Bike Sharing Demand Analysis in Urban Areas: Exploring the Impact of Weather and Local Events on Transportation Patterns

Nikhil Melkot (nvm35) Neil Melkot (nm1207)

1. Title and Team Information

Project Title: Bike Sharing Demand Analysis in Urban Areas: Exploring the Impact of Weather and Local Events on Transportation Patterns

Team Members:

- Neil Melkot (nm1207)
- Nikhil Melkot (nvm35)

2. Research Question and Motivation

Research Question:
How do various weather conditions and local events influence the demand for bike-sharing services in urban environments? Furthermore, can we develop a predictive model that accurately forecasts bike usage based on these influencing factors and temporal trends?

Motivation:

- **Practical Impact:** Bike-sharing systems are now a critical component of urban mobility, offering eco-friendly transportation and reducing congestion. Accurately forecasting bike demand will enable city planners and operators to optimize fleet distribution, minimize waiting times, and reduce maintenance costs by allocating resources dynamically.
- **Data-Driven Insights:** Integrating diverse datasets—such as historical usage records, comprehensive weather data, and local event schedules—allows us to uncover non-obvious patterns and relationships. This multifaceted analysis bridges the gap between theoretical research and practical application in urban planning.
- **Relevance and Context:** While past studies have often focused solely on historical usage data, our approach incorporates external variables like weather fluctuations and public events. This comprehensive analysis not only contributes to academic research but also offers actionable insights for improving urban transportation systems.

3. Data Sources

Primary Data Sets:

- **Bike Sharing Usage Data:** Use open datasets from city transportation agencies such as Capital Bikeshare (Washington, D.C.), Citi Bike (New York City), and Divvy Bikes (Chicago). These datasets include trip records, station locations, and usage statistics.
- **Weather Data:** Obtain historical weather data from sources like NOAA or OpenWeatherMap, which provide information on temperature, precipitation, humidity, wind speed, and visibility.
- **Event Data:** Collect data on local events through public APIs or web scraping from platforms such as Ticketmaster, Eventbrite, and city tourism websites. This data should cover event types, locations, expected attendance, and durations.

Data Acquisition and Preparation:

- **Obtaining the Data:** Download the datasets from public repositories and APIs. Automate data collection using Python scripts to ensure reproducibility and allow for periodic updates.
- **Cleaning and Preprocessing:** Standardize date and time formats, handle missing values through imputation or filtering, and normalize data to allow for integration. Address different data granularities (e.g., hourly usage vs. daily weather summaries) by appropriate aggregation or disaggregation.
- **Potential Challenges:** Merging data from diverse sources may lead to issues with inconsistent time formats and granularity. The unstructured nature of event data might require additional classification and processing.

4. Methodology and Analysis Plan

Data Exploration:

- Perform an extensive exploratory data analysis (EDA) using visualization libraries (e.g., Matplotlib, Seaborn) to detect trends, seasonal patterns, and outliers. This will involve generating time-series plots, histograms, and correlation heatmaps to understand the relationship among various factors.
- Generate statistical summaries (mean, median, variance) to understand data distributions and correlations. Outlier detection and anomaly analysis will be conducted to ensure data integrity.

Feature Engineering:

- Develop new features such as an "event intensity" metric, which quantifies the combined impact of multiple events in a day, and a "weather discomfort index" that aggregates adverse weather factors.
- Include temporal features like day-of-week, holiday indicators, and seasonal markers to capture cyclic usage patterns. This temporal context is expected to reveal periodic trends in bike-sharing usage.
- Incorporate lag features and moving averages to account for temporal dependencies and smooth out short-term fluctuations. These engineered features will be critical for improving the predictive power of our models.

Modeling Approach:

- Build a predictive model using regression techniques such as Linear Regression, Random Forest, and Gradient Boosting. We will test multiple algorithms to determine the best fit for our data.
- Explore clustering techniques (e.g., K-means) to segment days or stations based on usage patterns. This segmentation will help us understand localized trends and may uncover hidden patterns in the data.
- Conduct sensitivity analysis and feature importance evaluations to assess which factors most significantly influence bike-sharing demand. This step will guide the refinement of our models.

Tools and Libraries:

- **Programming Language:** Python
- **Data Manipulation:** Pandas, NumPy
- **Visualization:** Matplotlib, Seaborn
- **Modeling:** Scikit-Learn, with potential exploration of TensorFlow for advanced modeling approaches

In addition to these technical steps, our methodology will include iterative model validation and error analysis. We plan to partition the data into training and test sets and use cross-validation techniques to ensure that our models generalize well. By analyzing residuals and prediction errors, we will further refine our feature engineering and model selection processes. Detailed documentation and version control will be maintained throughout the project to ensure reproducibility.

5. Expected Outcomes and Evaluation

Anticipated Findings:

- Identification of key weather variables and event types that significantly affect bike-sharing demand, along with insights into how these factors interact.
- Development of a robust predictive model capable of forecasting bike usage with high accuracy. We expect that the model will provide actionable guidance for dynamic fleet management and service optimization.
- Discovery of temporal and spatial usage patterns that can inform strategies for optimizing bike distribution and scheduling maintenance. Our analysis may also reveal unforeseen correlations between seemingly unrelated factors.

Evaluation Criteria:

- **Model Performance:** Evaluate predictive accuracy using statistical metrics such as R-squared, Mean Squared Error (MSE), and Mean Absolute Error (MAE). Performance will be validated via cross-validation and by testing on holdout datasets.
- **Practical Relevance:** Assess the model's real-world applicability by comparing predictions against actual bike usage during similar weather conditions and events. This will involve simulated operational scenarios to test the model's recommendations.

- **Robustness and Scalability:** Perform sensitivity analysis to ensure the model maintains performance under varying conditions and over time. We will test the model on different subsets of data (e.g., different cities or time periods) to evaluate its generalizability.
- **Potential Extensions:** Consider the integration of additional data sources (such as public transit data or social media sentiment) to refine predictions further. This could open avenues for real-time forecasting and more dynamic decision-making in urban transportation management.

Furthermore, we will conduct a comprehensive error analysis to identify any systematic biases or limitations in our approach. Our evaluation process will not only focus on quantitative metrics but also on the practical implications of deploying such a model in urban settings. By aligning our outcomes with the needs of city planners and bike-sharing operators, we aim to demonstrate both the scientific and societal value of our research.