Bus Demand Prediction Project

Introduction

This project aims to predict the number of buses required at different times of the day for the route between Gandhipuram and Somanur in Coimbatore, considering factors such as population, holidays, and peak hours. The goal is to predict the demand for buses in a way that prevents overcrowding, particularly during peak hours and around public holidays.

Project Structure

The project is organized into several key components:

- **/data**: Contains input data files such as demographics, bus schedules, and holidays.
- **/src**: Contains Python scripts for data processing, feature engineering, model training, and prediction.
- **/models**: Contains trained machine learning models.
- **/visualizations**: Stores visualizations generated during the project.

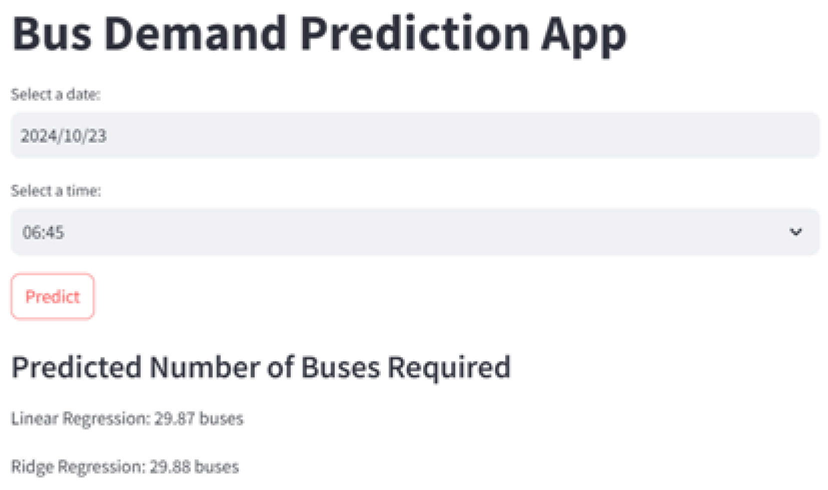
Assumptions

The following assumptions were made in this project:

1. A bus has a capacity of up to 45 sitting and 15 standing passengers.
2. Around 30% of the school population consists of hostelers who use public transport before and after holidays.
3. 45% of the school/college students use private transport.
4. 50% of the working population use public transport on a daily basis.
5. An increased number of buses is required 2 days before and after a public holiday.

Bus Demand Prediction Interface

The interface allows users to select a date and time to predict the number of buses required, considering various factors such as holidays and population.



Bus Demand Prediction App

Select a date:

2024/10/23

Select a time:

06:45

Predicted Number of Buses Required

Linear Regression: 29.87 buses

Ridge Regression: 29.88 buses

Workflow

The workflow for this project involves the following steps:

1. **Data Preprocessing**: The script `data_preprocessing.py` cleans and processes the raw data, handling missing values and normalizing the features.
2. **Feature Engineering**: The script `feature_engineering.py` generates new features from the existing data, such as holiday effects and time-based features.
3. **Model Training**: The script `model_training.py` trains multiple machine learning models (Linear Regression, Polynomial Regression, and Ridge Regression) on the processed data.
4. **Prediction**: Once trained, the models are used to predict the number of buses required at different times of the day, considering factors such as peak hours and holidays.

File Descriptions

- **demographics.csv**: Contains population data for schools, colleges, and workplaces, categorized by ward number.
- **bus_schedules.csv**: Contains bus schedule information, including bus numbers, start and end times, and passenger counts.
- **holiday.csv**: Contains the list of holidays to account for public transport variations during these days.
- **feature_engineered_bus_demand.csv**: Contains engineered features such as the total number of passengers, required buses, and holiday effects for specific time intervals.

Model Descriptions

Three models were trained in this project:

- **Linear Regression**: A simple regression model used to predict the number of buses based on input features.
- **Polynomial Regression**: A more complex model that captures non-linear relationships in the data, providing more accurate predictions during peak and off-peak hours.
- **Ridge Regression**: A regularized regression model that prevents overfitting by penalizing large coefficients. It was found to be the best-performing model in this project.

Results

The model evaluation results showed that Ridge Regression performed best in predicting the bus demand, with a high R^2 value, indicating a good fit to the data. The application interface provides real-time predictions for the number of buses required at different times of the day.