Technical Documentation on Seoul Bike Sharing Demand Prediction (Regression)

Introduction

Bike sharing systems have become increasingly popular in urban areas as an eco-friendly and cost-effective mode of transportation. Accurately predicting bike demand is crucial for managing these systems effectively. This technical document aims to outline the process and methodologies used for predicting bike sharing demand in Seoul using regression techniques.

Objective

The primary objective is to predict the demand for bikes in Seoul based on various features such as weather conditions, time, day of the week, and other relevant factors. This predictive model will assist in optimizing bike allocation, inventory management, and improving overall user experience.

Dataset

The dataset used for this analysis contains historical bike usage data in Seoul, coupled with weather-related features. It includes information on:

Date & Time: Timestamps indicating when the bike was rented.

Weather Features: Temperature, humidity, wind speed, weather conditions, etc.

Usage Information: Count of rented bikes.

Data Preprocessing

Handling Missing Values: Addressing any missing or null values in the dataset through imputation or removal.

Feature Engineering: Extracting useful features from timestamps, such as hour of the day, day of the week, etc.

Normalization/Scaling: Scaling numerical features to ensure they contribute uniformly to the model.

Regression Model

Feature Selection- Identifying the most influential features using techniques like correlation analysis, feature importance, or domain knowledge.

Model Selection

- **Linear Regression**: A baseline model to understand the linear relationship between features and bike demand.

- **Decision Trees/Ensemble Methods**: Utilizing decision tree-based methods like Random Forest, Gradient Boosting, or XGBoost for capturing non-linear relationships.

- **Evaluation Metrics**: Employing metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared to evaluate model performance.

Training and Validation

- Data Splitting: Dividing the dataset into training and validation sets.
- Model Training: Training the regression models on the training data.
- Hyperparameter Tuning: Fine-tuning model parameters to optimize performance.
- Cross-Validation: Employing k-fold cross-validation to ensure the model's robustness.

Model Evaluation

Performance Metrics

- Evaluating the models using metrics like RMSE, MSE, MAE (Mean Absolute Error), and R-squared on the validation dataset.

Visualizations

- Residual Plots: Analysing the distribution of residuals to identify model biases.
- Feature Importance Plots: Understanding the impact of unique features on the prediction.

Conclusion

This technical document demonstrates the methodology used for predicting bike sharing demand in Seoul. The model with the best performance will be selected for implementation in optimizing bike allocation and improving the efficiency of the bike sharing system.

This project not only provides a predictive model but also insights into the factors influencing bike demand, aiding in better decision-making for bike sharing system management.