BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Work Integrated Learning Programmes Division
M.Tech (AIML/ DSE)
Machine Learning

Assignment - 1

Objective

This Assignment is regarding the **Bike Sharing Demand Prediction Challenge**, where your goal is to predict the number of hourly bike rentals, using weather, time, and seasonal data. Through this problem, you will:

- Learn how to analyze a real-world dataset,
- Apply Linear Regression and its extensions,
- Understand nonlinearity and regularization,
- Evaluate models using a logarithmic error metric (RMSLE).

Dataset

Use the official Kaggle dataset: https://www.kaggle.com/c/bike-sharing-demand/data

(Will be later replaced with BITS Internal Link)

Files:

- train.csv – Training data with features and target
- test.csv – Test data without target (for optional submission)
- sampleSubmission.csv – Example format for predictions

Target column: count (number of bikes rented per hour)

Evaluation Metric

You will be graded primarily on RMSLE:

RMSLE = $\text{sqrt}((1/n) * \Sigma (\log(\text{pred}+1) - \log(\text{actual}+1))^2)$

This function is not available in the sklearn library. You can use the below code to compute RMLSE:

Task Breakdown

Exploratory Data Analysis (EDA) - 1 Mark

Q1. Examine dataset size, missing values, and feature types.
Q2. Visualize relationships between key features and the target variable (count).
Q3. Suggest which variables are likely to be most informative.

Feature Engineering - (Optional hints to improve performance)

Q4. You can try to derive features from datetime (hour, weekday, month, season), encode categorical variables, consider transformations to capture nonlinear trends to improve your model performance. If you do any of these, report it as answer to Q4. It is optional.

Regression Model - 2 Marks (Based on Leaderboard score/rank)

Q5. Split data into training and validation sets and build a simple Linear Regression model.

Q6. To improve model performance, you may try to:

- Extend feature space using polynomial transformations (degree 2 or 3)
- Apply Ridge and Lasso regression on polynomial features, Tune the regularization strength (α).

Model Comparison and Interpretation - 2 Marks

Q7. Summarize all results (of different models tried out) in one table (RMSLE, key observations).

Q8. Plot residuals for the best model.

Q9. Explain why the winning model performs better.

Reflection Questions (For your understanding – 0.25 bonus marks for attempting, subject to maximum assignment marks of 5)

Q10. Why does RMSLE penalize under-predictions more gently than RMSE?

Q11. What are the trade-offs between model simplicity and predictive power?

Q12. Why can't Linear Regression alone capture time-of-day effects effectively?

Submission Components

1. Report (Python Notebook): with the answers (theoretical or code) for the above 12 questions.
2. Submission.csv file in required format on test data