Final Project Report

Name: Vishva Chaudhary
Domain: Data Science
Date of submission: 10/06/2024

I. Overview

- Project Statement

The project focuses on analyzing and **predicting traffic patterns in a smart city** using data science techniques. The goal is to develop a model that can accurately predict traffic congestion at various junctions in the city, which can then be used to improve traffic management and reduce congestion.

- Need for the Project

With the increasing number of vehicles and urbanization, managing traffic efficiently has become a critical issue for city planners. By leveraging data science and machine learning, we can predict traffic patterns and help in optimizing traffic flow, reducing congestion, and improving the overall efficiency of urban transportation systems. This project aims to contribute to the development of smarter cities by providing actionable insights through data analysis.

II. Dataset Description

The dataset used in this project comprises traffic data collected from four junctions in a smart city. The data includes:

- Datetime: The timestamp of the recorded traffic data.
- Junction: The identifier for the junction where the data was recorded.
- Vehicles: The number of vehicles recorded at each junction at each timestamp.
-ID: The ID of every vehicle passe by the junction is given.

Additional features were derived from the datetime column, including the day, month, year, and hour, to better understand and model the traffic patterns.

III. Process Overview

Data Preprocessing

1. Missing Values: Checked for any missing or null values in the dataset and handled them appropriately.
2. Datetime Transformation: Extracted additional features such as date, day of the week, month, year, and hour from the datetime column.
3. Data Cleaning: Ensured the data was clean and free from any inconsistencies.

Exploratory Data Analysis (EDA)

1. Visualizations:
   - Histogram: To understand the distribution of vehicles.
   - Time-Series Plot: To observe traffic trends over time.
   - Count Plot: To see the number of vehicles at different times of the day and different days of the week.
   - Scatter Plot: To identify any correlations between variables.

2. Observations:
   - Traffic patterns vary significantly at different times of the day and days of the week.
   - Peak traffic hours were identified, which can help in better traffic management.
   - Each junction has unique traffic patterns, requiring junction-specific analysis which is done with the help of histogram.

Model Training

1. Linear Regression: Initially used for modeling, but the accuracy was not satisfactory.
2. LGBM (Light Gradient Boosting Machine): Improved accuracy over linear regression.
3. Random Forest: Provided better accuracy and robustness compared to linear regression.

Model Evaluation

- Metrics: Evaluated the models using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2) to determine the accuracy and performance of the models.

The accuracy for Linear Regression is:

```
Mean Squared Error: 201.73482504438113
Mean Absolute Error: 10.300469987361438
R2 Score: 0.5422434446758574
```

The accuracy for LGBM is:

```
Mean Squared Error: 97.07228093821986
Mean Absolute Error: 6.448586650344218
R2 Score: 0.7797332566156524
```

The accuracy for Random Forest is:

```
Mean Squared Error: 94.98263740355833
Mean Absolute Error: 6.326203056741734
R2 Score: 0.7844748674211819
```

- Comparison: Compared the performance of different models to select the best one for traffic prediction, almost LGBM and Random Forest have significantly same accuracy.

IV. Achievements

1. Skill Development:

- Enhanced skills in data preprocessing, EDA, and model training.
- Learned Flask for web application development.
- Improved communication, and public speaking skills through soft skill videos provided by upskill.

2. Project Contributions:

- Thoroughly explored and understood the project requirements.
- Evaluated and fine-tuned the model for better performance.
- Ensured the project was well-documented and all necessary changes were implemented.

V. Challenges

1. Time Constraints: Faced challenges in managing time due to external exams.
2. Complexity: The complexity of the data science project required careful analysis and iterative model training to achieve satisfactory results.

VI. Learning Resources

1. USC_TIA Documentation:
  - Utilized official documentation for reference and troubleshooting.
  - Attended webinars and online tutorials and material provided by upskill to enhance understanding.

2. Data Science Learning Resources:
  - Engaged with platforms like Kaggle, Medium, and YouTube to strengthen Python and machine learning skills.
  - Referred to "Introduction to Probability and Statistics" and “Introduction To Machine Learning” for foundational knowledge.

VIII. Additional Comments

The guidance from both UNICONVERGE and UPSKILL CAMPUS was instrumental in understanding the platform and the company, helping me get started with the internship. As of today's date, 10/06/2024, I have developed a smart city traffic pattern prediction model based on the given dataset. The complete code and report for the project are available on GitHub.

IX. GitHub Repository

The complete code and report for the project can be accessed at the following GitHub link: https://github.com/vish3101/upskill_campus

---