To

IITD-AIA Foundation of Smart Manufacturing

Date:23-07-2023

Subject: ***Weekly Progress Report for Week-7.***

Dear Sir,

Following is the required progress report of this week dated from 17-07-2023 to 23-07-2023.

Weekly Progress:

17 July :

Topics covered:

- I have performed model generalization.
- Model generalization refers to the ability of a machine learning model to perform well on unseen or new data, beyond the training dataset.
- Achieving good generalization is crucial to ensure the model's reliability and applicability in real-world scenarios.
- The dataset is split into training and test sets using the train_test_split function from scikit-learn.
- The model is trained on the training data (X_trainand y_train), and its predictions are evaluated on both the training data and the test data (X_test and y_test).
- Good generalization is indicated by similar performance metrics (e.g., R-squared, Mean Squared Error) on both the training and test datasets.
- If the model performs well on the training data but poorly on the test data, it might be overfitting to the training data and not generalizing well to unseen data.

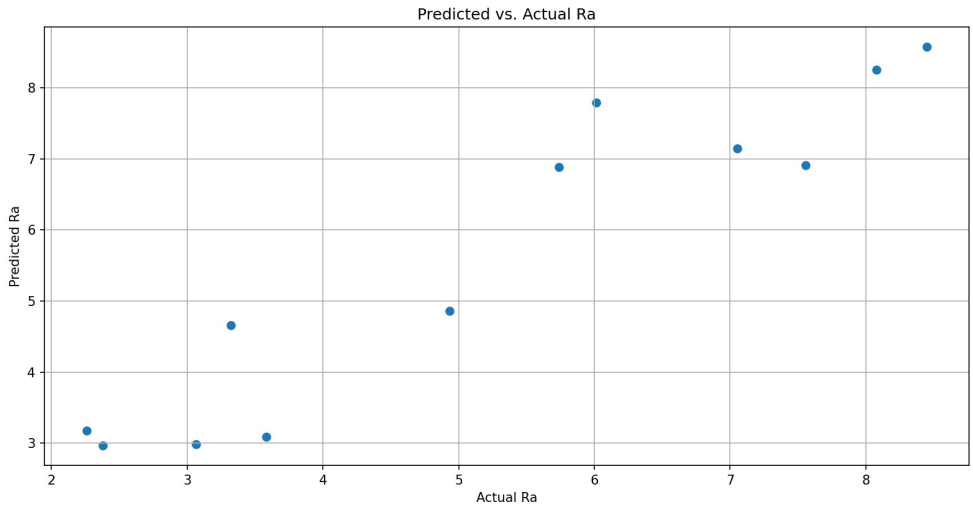
```
Model Generalization Performance:
Training Mean Squared Error: 0.20789145958944 73
Training R-squared: 0.9537100908171043
Test Mean Squared Error: 0.6829179039008436
Test R-squared: 0.8541254416847537
```

18 July:

Topics covered:

- Interpreting the results of a machine learning model involves visualizing and analyzing various aspects of the model's performance.
- We visualize the predicted versus actual values to understand how well the model performs on the test data.

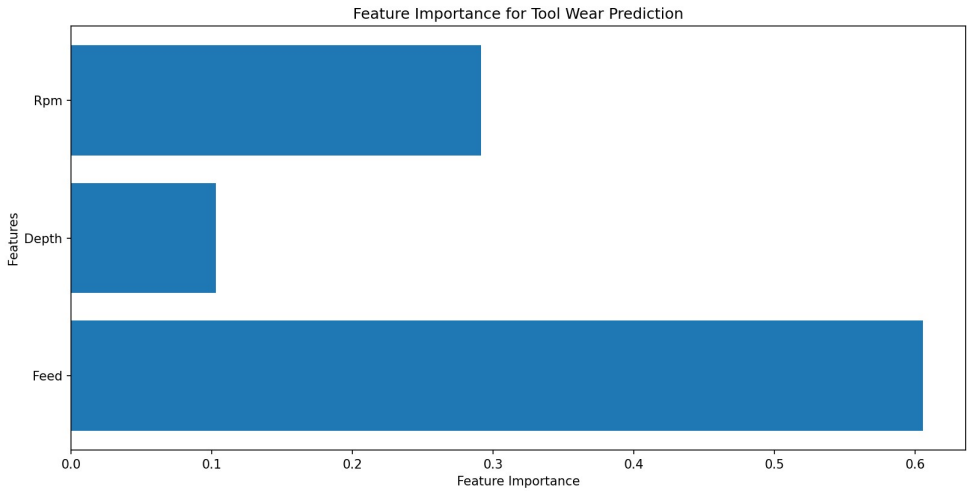
- Additionally, we plot the feature importances to identify the most influential input parameters for tool wear prediction.
- We also perform residual analysis to assess the model's prediction errors.
- Interpretation may involve further analysis based on your specific project's requirements and domain knowledge.
- It provides a starting point for understanding the model's performance and gaining insights into the relationships between input parameters and the target variable.



19 July:

Topics covered:

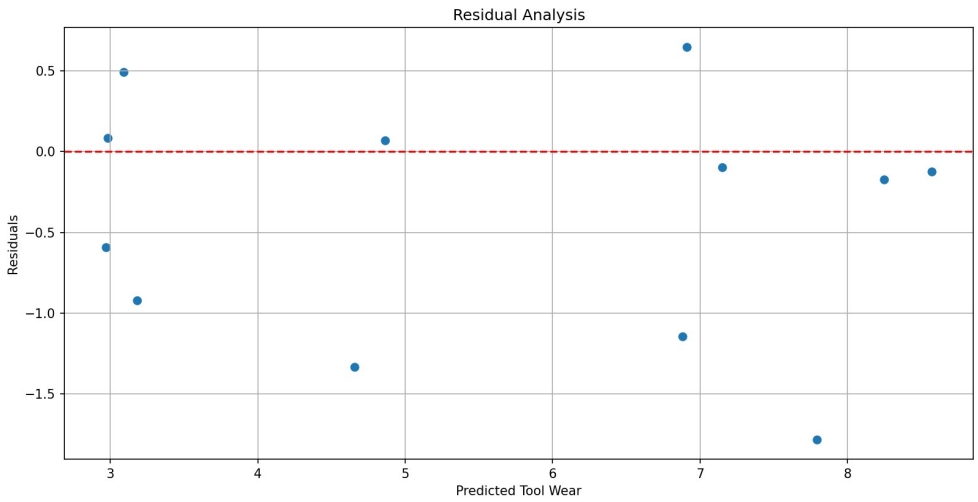
- Interpreting the results of a machine learning model involves visualizing and analyzing various aspects of the model's performance.
- Feature importance is identified to understand the relative influence or contribution of each input feature (independent variable) in predicting the target variable (dependent variable) in a machine learning model.
- It helps in answering questions like: Which features are most important in determining the target outcome? What factors have the most significant impact on the prediction?
- It's important to note that feature importance is specific to the chosen machine learning model and the dataset used for training.
- Different models may assign different levels of importance to features.
- Additionally, feature importance should be interpreted alongside domain knowledge and subject matter expertise to make informed decisions in practical applications.



20 July:

Topics covered:

- Residual analysis is a crucial step in various statistical and machine learning projects to assess the quality of a model's fit and identify potential issues or patterns in the model's residuals.
- Residuals are the differences between the observed values and the predicted values from the model.
- Before conducting residual analysis, you need to have a model already built on your data.
- This could be a linear regression, logistic regression, decision tree, neural network, or any other suitable model for your project.
- We then calculate the residuals by subtracting the predicted values from the actual observed values in your dataset.
- Create a residual plot to visualize the distribution and patterns of the residuals.
- The most common type of residual plot is a scatter plot with the predicted values on the x-axis and the residuals on the y-axis. Look for any specific patterns or trends in the residual plot.
- Check the distribution of residuals.
- A histogram or a kernel density plot can help you determine if the residuals are approximately normally distributed.
- Normality of residuals is important for many statistical assumptions.



21 July & 22 July:

Topics covered:

- Insights into Input-Output Relationships.
- In the context of predicting tool wear and surface roughness in a lathe machine, gaining insights into the input-output relationships is essential for understanding the impact of machining parameters on tool wear and surface quality.
- These insights provide valuable information for optimizing cutting conditions and improving the efficiency of the machining process.
- Analyzing the relationship between feed rate and tool wear/surface roughness can reveal the trade-off between productivity and tool life.
- Understanding how the depth of cut affects tool wear and surface roughness can help determine the optimal balance between material removal and tool life.
- Deeper cuts may lead to higher tool wear and potential vibrations, while shallower cuts may result in better surface quality but may require more passes to complete the machining.
- Examining the effect of RPM on tool wear and surface roughness can provide insights into the optimal spindle speed for a given machining operation.
- Analyzing how the input parameters influence the surface roughness can help identify the critical factors affecting surface quality.

23 July:

Topics covered:

- Identifying Outliers and Anomalies.
- Identifying outliers and anomalies is a crucial step in data analysis, as these data points can significantly affect the accuracy and reliability of our predictive models.

- Outliers are data points that deviate significantly from the rest of the data, while anomalies are data points that do not conform to the expected pattern of the majority of the data.
- some common methods for identifying outliers and anomalies in your data:
    - Visual Inspection: Plotting the data using scatter plots, box plots, or histograms can help you visually identify any data points that appear far from the main cluster of points or do not follow the general distribution pattern.
    - Z-Score Method: Calculate the z-score of each data point, which represents how many standard deviations a data point is from the mean. Data points with a z-score greater than a certain threshold (e.g., 3 or -3) can be considered outliers.
    - DBSCAN (Density-Based Spatial Clustering of Applications with Noise): A clustering algorithm that can identify outliers as data points not belonging to any cluster.
    - Isolation Forest: An algorithm that isolates outliers by recursively partitioning the data until each data point is in its own partition.
    - Local Outlier Factor (LOF): A method that calculates the density deviation of a data point compared to its neighbors, identifying outliers as points with significantly lower density.
    - Elliptic Envelope: A model that fits an ellipse around the data, identifying data points outside the ellipse as outliers.