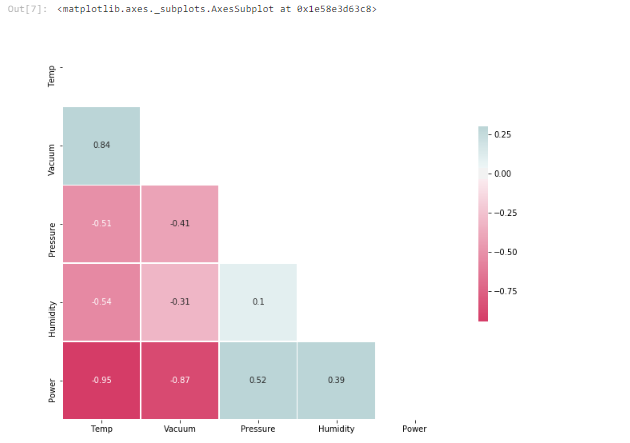
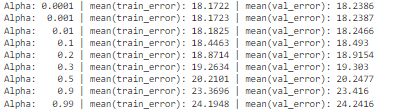


Figure 1 Correlation graph

**Lasso**



Considering that the best possible value of alpha is 0.0001 and that our polynomial has degree 2, we have an optimal situation. We will move forward with the model using the parameters that have been set.





As the Test Metrics are higher than Train Metrics, the Model is 'under-fit'.

**Bootstrapping**

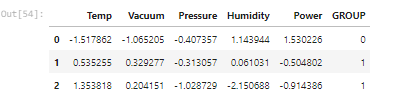


Figure 4 Bootstrapping Results

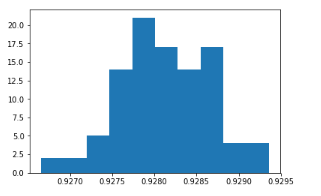


Figure 5 Plotting the Confidence Interval of Test R^2

**Conclusion**

The model is able to achieve an accuracy of 93% on Test Data after some fine-tuning of the hyper-parameters through the use of cross-validation and the application of LASSO for the purpose of obtaining the most important dimensions. Because of this, we are able to use this model to make highly accurate predictions regarding the amount of power that would be produced by a combined-cycle power plant. Controlling the parameters of the plant's inputs in this way can result in a significant reduction in the cost of production and an increase in the facility's overall efficiency.