Traffic Prediction

Anushka Singh Hari Shubha Mohanshi Jain Pari Sharma
Aastha Bhore Pranaya Jayaprakash

IIT Jodhpur
{b22ai008, b22ai021, b22es014, b22cs039, b22ee002, b22ee086}@iitj.ac.in

Abstract

Traffic congestion is rising in cities around the world driven by factors such as growing urban populations, aging infrastructure, poorly synchronized traffic signals and a lack of real-time data.The impacts are significant. Traffic data and analytics company INRIX estimates that traffic congestion cost U.S. commuters 305 billion in 2017 due to wasted fuel, lost time and the increased cost of transporting goods through congested areas. Given the physical and financial limitations around building additional roads, cities must use new strategies and technologies to improve traffic conditions. One key approach is 'Traffic Prediction'.

Traffic prediction means forecasting the volume and density of traffic flow to manage vehicle movement, reduce congestion, and find the optimal (least time or energy consuming) route. The task of traffic prediction is to detect traffic conditions for upcoming periods such as next day, week etc. Utilizing historical traffic data, time-of-day patterns, and other relevant factors, predictive models can anticipate traffic congestion hot spots and enable the timely deployment of resources to mitigate its impact.

Contents

1 Introduction

Traffic congestion is a prevalent problem in urban areas worldwide, imposing significant economic, environmental, and social costs. The inefficiencies caused by congested road networks lead to wasted time, increased fuel consumption, excessive air pollution and decreased quality of life for residents.

One promising approach to tackle traffic congestion is through the use of predictive modeling techniques. By using historical traffic data, time-of-day patterns and other relevant factors it is possible to develop models capable of forecasting future traffic conditions. These predictive models enable transportation authorities to adjust traffic signal timings, reroute traffic and implement other interventions to alleviate congestion and improve overall transportation efficiency.

This project aims to explore the effectiveness of various machine learning algorithms in predicting traffic patterns. Specifically, we investigate the performance of Linear Regression, K-Nearest Neighbors (KNN), Decision Tree, Gradient Boosting and Random Forest algorithms in forecasting traffic congestion

levels. By comparing the strengths and weaknesses of these different approaches, we aim to identify the most suitable algorithm for traffic prediction tasks.

The findings of this project have implications for urban transportation planning and management. By accurately predicting traffic conditions, cities can implement proactive measures to mitigate congestion and enhance the overall mobility experience. Moreover, the development of reliable traffic prediction models lays the foundation for the future implementation of intelligent transportation systems.

In this, we provide a detailed analysis of our experimental methodology, results, and conclusions. By sharing our insights and findings, we aim to contribute to the growing body of knowledge in the field of traffic prediction and support efforts to build smarter and more efficient transportation systems for urban environments.

1.1 Figures

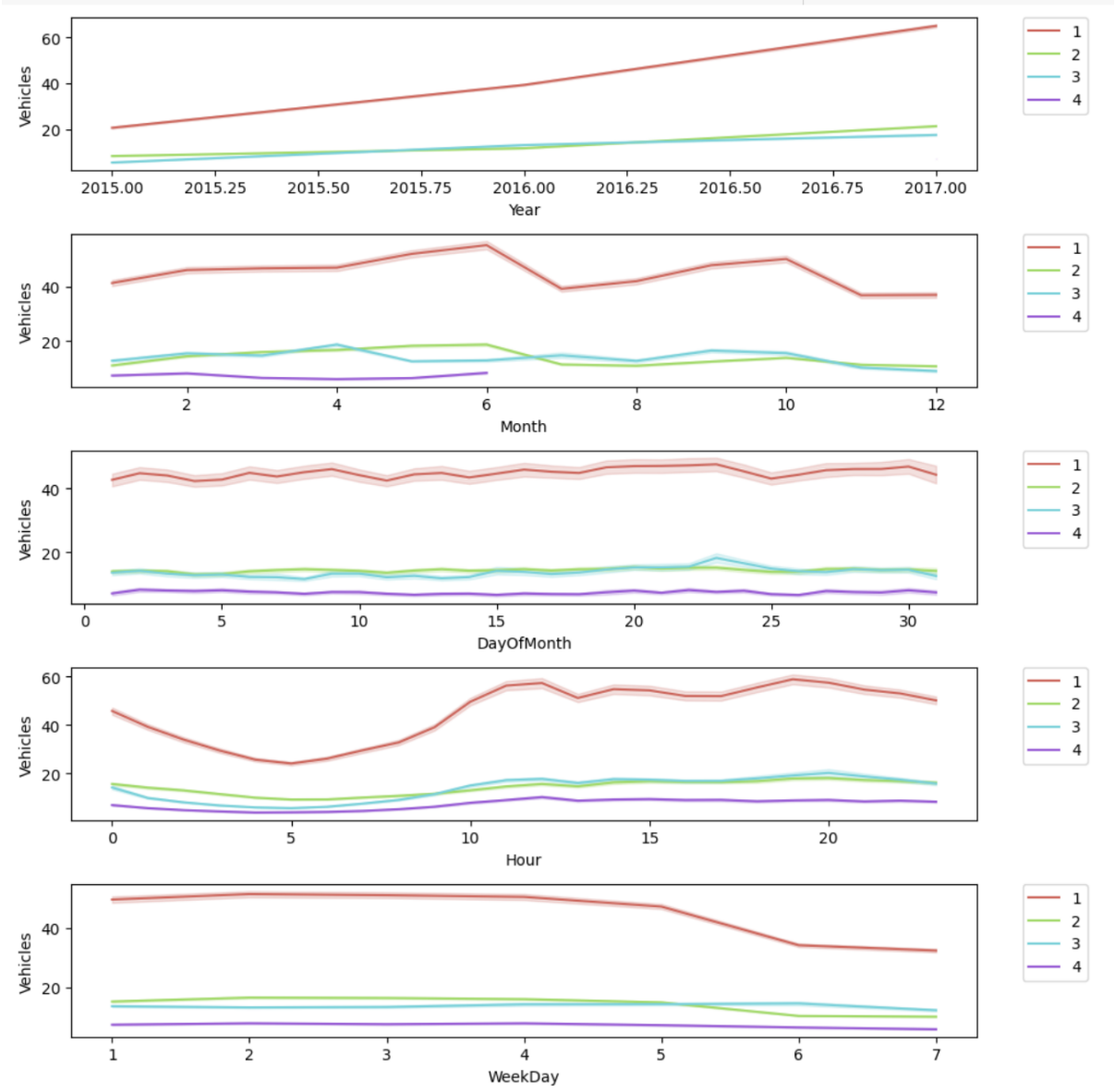

Figure 1: Variation in Number of Vehicles with Time

The Figure 1 shows the variation in the response variable that is Number of Vehicles or Traffic with time in different units like Hours, Days in Week, Days in Month, Months and Years.

2 Approaches Tried

Several machine learning algorithms were explored to tackle the traffic prediction problem. Here, we used KNN[2], Linear Regression[3], Decision Tree[4], Gradient Boosting[5] and Random Forest[6].

1. KNN was chosen for its simplicity and ease of implementation. In the context of traffic prediction, KNN works by finding the k nearest data points to a given sample and predicting the traffic condition (number of vehicles) based on the majority class among its neighbors. Despite its simplicity, KNN was effective for capturing local patterns in the data.

2. Linear Regression is a classic statistical method used to model the relationship between a dependent variable (number of vehicles) and one or more independent variables (e.g., time of day, month, etc). In traffic prediction, Linear Regression provides insights into how changes in predictor variables affect traffic patterns. While Linear Regression assumes a linear relationship between the features and the target variable, it can still offer valuable insights into traffic dynamics. However, It does not capture nonlinear relationships or interactions between variables as effectively as other machine learning techniques.

3. Decision Trees are versatile models that partition the feature space into hierarchical decisions. Each node of the tree represents a decision based on a particular feature, leading to a tree structure where the leaves correspond to the predicted outcomes. In traffic prediction, decision trees can capture complex relationships between traffic flow, time-of-day, weather conditions, and other relevant features. However, decision trees are prone to overfitting, especially when the dataset is large or contains noisy attributes.

4. Gradient Boosting is another ensemble learning technique that builds a series of weak learners sequentially, with each learner focusing on the errors made by its predecessors. By iteratively refining the model to reduce the residual errors, Gradient Boosting produces a strong predictive model capable of capturing intricate dependencies in the data. Gradient Boosting methods are adept at capturing complex patterns and achieving high predictive accuracy, making them well-suited for traffic prediction tasks with diverse and dynamic datasets.

5. Random Forest is an ensemble learning method that constructs multiple decision trees and combines their predictions through voting or averaging. By training each tree on a random subset of the data and features, Random Forest mitigates overfitting and improves generalization performance. Random Forest is robust to noisy data and can capture complex relationships in the dataset. It is widely used in traffic prediction due to its high predictive accuracy and scalability.

3 Experiments and Results

3.1 Dataset

The dataset contains the following :-

- Number of Instances/samples : 48120
- Number of Attributes : 4 DateTime, Junction, Vehicles, ID
- Out of these features: number of vehicles, year, month, day of the month, hour and which day of the week are of importance.

An example of the first few rows of the dataset :

DateTime	Junction	Vehicles	ID
2015-11-01 00:00:00	1	15	20151101001
2015-11-01 01:00:00	1	13	20151101011
2015-11-01 02:00:00	1	10	20151101021
2015-11-01 03:00:00	1	7	20151101031
2015-11-01 04:00:00	1	9	20151101041

3.2 Experimental setting

The experiments were conducted using a dataset[1] containing historical traffic data obtained from different junctions and time-related features after preprocessing.

Junction	Vehicles	Year	Month	DayOfMonth	Hour	WeekDay
1	15	2015	11	1	0	7
1	13	2015	11	1	1	7
1	10	2015	11	1	2	7
1	7	2015	11	1	3	7
1	9	2015	11	1	4	7

We split the dataset into training and testing sets. As it is regression task, we evaluated the performance of each based on metrics such as mean squared error, root mean squared error, mean absolute error and R-squared score[7] (measurement of the amount of variance captured by the model in the dataset).

3.3 Results

The results revealed that Random Forest and Gradient Boosting along with KNN Regressor and Decision Tree outperformed Linear Regression in terms of predictive accuracy and robustness. These ensemble methods effectively captured the intricate dependencies within the data, leading to more accurate traffic predictions.

1. KNN: With *"n-neighbours = 3, 5, 7, 9"* best performance was achieved for k = 7 with Root Mean Squared Error- 4.747739490924929 and R-squared Score- 0.9476914779853474.

2. Linear Regression: With default parameters Mean Squared Error- 175.4272217826185 and R-squared Score- 0.5929050899107802.

3. Decision Tree: With default parameters the Root Mean Squared Error: 4.881950147249841 and R-squared Score: 0.9446923294353229

4. Gradient Boosting: With "n-estimators" = [100, 200, 300], learning-rate = [0.05, 0.1, 0.2], max-depth = [3, 4, 5] best performance was achieved with 'learning-rate': 0.2, 'max-depth': 5, 'n-estimators': 300 with Root Mean Squared Error: 5.062888763687377 and R-squared Score: 0.9405166446736816

5. Random Forest: With n-estimators - 288 and n-iter = 50 and 100 Root Mean Squared Error: 4.298507564463358 and R-squared Score: 0.9571220436781024 and Root Mean Squared Error: 4.298507564463358 and R-squared Score: 0.9571220436781024

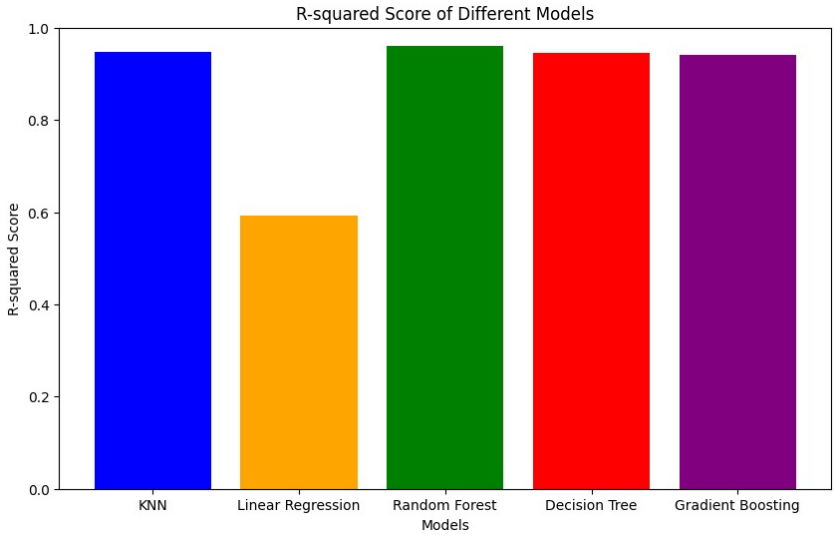

Figure 2: Comparision of R-squared score of models

4 Summary

In conclusion, this project demonstrates the effectiveness of machine learning algorithms like KNN, Linear Regression, Random Forest, Decision Tree, Gradient Boosting and others in traffic prediction. By leveraging data and relevant features, we can develop models capable of forecasting traffic conditions with high accuracy. The findings highlight the potential for implementing intelligent transportation systems to alleviate congestion and improve overall urban mobility.

5 Bibilography

1. **Dataset:** https://www.kaggle.com/datasets/fedesoriano/traffic-prediction-dataset/data
2. **KNN:** https://www.geeksforgeeks.org/k-nearest-neighbors-with-python-ml/amp/
3. **Linear Regression:** https://www.geeksforgeeks.org/ml-linear-regression/
4. **Decision Tree:** https://medium.com/analytics-vidhya/decision-tree-fce5018f3278
5. **Gradient Boosting:** https://www.datacamp.com/tutorial/guide-to-the-gradient-boosting-algorithm
6. **Random Forest:** https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/
7. **R-squared score :** https://statisticsbyjim.com/regression/interpret-r-squared-regression/