M2 Project

Instructions

1. You are required to submit a report in PDF format that covers all tasks/questions provided in this project document.
2. The report should include sections and subsections for better clarity and evaluation of the methods used.
3. Ensure that the code files (.py or .ipynb) are submitted along with the report. Submissions without code will not be evaluated.
4. Plagiarism will be penalized. Please provide original work.
5. Include any visualizations or tables that support your analysis and findings.
6. Write your name on the first page of the report.

Abstract

This project focuses on regression analysis using a Bike Sharing dataset. We implement and compare linear, multiple linear, and nonlinear regression models to predict bike rental demand.

1 Introduction

In this project, we analyze a Bike Sharing dataset to predict the demand for bike rentals using various regression models. The dataset contains hourly and daily counts of rental bikes in Washington, D.C., USA, over a period of two years. The features include temperature, humidity, windspeed, and weather conditions, among others. The target variable is the count of rented bikes.

2 Tasks

Task 1: Simple Linear Regression

Implement a simple linear regression model to predict the demand for bike sharing using the 'temp' feature.

(a) Load the dataset and create a scatter plot of the 'temp' variable against the 'cnt' (rental count) variable.

(b) Fit a simple linear regression model using the 'temp' variable as the predictor.

(c) Plot the regression line on the scatter plot and calculate the R-squared value.

Task 2: Multiple Linear Regression

Extend the simple linear regression model by including additional features ('atemp', 'hum', and 'windspeed').

(a) Train a multiple linear regression model on the dataset.

(b) Predict the number of bike rentals using the trained model.

(c) Calculate the Mean Squared Error (MSE) and plot the residuals.

Task 3: Nonlinear Regression

Implement a nonlinear regression model using polynomial features.

(a) Create a polynomial feature ('temp2') by squaring the 'temp' variable.

(b) Fit a nonlinear regression model using both 'temp' and 'temp2' as predictors.

(c) Plot the regression curve and compare the R-squared value with the multiple linear regression model.

3 Methodology

3.1 Dataset

The dataset used in this project is the Bike Sharing dataset, which is publicly available from the UCI Machine Learning Repository. It contains hourly and daily records of bike rentals, along with weather-related features.

- Dataset Link: `https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset`
- Features: Temperature, apparent temperature, humidity, windspeed, and weather conditions.
- Target Variable: Count of rented bikes ('cnt').

3.2 Data Preprocessing

The dataset is loaded into a Pandas DataFrame. Missing values are handled, and features are scaled as needed. Exploratory data analysis (EDA) includes visualizing feature distributions and relationships.

3.3 Model Implementation

We implemented three types of regression models:

- Simple Linear Regression: Uses 'temp' as the predictor.
- Multiple Linear Regression: Uses 'temp', 'atemp', 'hum', and 'windspeed' as predictors.
- Nonlinear Regression: Uses polynomial features ('temp' and 'temp2').

4 Results and Discussion

4.1 Simple Linear Regression

The simple linear regression model shows a strong positive relationship between temperature and bike rental demand. The R-squared value indicates the percentage of variance explained by the model.

4.2 Multiple Linear Regression

The multiple linear regression model improves upon the simple model by considering additional features. The Mean Squared Error (MSE) is used to evaluate the model's performance.

4.3 Nonlinear Regression

The nonlinear regression model captures more complex patterns in the data, leading to a higher R-squared value compared to the multiple linear regression model. The polynomial regression curve fits the data well.

5 Conclusion

In this project, we compared the performance of linear, multiple linear, and nonlinear regression models for predicting bike rental demand. The nonlinear model provided the best fit, suggesting that temperature alone is not sufficient to capture the demand patterns. Including additional features and using polynomial regression improved the prediction accuracy.

References

1. Bike Sharing Dataset: `https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset`
2. Scikit-learn Documentation: `https://scikit-learn.org/stable/`