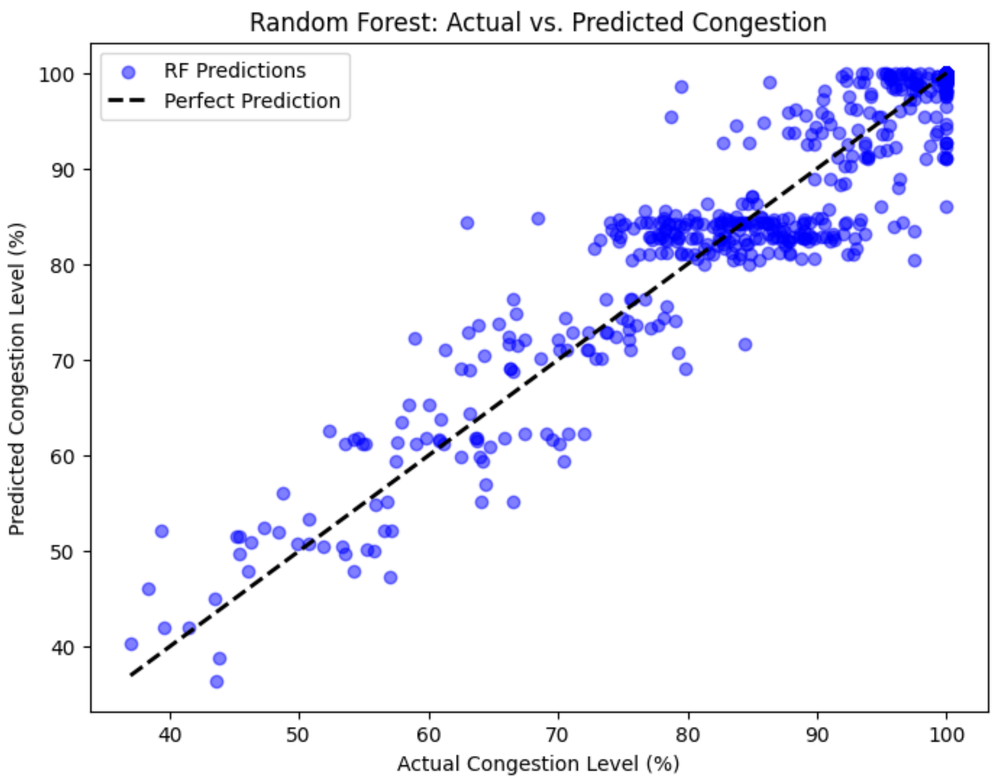
MACHINE LEARNING FOR TRANSPORTATION ANALYSIS

JIAYANG HONG

STATS 201, 2025

Introduction

This project examines how machine learning techniques, including supervised learning, and regression discontinuity (RD) design, can be used to analyze transportation policies and their effects. The project focuses on understanding public predicting congestion levels, and evaluating the causal effects of policy interventions.

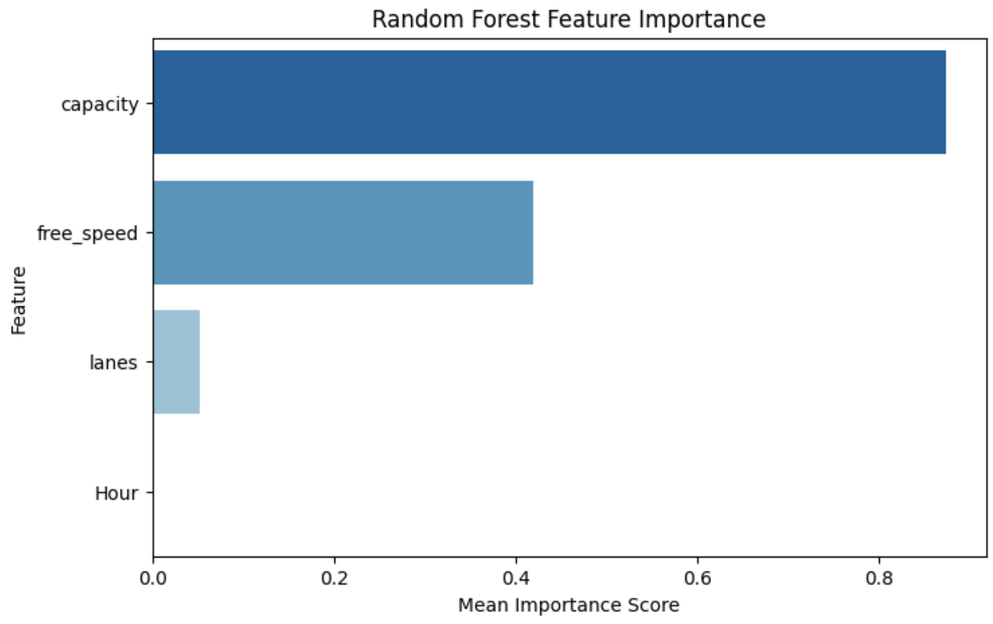


Methods

This study uses multiple data sources, including the General Modeling Network Specification (GMNS) dataset for transportation networks, and a simulated dataset for evaluating the impact of congestion pricing.

For prediction, supervised machine learning models, including Random Forest and XGBoost, are trained to predict congestion levels. Features such as road capacity, congestion pricing, and adaptive traffic signals are used for training. Model performance is evaluated using RMSE, R^2, and MAE.

For causal inference, the study applies RD design to examine the impact of congestion pricing exemptions on travel speed. Vehicles above a specific emissions threshold are exempt from congestion pricing, creating a natural cutoff for RD analysis. Robustness checks and sensitivity analyses are conducted to validate the results.



Results

Congestion prediction experiments show that Random Forest slightly outperforms XGBoost, achieving an RMSE of 4.04, an R^2 of 0.913, and an MAE of 2.40 (versus XGBoost's RMSE of 4.07, R^2 of 0.912, and MAE of 2.43). Feature importance analysis reveals that capacity and free speed are the most influential predictors, with lanes having a smaller yet notable effect. Regression Discontinuity (RD) analysis indicates that vehicles exempt from congestion pricing travel roughly 2 km/h faster than those subject to fees. Sensitivity checks across multiple bandwidths confirm the robustness of this causal effect, demonstrating a clear policy impact at the emission score cutoff.

Conclusion

Machine learning provides valuable tools for analyzing policy impacts in transportation. supervised learning improves congestion forecasting. RD analysis confirms the causal effects of congestion pricing policies on travel speed.

Future research can expand the dataset to include additional policy interventions and explore deep learning approaches for enhanced NLP analysis. The integration of real-world mobility data can further strengthen causal inference studies in transportation planning.

昆山杜克大学
DUKE KUNSHAN UNIVERSITY