Final Portfolio Project: End-to-End Machine Learning

Module: 5CS037 Concepts and Technologies of AI

Student Name: Prajita Banjara

Student ID: 2513644

Date: February 10, 2026

Regression Task: Seoul Bike Demand Prediction

Contents

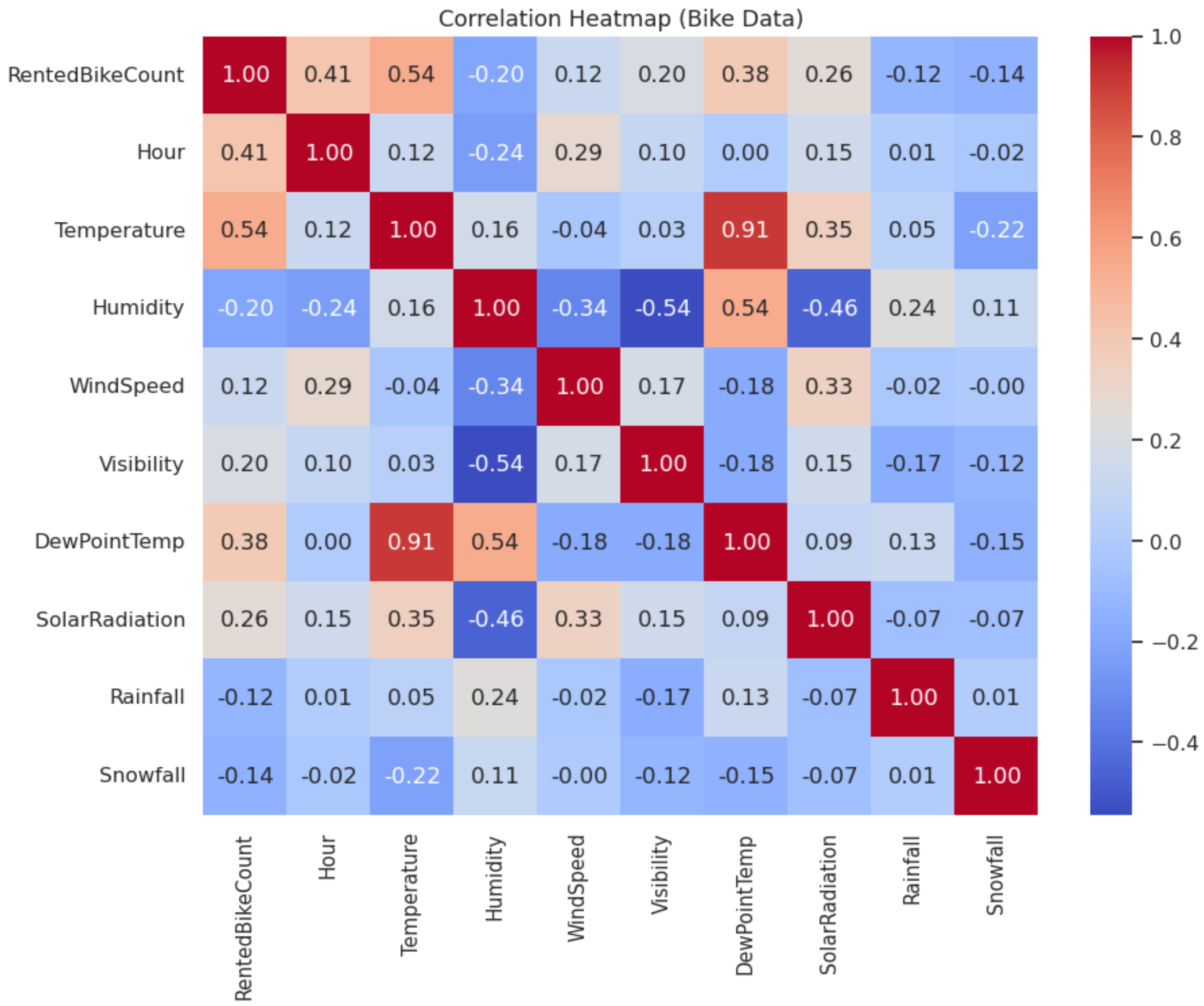
1.1 Problem & Dataset

The goal is to forecast the number of rented bikes using weather and temporal data. The dataset was obtained from the UCI Machine Learning Repository and includes features such as Temperature, Humidity, Hour, and Seasons.

1.2 Exploratory Data Analysis (EDA)

An exploratory analysis was conducted to gain insights into the data distribution.

- Target Variable: The distribution of bike rentals exhibits a positive skew.
- Correlations: The analysis of correlations indicates that Temperature and Hour are strongly positively correlated with bike rentals.
- Seasonality: The highest number of bike rentals occurs in Summer, while Winter shows the lowest numbers, highlighting the influence of weather conditions



	RentedBikeCount	Hour	Temperature	Humidity	WindSpeed	Visibility	DewPointTemp	SolarRadiation	Rainfall	Snowfall
RentedBikeCount	1.00	0.41	0.54	-0.20	0.12	0.20	0.38	0.26	-0.12	-0.14
Hour	0.41	1.00	0.12	-0.24	0.29	0.10	0.00	0.15	0.01	-0.02
Temperature	0.54	0.12	1.00	0.16	-0.04	0.03	0.91	0.35	0.05	-0.22
Humidity	-0.20	-0.24	0.16	1.00	-0.34	-0.54	0.54	-0.46	0.24	0.11
WindSpeed	0.12	0.29	-0.04	-0.34	1.00	0.17	-0.18	0.33	-0.02	-0.00
Visibility	0.20	0.10	0.03	-0.54	0.17	1.00	-0.18	0.15	-0.17	-0.12
DewPointTemp	0.38	0.00	0.91	0.54	-0.18	-0.18	1.00	0.09	0.13	-0.15
SolarRadiation	0.26	0.15	0.35	-0.46	0.33	0.15	0.09	1.00	-0.07	-0.07
Rainfall	-0.12	0.01	0.05	0.24	-0.02	-0.17	0.13	-0.07	1.00	0.01
Snowfall	-0.14	-0.02	-0.22	0.11	-0.00	-0.12	-0.15	-0.07	0.01	1.00

Figure 1: Correlation Heatmap of Bike Data Features

1.3 Data Preprocessing

- Encoding: Categorical variables such as Seasons, Holiday, and Functioning Day were converted into a numerical format using One-Hot Encoding.

- Scaling: Numerical features were standardized with the help of StandardScaler. This process was essential for the Neural Network to converge effectively, given that deep learning models are sensitive to data that isn't scaled.

- Splitting: The dataset was divided into 80% for training and 20% for testing.

1.4 Model Selection & Evaluation

Three different models were developed to assess their performance:

1. Neural Network (MLP Regressor): A Multi-Layer Perceptron consisting of two hidden layers with sizes 64 and 32.

2. Linear Regression: A traditional baseline model.

3. Random Forest Regressor: An ensemble model based on decision trees that is recognized for its ability to manage non-linear data.

Initial Results (RMSE):

[Neural Network] RMSE: 316.38 | R2 Score: 0.7598

[Linear Regression] RMSE: 440.78 | R2 Score: 0.5337

[Random Forest] RMSE: 240.32 | R2 Score: 0.8614

The Random Forest model outperformed the Linear Regression model in capturing non-linear relationships, and it required less tuning than the Neural Network.

1.5 Enhancement

Using GridSearchCV, we improved the Random Forest model to determine the optimal hyperparameters (n_estimators, max_depth), and then we used SelectKBest to choose the top ten features.

- Final Tuned R2 Score: 258.81 | R2 Score: 0.8392

2.Final Thoughts

The entire machine learning pipeline was shown in this project.

- Key Findings: Random Forest performed better in the regression task than both the Neural Network and Linear Regression models. It showed great resilience to noise and outliers in the dataset and successfully handled the combination of numerical and categorical features. Although the Neural Network model produced encouraging results, Random Forest produced the most accurate bike demand forecasts and was superior at capturing non-linear relationships. Although it worked well as a baseline, the Linear Regression model had trouble capturing the intricacies of the data.

- Difficulties: Preprocessing the categorical data (e.g., encoding seasonal data and public holidays) and making sure the Neural Network received appropriately scaled data were the main challenges in this task. For the Neural Network to correctly converge during training, extensive hyperparameter tuning was necessary, and appropriate scaling of input features was essential.

- Future Work: For future improvements, exploring deeper deep learning models like LSTM (Long Short-Term Memory) could be beneficial, especially if the dataset is expanded to include more temporal data or additional features like historical bike demand. Feature engineering could be another avenue for enhancement, such as creating features based on weather forecasts or events that might impact bike demand. Model performance could also be improved by experimenting with hyperparameter optimization methods other than GridSearchCV, such as Bayesian Optimization.

3. References

1. Seoul Bike Sharing Demand Dataset. UCI Machine Learning Repository.
2. Scikit-Learn Documentation.