Enhancing Bike-Sharing Efficiency through Predictive Analytics

Zihan Liu
Sianna Fang
Yunjing Yao

1. Introduction:

In many urban centers, bike-sharing systems offer a sustainable transit option but frequently suffer from inefficiencies such as underutilization and poor station placements. This project aims to use machine learning to dynamically adjust bike allocations and station locations in response to real-time demand influenced by environmental factors like weather and time of year. By analyzing datasets from Europe, Asia, and the US, the initiative will develop predictive models to enhance the operational efficiency of bike-sharing systems globally, thereby optimizing resource use and reducing urban congestion.

2. Objective:

- To analyze historical bike-sharing usage data across multiple cities in relation to weather conditions and temporal factors.
- To analyze the data by omitting the unimportant and irrelevant or missed values. Also, some columns are combined to form new features.
- To develop a predictive model that forecasts bike-sharing demand to optimize bike distribution and station placements.
- To generalize the predictive model that can adapt to different urban settings by learning from data across several cities.

3. Data Collection:

The project will utilize the following datasets:

- **Bike Sharing Dataset from UCI**: Covering hourly and daily bike rental data along with weather and seasonal information. This can be retrieved from UCI's website, and we can import from libraries.
- **Seoul Bike Sharing Demand Dataset from UCI**: Includes factors like temperature, humidity, and precipitation. This can be retrieved from UCI's website, and we can import from libraries.
- **London Bike Sharing Dataset from Kaggle**: Contains timestamped information on bike rentals with associated weather and seasonal conditions. This can be downloaded from the kaggle dataset.

Data will be concatenated and harmonized from these cities to form a robust dataset capable of training a generalized predictive model. None of the data from the webpage will be used.

4. Methodology:

Data Preprocessing: Normalize data formats and handle missing values. Prepare and transform raw data into a clean, consolidated format suitable for analysis.

Tools and Libraries:

- Pandas: For data manipulation and aggregation.
- NumPy: For numerical operations.
- Scikit-learn: For preprocessing techniques.

Steps:

- Normalize Data Formats: Use pandas to standardize the formats across different datasets, ensuring consistency in data types and structures.
- Handle Missing Values: Identify missing values using pandas (DataFrame.isna()) and just delete them.
- Consolidate Datasets: Merge data from different sources into a single DataFrame to create a unified dataset using pandas' merge() or concat() functions.
- Create weather classifications (e.g., "sunny", "rainy") based on weather condition data. This might involve mapping specific weather condition codes to broader categories.
- Normalize/Scale features using scikit-learn's StandardScaler or MinMaxScaler to prepare for machine learning.

Data Analysis and Visualization: Identify patterns, trends, and correlations that can influence bike-sharing demand.

Tools and Libraries:

- Matplotlib and Seaborn: For creating visualizations.
- Pandas: For data exploration and manipulation.

Steps:

- Descriptive Statistics: Use pandas to calculate statistics such as mean, median, mode, min, max, and standard deviation for relevant columns.
- Distribution Analysis: Analyze the distribution of key variables using histograms and boxplots to understand their spread and identify outliers.
- Trend and Correlation Analysis: Plot time series data of bike usage to observe trends over time. Use seaborn's pointplot or matplotlib's plot function to visualize trends and seasonal patterns. Analyze correlations between variables visualized through seaborn's heatmap.

Regression Analysis: Develop a predictive model to forecast bike-sharing demand based on explanatory variables using scikit-learn.

Tools and Libraries:

- Scikit-learn: For building, training, and evaluating regression models.

Steps:

- Model Selection: Logistic Regression: Implement logistic regression as a baseline model to understand the influence of predictors on the likelihood of certain levels of bike-sharing demand. Random Forest Regression: Use Random Forest to capture non-linear relationships and feature interactions which logistic regression might miss.
- Model Training: Data Splitting: Use scikit-learn's train_test_split function to divide the data into training and testing sets, typically with a split of 80-20. Then start training.
- Model Evaluation: Evaluate both models using the Root Mean Squared Error (RMSE) and R-squared (R^2) metrics provided by scikit-learn to measure accuracy and explainability of the models.
- Model Selection: Compare the performance metrics of both models to select the one with the best accuracy and generalization performance for predicting bike-sharing demand.

5. Conclusion:

Seasonal weather changes and varying demand during peak seasons significantly impact bike-sharing systems, requiring more frequent maintenance and strategic bike redistribution. In winter, adverse conditions can damage bikes and infrastructure, while summer sees increased usage, particularly near parks and recreational areas, accelerating wear and tear. To address these challenges, data science is crucial. Predictive analytics help forecast demand and optimize bike distribution, ensuring bikes are available where and when needed. Similarly, using weather and usage data for predictive maintenance scheduling minimizes downtime and maintains system reliability. This data-driven approach not only enhances operational efficiency but also ensures the bike-sharing service adapts effectively to environmental changes and user needs.