Project Title:

Comparative Evaluation of Regression Models

Subject:

Machine Learning (CS-584-04-05)
Professor: Stephen Avsec (savsec@iit.edu)

Student Details:

Name: Gowtham Kumar Kamuni
Student ID - A20549435
Student mail - gkamuni@hawk.iit.edu

Table of Contents:

Project Overview:

In this project, we aim to conduct a comparative evaluation of five different machine learning models applied to a specific dataset. The primary objective is to determine the most effective model for the given task based on their performance metrics, particularly the R^2value.

The dataset selected for this analysis serves as the foundation upon which the models will be trained, tested, and evaluated. Each model will undergo rigorous testing and validation processes to assess its predictive capabilities and generalization performance.

To evaluate the performance of each model, we will utilize the coefficient of determination (R^2 value), a widely used metric in regression analysis. The R^2 value provides insight into the proportion of variance in the dependent variable that is explained by the independent variables. Higher R^2 values indicate better model fit and predictive accuracy.

After conducting extensive experimentation and analysis, we will compare the performance of the five models based on their respective R^2 values. The model with the highest R^2value will be deemed the most effective for the given dataset and task.

Throughout the project, we will maintain a rigorous methodology, ensuring reproducibility and reliability of results. Additionally, we will document our findings comprehensively, providing insights into the strengths and weaknesses of each model and offering recommendations for future research or applications.

By the conclusion of this project, we aim to not only identify the best-performing model but also deepen our understanding of regression modeling techniques and their applications in real-world scenarios.

We will be using the following models:

1. Multiple Linear Regression
2. Polynomial Regression
3. Support Vector Regression (SVR)
4. Decision Tree Regression
5. Random Forest Regression

Dataset Location:

The dataset used for this project:
Link: https://archive.ics.uci.edu/dataset/294/combined+cycle+power+plant

QR Code:

The Dataset can be found at **UC Irvine Machine Learning Repository**.

The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms.
The archive was created as an ftp archive in 1987 by UCI PhD student David Aha. Since that time, it has been widely used by students, educators, and researchers all over the world as a primary source of machine learning datasets.
Many people deserve thanks for making the repository a success. Foremost among them are the donors and creators of the databases and data generators. Special thanks should also go to the past librarians of the repository: David Aha, Patrick Murphy, Christopher Merz, Eamonn Keogh, Cathy Blake, Seth Hettich, David Newman, Arthur Asuncion, Moshe Lichman, Dheeru Dua, Casey Graff. The current librarians are Kolby Nottingham, Rachel Longjohn, Markelle Kelly. The current version of the web site was released in 2023. Funding support from the National Science Foundation is gratefully acknowledged.

Dataset Description:

The dataset 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.
For comparability with our baseline studies, and to allow 5x2 fold statistical tests be carried out, we provide the data shuffled five times. For each shuffling 2-fold CV is carried out and the resulting 10 measurements are used for statistical testing.
We provide the data both in .ods and in .xlsx formats.

Missing Values : NO

Variables Information:

Variables Table

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

Additional Variable Information

Features consist of hourly average ambient variables
- Temperature (T) in the range 1.81°C and 37.11°C,
- Ambient Pressure (AP) in the range 992.89-1033.30 milibar,
- Relative Humidity (RH) in the range 25.56% to 100.16%
- Exhaust Vacuum (V) in teh range 25.36-81.56 cm Hg
- Net hourly electrical energy output (EP) 420.26-495.76 MW

The averages are taken from various sensors located around the plant that record the ambient variables every second. The variables are given without normalization.

Machine Learning Models and Evaluation metrics:

Models:

1. MULTIPLE LINEAR REGRESSION

Multiple Linear Regression is a statistical algorithm used to model the relationship between multiple independent variables $X_1, X_2, \ldots, X_n$ and a single dependent variable Y. The algorithm aims to estimate the coefficients $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ that minimize the difference between the observed values of the dependent variable Y and the values predicted by the model.

The formula for multiple linear regression is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon$$

The algorithm estimates the coefficients $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ using a method such as ordinary least squares (OLS), which minimizes the sum of the squared differences between the observed and predicted values of the dependent variable.

Once the coefficients are estimated, the model can be used to predict the dependent variable Y for new observations by substituting the values of the independent variables $X_1, X_2, \ldots, X_n$ into the equation.

Multiple Linear Regression is widely used in various fields such as economics, finance, social sciences, and engineering for tasks like forecasting, prediction, and understanding the relationships between variables in complex systems.

2. POLYNOMIAL REGRESSION

Polynomial Regression is a variation of linear regression used when the relationship between the independent and dependent variables is nonlinear. Unlike simple linear regression, which fits a straight line to the data, polynomial regression fits a polynomial curve to capture more complex patterns in the data.

The algorithm models the relationship between the independent variable X and the dependent variable Y using a polynomial function of degree n:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_n X^n + \epsilon$$

The degree of the polynomial (\(n\)) determines the flexibility of the curve. Higher degrees allow the model to fit more complex patterns but may also lead to overfitting, where the model captures noise in the data rather than the underlying trend.

Polynomial regression is useful when the relationship between variables cannot be adequately captured by a straight line, such as in cases where the data exhibits curvature or nonlinear patterns. It's commonly used in fields like engineering, physics, and economics to model complex relationships and make predictions based on nonlinear data.

3. SUPPORT VECTOR REGRESSION (SVR):

Support Vector Regression (SVR) is a powerful algorithm used for regression tasks, particularly when dealing with complex and high-dimensional datasets. SVR extends the concepts of Support Vector Machines (SVM) from classification to regression.

The core idea behind SVR is to find the optimal hyperplane that best fits the data while minimizing the error or deviation of data points from the hyperplane. Unlike traditional regression methods, SVR aims to minimize the margin of error rather than fitting the data precisely.

SVR achieves this by transforming the input data into a higher-dimensional feature space using a kernel function. In this feature space, SVR identifies the hyperplane that maximizes the margin around the data points within a specified margin of tolerance (epsilon). Data points that lie within the margin or violate the margin are treated as support vectors, influencing the position and orientation of the hyperplane.

The SVR algorithm aims to find the optimal hyperplane by solving a convex optimization problem, balancing the trade-off between maximizing the margin and minimizing the error. The solution typically involves minimizing the norm of the weight vector (parameters) subject to constraints defined by the margin and the epsilon-insensitive loss function.

SVR is highly versatile and can handle nonlinear relationships between variables by using different kernel functions, such as linear, polynomial, or radial basis function (RBF) kernels. This flexibility allows SVR to capture complex patterns in the data and generalize well to unseen examples.

Overall, SVR is well-suited for regression tasks where traditional linear models may not suffice, especially in scenarios with high-dimensional data or nonlinear relationships. It's widely used in fields like finance, engineering, and bioinformatics for tasks such as stock price prediction, time series forecasting, and function approximation.

4. DECISION TREE REGRESSION:

Decision Tree Regression is a versatile algorithm used for regression tasks, capable of capturing complex relationships between input features and target variables. It builds a tree-like structure where each internal node represents a decision based on the value of a feature, and each leaf node represents the predicted output.

At each step of building the tree, Decision Tree Regression selects the feature that best splits the data into subsets, aiming to minimize the variance of the target variable within each subset. This process continues recursively until a stopping criterion is met, such as reaching a maximum tree depth or having a minimum number of samples in each leaf node.

To predict the target variable for a new instance, Decision Tree Regression traverses the tree from the root node to a leaf node based on the values of its features. The predicted output is then determined by the average (or weighted average) of the target variable values in the leaf node.

Decision Tree Regression is interpretable and easy to understand, as the resulting tree structure can be visualized and analyzed. It can handle both numerical and categorical features and automatically handles missing values and feature scaling.

However, Decision Trees are prone to overfitting, especially when the tree depth is not properly controlled or when the dataset is noisy. Techniques such as pruning, limiting the maximum tree depth, and setting minimum sample sizes in leaf nodes can help mitigate overfitting.

Despite its limitations, Decision Tree Regression is widely used in various domains due to its simplicity, flexibility, and ability to capture nonlinear relationships in the data. It serves as the basis for ensemble methods like Random Forests and Gradient Boosting, which further enhance its predictive performance.

5. RANDOM FOREST REGRESSION:

Random Forest Regression is a robust algorithm for regression tasks that combines the power of multiple decision trees to make accurate predictions. It constructs an ensemble of decision trees, each trained on a random subset of the data and features. By averaging the predictions of these trees, Random Forest Regression reduces overfitting and improves generalization performance. It's effective for handling complex relationships in data and is widely used in various fields for predictive modeling tasks.

Evaluation Metrics:

R^2 *(COEFFICIENT OF DETERMINATION) EVALUATION METRIC:*

The R^2 (pronounced "R-squared") technique, also known as the coefficient of determination, is a statistical measure that evaluates the goodness of fit of a regression model to the observed data. It represents the proportion of the variance in the dependent variable that is explained by the independent variables in the model.

The R^2 value ranges from 0 to 1, where:
- $R^2 = 0$ indicates that the model does not explain any of the variability in the dependent variable.
- $R^2 = 1$indicates that the model perfectly explains all the variability in the dependent variable.

In the context of evaluating regression models, a higher R^2 value indicates a better fit of the model to the data. However, it's essential to interpret R^2 alongside other metrics and consider factors such as the complexity of the model and the context of the data.

Expected Outcome:

The expected outcome of this project is to obtain and compare the R^2 values generated by five different regression models applied to the dataset. These R^2 values serve as quantitative measures of how well each model explains the variability in the dependent variable.

By evaluating the performance of each model using their respective R^2 values, we aim to identify the model that provides the best fit to the data. The model with the highest R^2 value will be selected as the most effective in capturing the underlying patterns and relationships within the dataset.

This outcome will not only determine the optimal model for predicting the dependent variable but also provide valuable insights into the strengths and limitations of each regression technique. Additionally, it will contribute to our understanding of how different models perform in various scenarios, aiding in future model selection and decision-making processes.

Ultimately, the selection of the best model based on the R^2 values will enable us to make informed decisions and draw reliable conclusions in our analysis, enhancing our understanding of the dataset and its underlying characteristics.

Project Outcomes and Conclusions:

The project will output the R squared values of all the 5 models used:

Results - r2 values of all the models

```
print('Multiple Linear Regression')
print('r2 value =', MLR_r2 )
```

```
Multiple Linear Regression
r2 value = 0.9325315554758247
```

```
print('Polynomial Regression')
print('r2 value =', PR_r2 )
```

```
Polynomial Regression
r2 value = 0.9067904965380815
```

```
print('Support Vector Regression - SVR')
print('r2 value =', SVR_r2 )
```

```
Support Vector Regression - SVR
r2 value = 0.9480784049986258
```

```
print('Decision Tree Regression')
print('r2 value =', DTR_r2 )
```

```
Decision Tree Regression
r2 value = 0.924737301342319
```

```
print('Random Forest Regression')
print('r2 value =', RFR_r2 )
```

```
Random Forest Regression
r2 value = 0.9615908334363876
```

Based on the outputs obtained from the analysis, it is concluded that the Random Forest Regression model yields the highest R^2 value = 0.961590 among the five models evaluated. This indicates that the Random Forest Regression model provides the best fit to the dataset compared to the other models considered.

Therefore, it is recommended to utilize the Random Forest Regression model for the project. By leveraging the strengths of Random Forests, such as robustness to overfitting, handling of nonlinear relationships, and ability to capture complex patterns in the data, we can enhance the accuracy and reliability of predictions for the dependent variable.

Implementing the Random Forest Regression model in the project will likely result in improved predictive performance and provide valuable insights into the underlying relationships within the dataset. This decision aligns with the objective of selecting the most effective model based on the R^2 value, thereby optimizing the project's outcomes and achieving the desired objectives.