A logo of a company

Description automatically generated A close-up of a logo

Description automatically generated

**Joint Tech Internship Community Program**

**Assignment: Predicting Daily Passenger Volume on Public Transportation (Train/Bus)**

**Problem Statement:**

A public transportation authority wants to predict the daily passenger volume on trains or buses to optimize scheduling, manage crowding, and improve service delivery. Accurate predictions can help in better resource allocation and enhance passenger experience. Your task is to develop a machine learning model that predicts the daily passenger volume given various factors such as weather conditions, day of the week, holidays, and special events.

**Objective:**

Build a predictive model to estimate the daily passenger volume on public transportation (train or bus). Evaluate the model using appropriate metrics and provide insights into the factors that most influence passenger volume.

**Dataset:**

You are provided with a dataset containing the following columns:

1. **Date**: The date for which the passenger volume is being recorded.
2. **TransportType**: The type of public transportation (Train/Bus).
3. **Station/StopID**: Unique identifier for each station or stop.
4. **Location**: The location of the station or stop (e.g., Urban, Suburban, Rural).
5. **DayOfWeek**: The day of the week (e.g., Monday, Tuesday).
6. **IsHoliday**: Whether the day is a public holiday (Yes/No).
7. **WeatherCondition**: The weather conditions on that day (e.g., Clear, Rainy, Snowy).
8. **Temperature**: The average temperature on that day (in degrees Celsius).
9. **SpecialEvents**: Whether there is a special event happening on that day (Yes/No).
10. **PreviousDayPassengerVolume**: The passenger volume recorded on the previous day.
11. **PassengerVolume**: The target variable representing the number of passengers recorded on that day.

**Tasks:**

1. **Data Exploration and Preprocessing:**
   * Load the dataset and perform initial exploration to understand the data.
   * Identify and handle any missing values appropriately.
   * Convert categorical variables into numerical ones using techniques such as One-Hot Encoding or Label Encoding.
2. **Feature Engineering:**
   * Perform feature scaling (e.g., Standardization or Normalization) on continuous variables.
   * Create new features if relevant, such as the difference in passenger volume between weekdays and weekends.
   * Use dimensionality reduction techniques (e.g., PCA) to reduce the feature space if necessary.
3. **Model Building:**
   * Split the dataset into training and testing sets (e.g., 80-20 split).
   * Train different regression models (e.g., Linear Regression, Decision Trees, Random Forest, Gradient Boosting).
   * Perform hyperparameter tuning using techniques like Grid Search or Random Search to optimize model performance.
4. **Model Evaluation:**
   * Evaluate your models using appropriate metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared.
   * Compare the performance of different models and select the best one.
   * Analyze feature importance to understand the most significant factors contributing to passenger volume.
5. **Insights and Recommendations:**
   * Provide insights based on your model analysis regarding the factors that influence passenger volume.
   * Suggest actionable strategies for the transportation authority to manage crowding and optimize service delivery based on the findings.
6. **Documentation:**
   * Document your process, including data exploration, preprocessing steps, model selection, and evaluation.
   * Include visualizations where necessary to support your findings.

**Deliverables:**

* A Jupyter notebook (or Python script) with the entire workflow.
* A report summarizing your findings, including the model's performance and recommendations for optimizing public transportation services.

This assignment problem focuses on predicting daily passenger volume on public transportation, a linear regression problem in the transportation domain. The problem involves data preprocessing, feature engineering, model training, and evaluation, with a focus on understanding the factors that influence passenger volume and helping transportation authorities improve service delivery.