**Passenger Volume Prediction Project**

**Executive Summary:**

This project tackles the challenge of predicting daily passenger volume on public transportation, essential for optimizing scheduling and improving service delivery. By using machine learning models, we analyse factors such as weather, holidays, and special events to forecast passenger demand. The insights from this project will help transportation authorities manage resources efficiently, reduce crowding, and enhance the passenger experience.

**Project Objectives:**

* Accurately predict daily passenger volume on public transportation using machine learning.
* Analyse key factors like weather, holidays, and special events that influence passenger demand.
* Optimize resource allocation and scheduling to manage crowding effectively.
* Identify patterns in passenger demand across different days of the week and locations.
* Implement feature engineering to enhance model performance.
* Perform hyperparameter tuning to further improve model accuracy.

**Dataset Overview:**

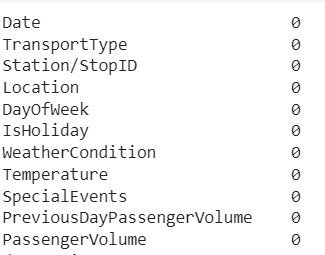
The dataset contains 5,000 records and consists of the following features related to daily passenger volume on public transportation:

|  |  |
| --- | --- |
| **Feature** | **Description** |
| Date | The date of the observation. |
| TransportType | Type of public transportation (Train/Bus). |
| Station/StopID | Unique identifier for each station or stop. |
| Location | Location type of the station/stop (Urban/Suburban/Rural). |
| DayOfWeek | Day of the week (e.g., Monday, Tuesday). |
| IsHoliday | Indicates whether the day is a public holiday (Yes/No). |
| WeatherCondition | Weather condition on the day (e.g., Clear, Rainy, Snowy). |
| Temperature | Average temperature on the day (in degrees Celsius). |
| SpecialEvents | Indicates if a special event was happening on that day (Yes/No). |
| PreviousDayPassengerVolume | Passenger volume recorded on the previous day. |
| PassengerVolume | Target variable representing the passenger volume on that day. |

**Data preprocessing steps:**

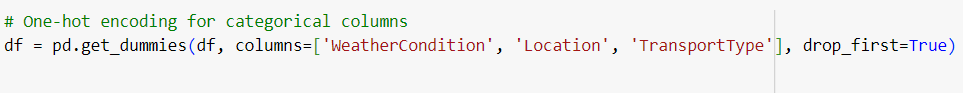
**Handling missing values:**

Identifying and filling or removing incomplete data points to maintain dataset integrity for analysis or modeling.



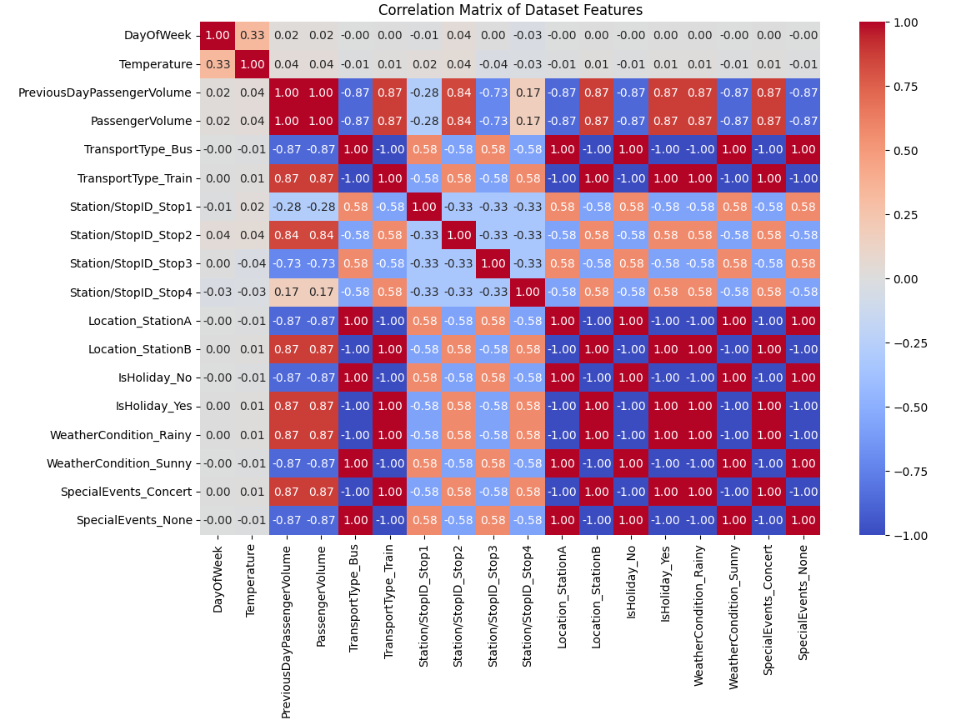
**Feature Encoding:**

Process of converting categorical data into numerical format so that machine learning algorithms can interpret and process it.

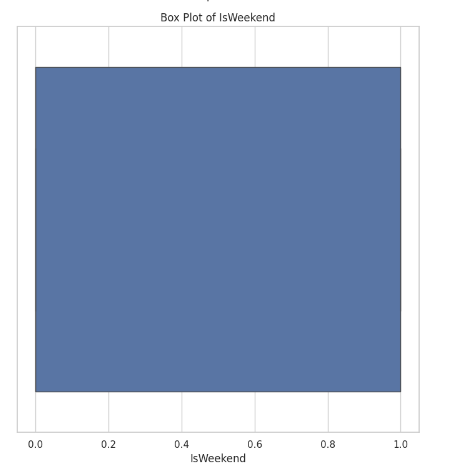
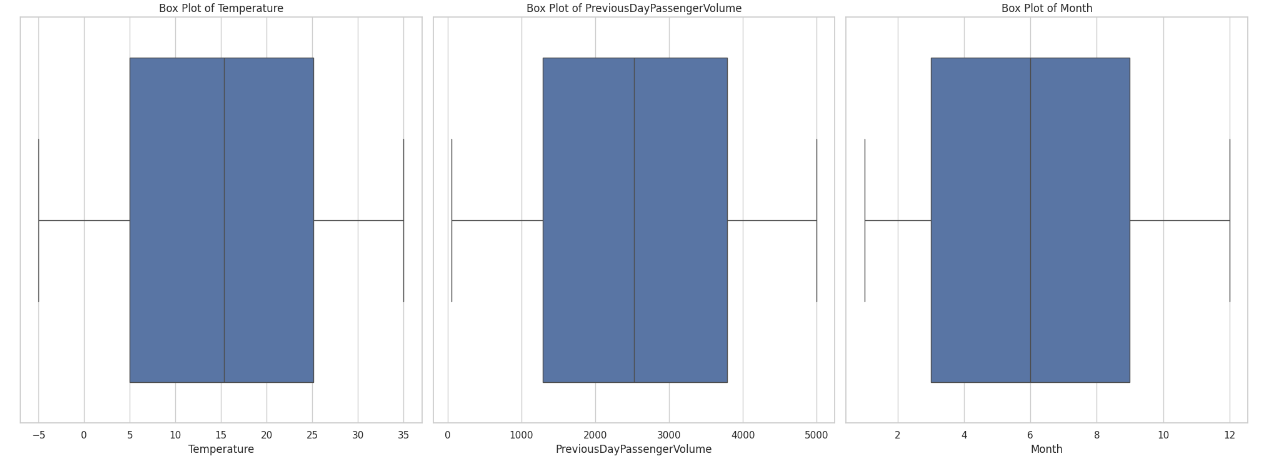


**Outlier Detection and Removal:**

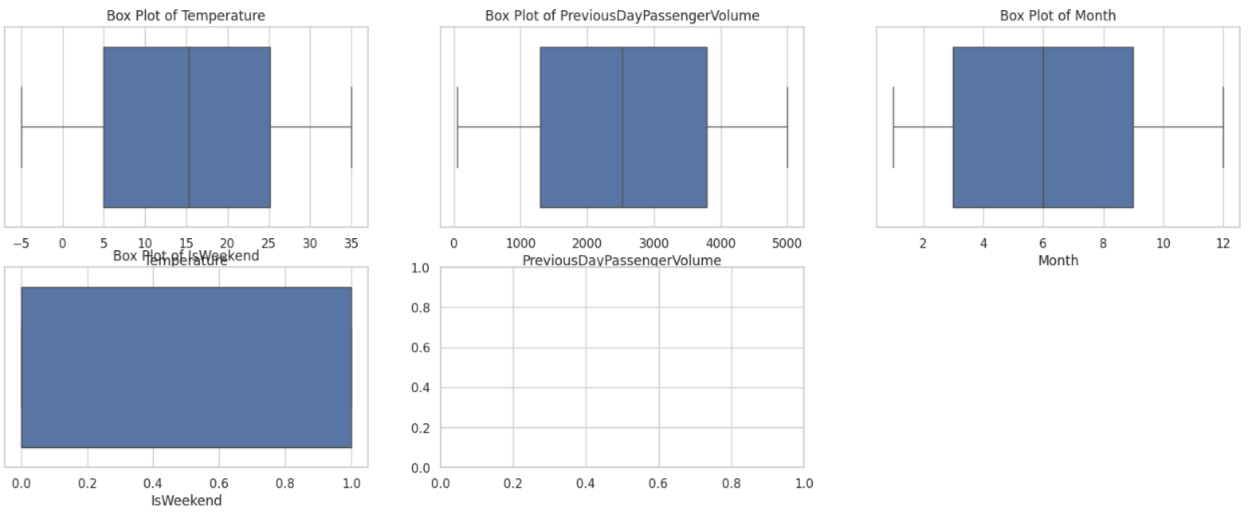
Process of Identifying and eliminating data points that significantly deviate from the normal distribution to improve model accuracy and performance.



**Before removing outlier:**



**After removing outlier:**



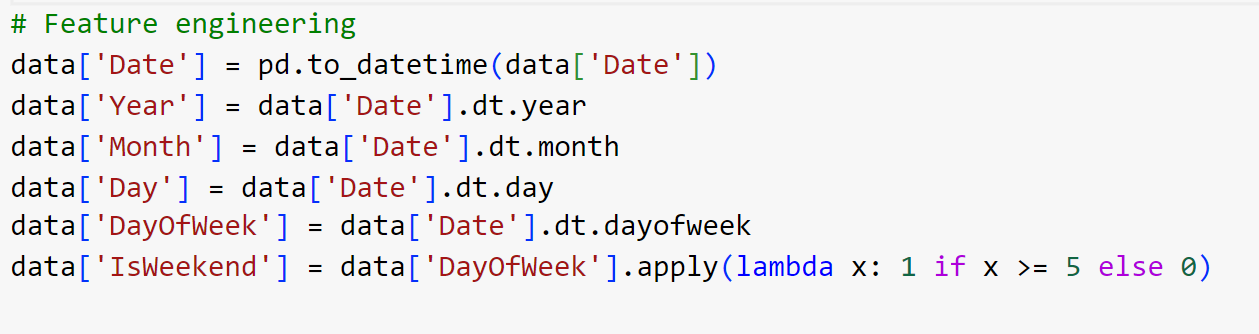
**Feature Scaling:**

Standardizes or normalizes data to ensure all features contribute equally to model performance.

**Feature Engineering:**

It involves:

* Creating new features
* Transforming features
* Handling time-based data
* Encoding categorical variables
* Aggregating data



**Model Development and Evaluation:**

A range of machine learning models were implemented to assess their effectiveness in predicting population growth:

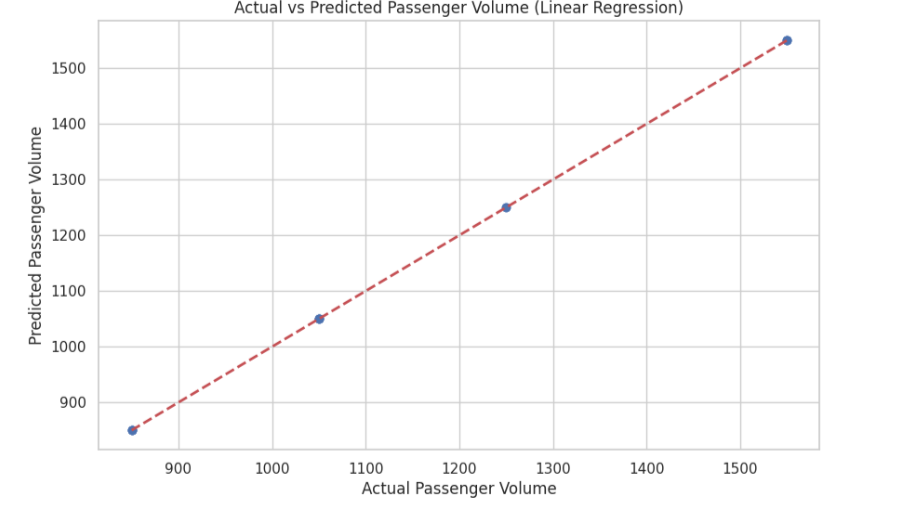
**Models Explored:**

1. Linear Regression
2. Decision Tree Regressor
3. Random Forest Regressor
4. XGBoost Regressor

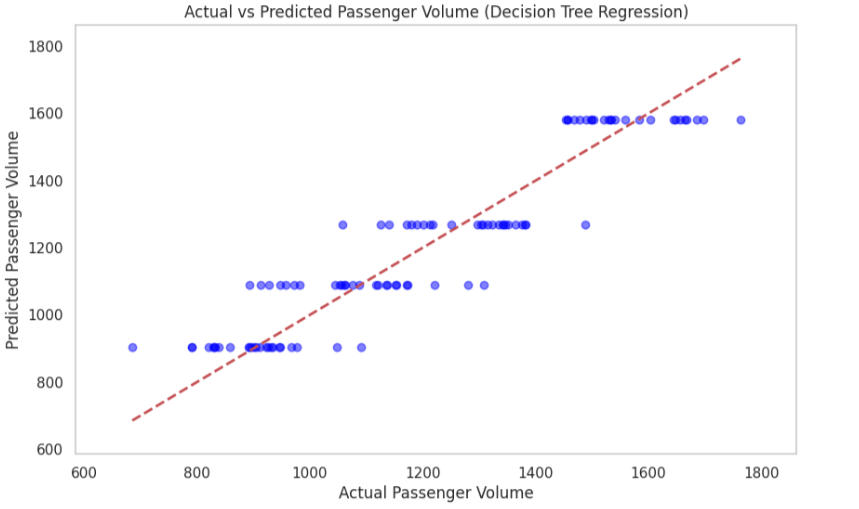
**Model Performance Summary:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Mean Squared Error (MSE)** | **Root Mean Squared Error (RMSE)** | **R-squared (R²)** |
| Linear Regression | 3.8776 | 3.4106 | 1.0 |
| Decision Tree | 7488.9766 | 71.6428 | 0.8767 |
| Random Forest | 9120.0797 | 82.0923 | 0.9114 |
| XGBoost | 0.0195 | 0.0653 | 0.999 |

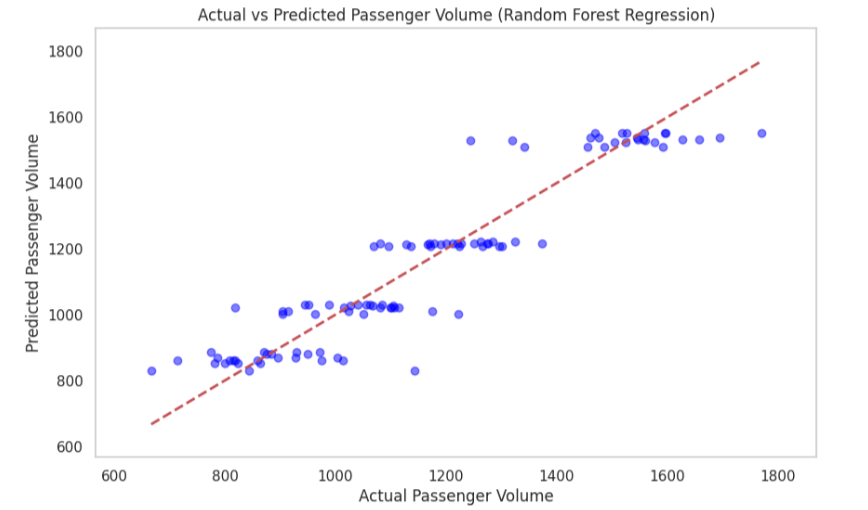
**Graphs:**



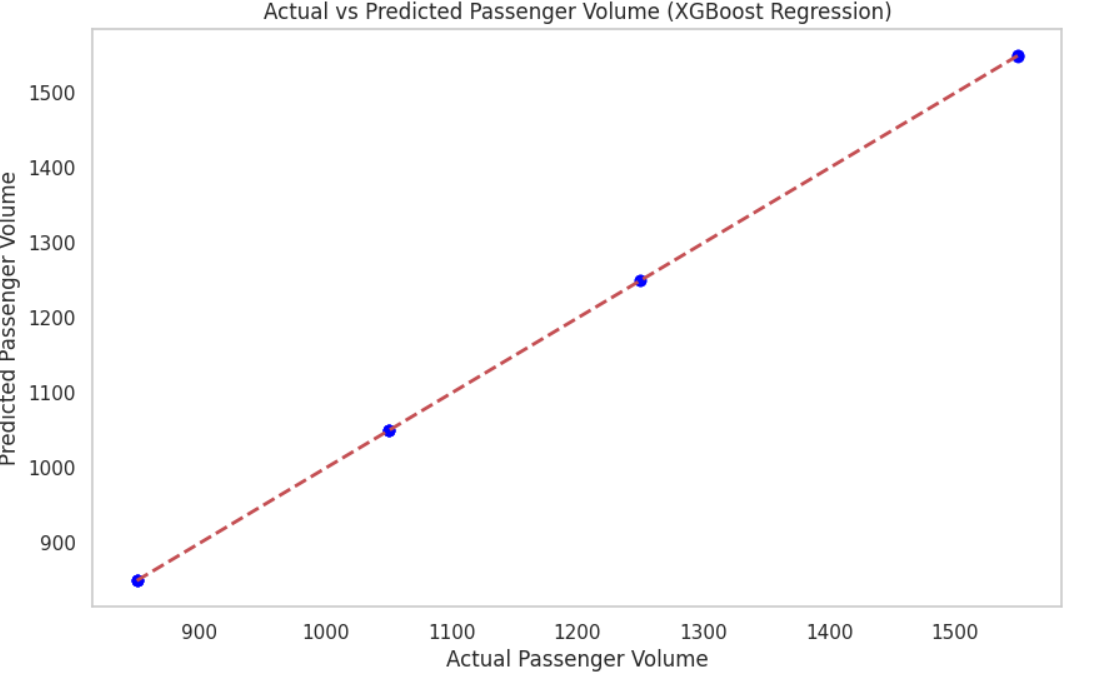
**Linear Regression**



**Decision Tree**



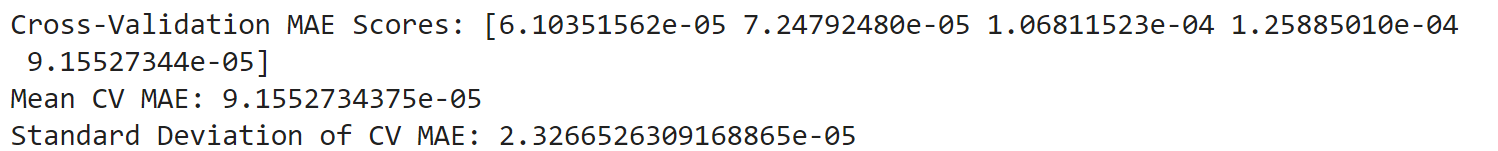
**Random Forest**

  **XGBoost**

* **Top Performer**: The **XGBoost** model outperformed all others in terms of MSE and RMSE, indicating a suitable fit for the dataset.
* **Underperforming Model**: The **Random Forest** model yielded a very large MSE, suggesting inadequate predictive power.
* **Linear Regression** demonstrated competitive results but fell short compared to XGBoost.

**Cross validation:**

statistical method used to assess the generalizability and performance of a model by dividing the dataset into multiple subsets, training the model on some subsets and validating it on others.



**Strategic Recommendations:**

1. **Model Selection**: Given its performance, XGBoost is recommended for predicting population growth.
2. **Further Enhancements**:
   1. Explore hyperparameter tuning for existing models to optimize performance.
   2. Investigate additional socio-economic features or data transformations.
   3. Consider advanced ensemble methods for improved predictive accuracy.

**Conclusion:**

This project demonstrates the potential of machine learning in predicting daily passenger volume on public transportation. Our model provides accurate forecasts, helping transportation authorities optimize scheduling, manage crowding, and improve service delivery.

**Future Work:**

1. Explore More Algorithms: Test advanced models like Gradient Boosting and Neural Networks for enhanced accuracy.
2. Refine Model Tuning: Apply more sophisticated hyperparameter tuning methods like Random Search to boost performance.
3. Expand Feature Set: Incorporate additional features like fuel prices and traffic conditions to further improve predictions.
4. Long-Term Trends: Analyse historical data to understand seasonal trends and their impact on passenger volume.
5. Develop a Dashboard: Create an interactive tool for real-time forecasting and actionable insights for transportation authorities.

**Code Repository:**

The complete implementation is available in the accompanying Colab Notebook or Python script, accessible via the project's GitHub repository.