MACHINE LEARNING FOR TRANSPORTATION ANALYSIS

JIAYANG HONG

STATS 201, 2025

Introduction

Machine learning is increasingly applied in social science research to enhance explanation, prediction, and causal inference. This study examines how machine learning techniques, including natural language processing (NLP), supervised learning, and regression discontinuity (RD) design, can be used to analyze transportation policies and their effects. The project focuses on understanding public sentiment toward policies, predicting congestion levels, and evaluating the causal effects of policy interventions.

Methods

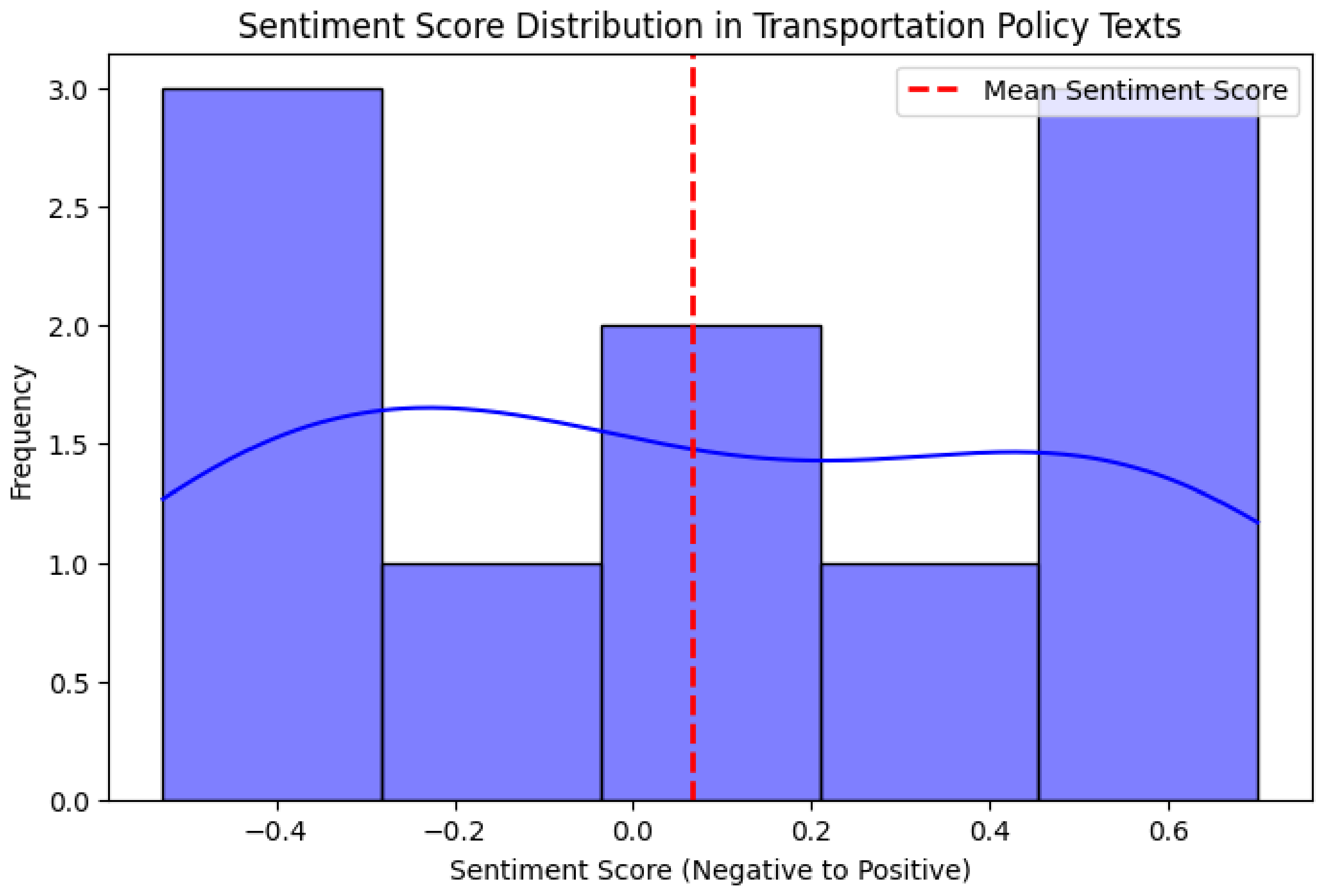
This study uses multiple data sources, including the General Modeling Network Specification (GMNS) dataset for transportation networks, policy-related text data for NLP analysis, and a simulated dataset for evaluating the impact of congestion pricing.

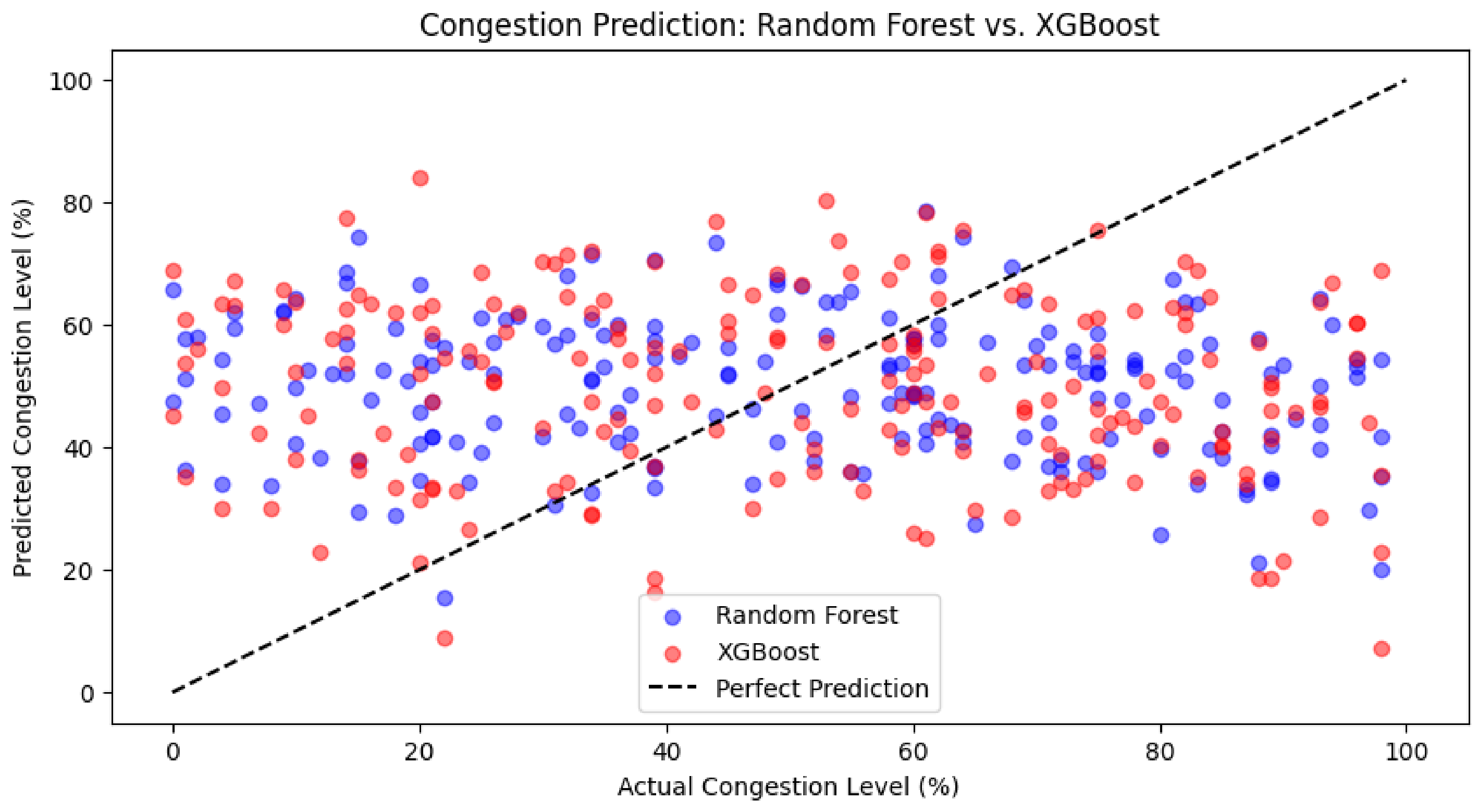
For explanation, NLP techniques such as sentiment analysis and topic modeling are applied to extract insights from policy-related text. The VADER sentiment analyzer is used to classify public sentiment, while Latent Dirichlet Allocation (LDA) and BERT are employed for topic modeling.

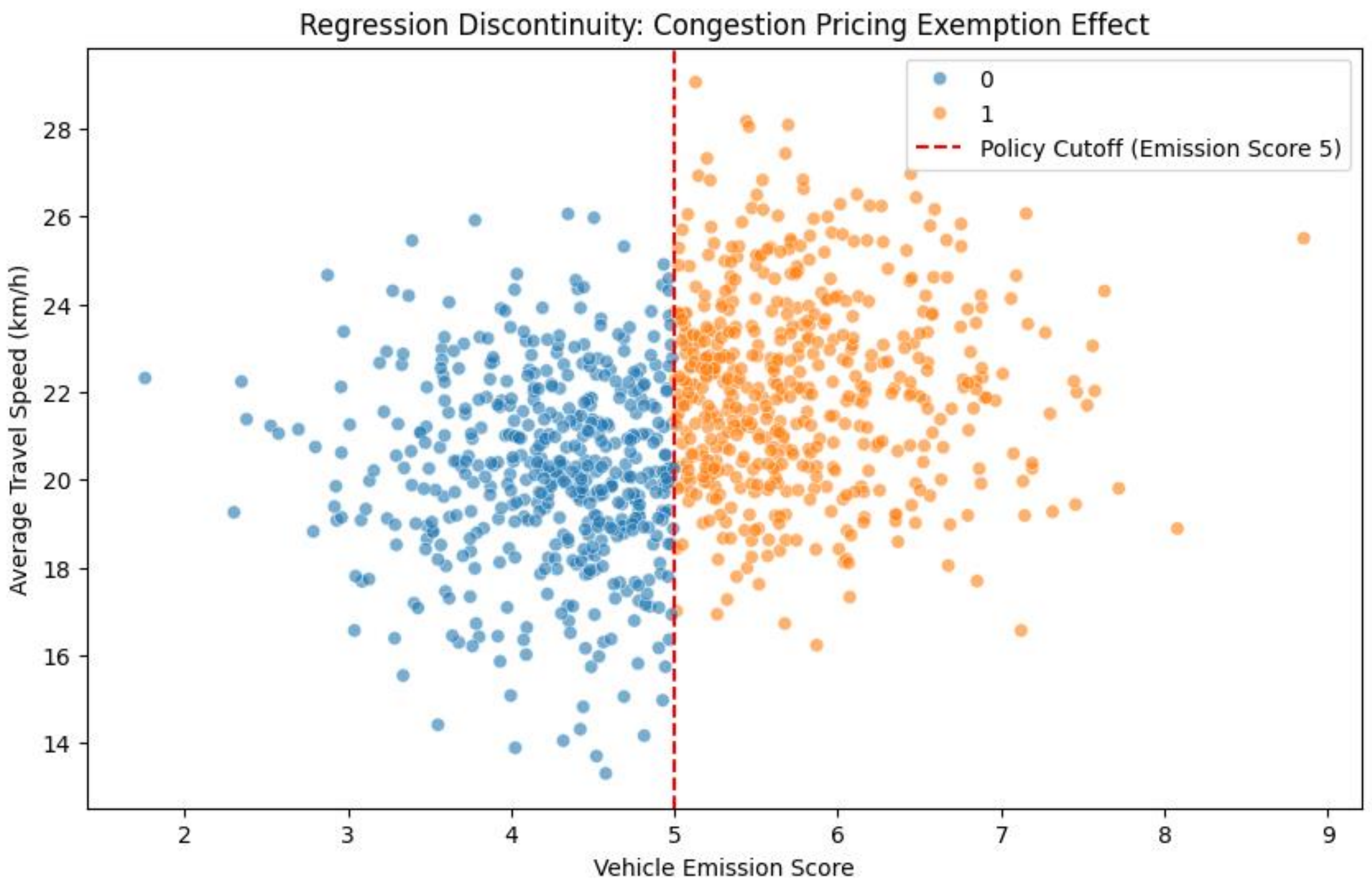
For prediction, supervised machine learning models, including Random Forest and XGBoost, are trained to predict congestion levels. Features such as road capacity, congestion pricing, and adaptive traffic signals are used for training. Model performance is evaluated using RMSE, R^2, and MAE.

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.

昆山杜克大学
DUKE KUNSHAN
UNIVERSITY


Sentiment Score Distribution in Transportation Policy Texts


Congestion Prediction: Random Forest vs. XGBoost


Regression Discontinuity: Congestion Pricing Exemption Effect

Results

NLP analysis shows that public sentiment toward transportation policies is polarized, with strong opinions both in support of and against congestion pricing. Topic modeling identifies congestion, equity, and infrastructure development as dominant themes in policy discussions.

Congestion prediction results indicate that Random Forest performs better than XGBoost, with lower RMSE and higher predictive accuracy. Feature importance analysis shows that congestion pricing and adaptive signals are key factors influencing congestion levels.

RD analysis finds that vehicles exempt from congestion pricing experience a statistically significant increase in travel speed of approximately 2 km/h. Sensitivity tests confirm the robustness of this finding across different bandwidths.

Conclusion

Machine learning provides valuable tools for analyzing policy impacts in transportation. NLP techniques help reveal public sentiment and major policy concerns, while 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.