

A BOOSTED GLIMPSE: USING A GRADIENT BOOSTING MACHINE MODEL TO UNRAVEL THE FACTORS INFLUENCING TRANSPORT IN HANOI

Eric Wanjau¹, Minh Kieu⁴, Nick Malleon^{1, 2}, Alexis Comber², Kristina Bratkova¹, Phe Hoang Huu³, Thanh Bui Quang⁵ and Hang Nguyen Thi Thuy⁶

¹Leeds Institute for Data Analytics, University of Leeds, UK | ²School of Geography, University of Leeds, UK | ³R&D Consultants, Hanoi City, Vietnam | ⁴Faculty of Engineering, University of Auckland, New Zealand | ⁵Faculty of Geography, VNU University Science, Hanoi, Vietnam | ⁶VNU Vietnam Japan University, Vietnam National University, Hanoi

1. Introduction

Objective 1:

Exploratory Data Analysis

An overwhelming majority in Hanoi city use motorbikes as their primary means: around 2 motorbikes per person.

Objective 2:

Predict attitudes towards ban

Implications are serious traffic congestion, air and noise pollution

Transport survey which currently has 30K responses has been undertaken to capture:

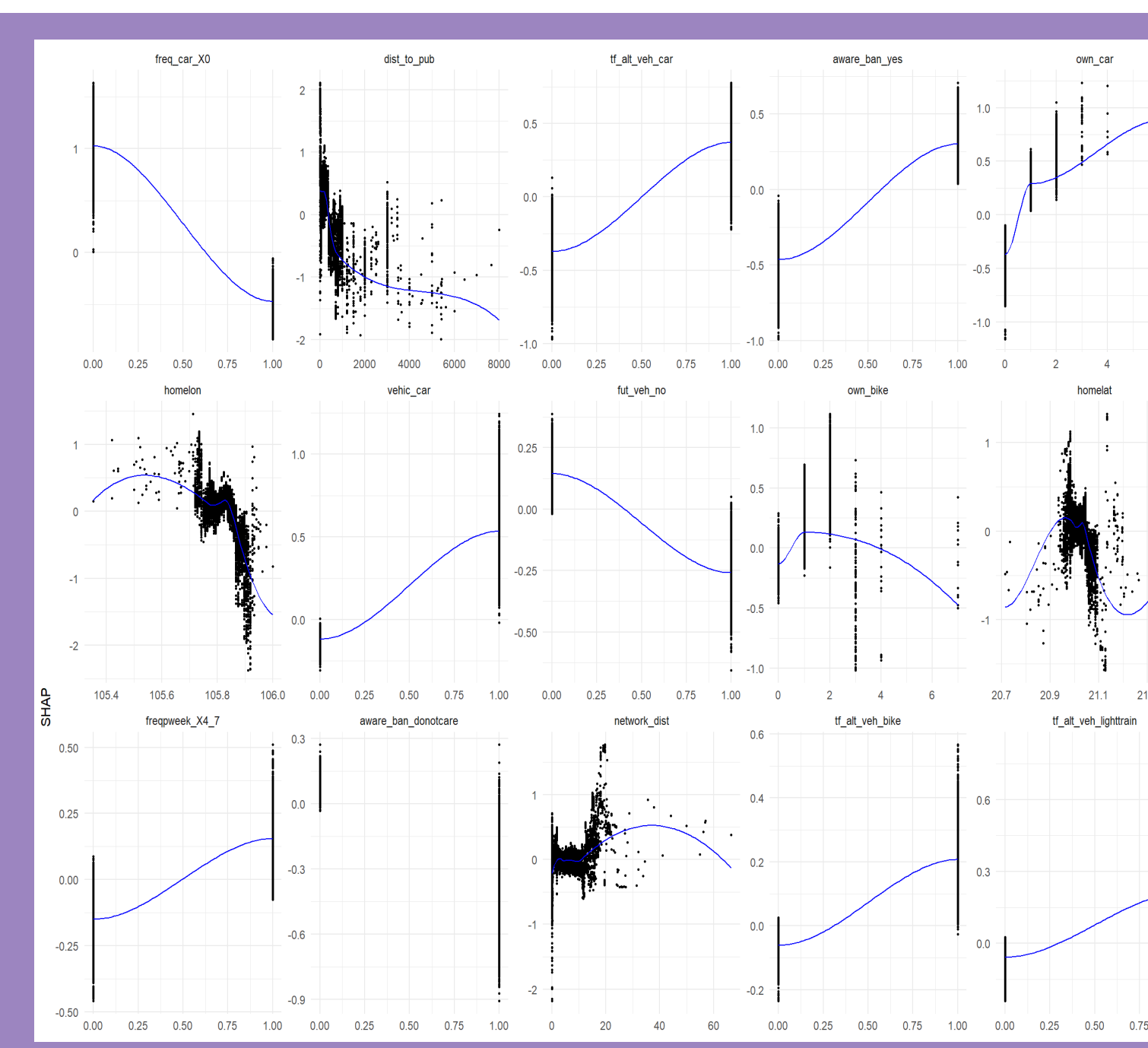
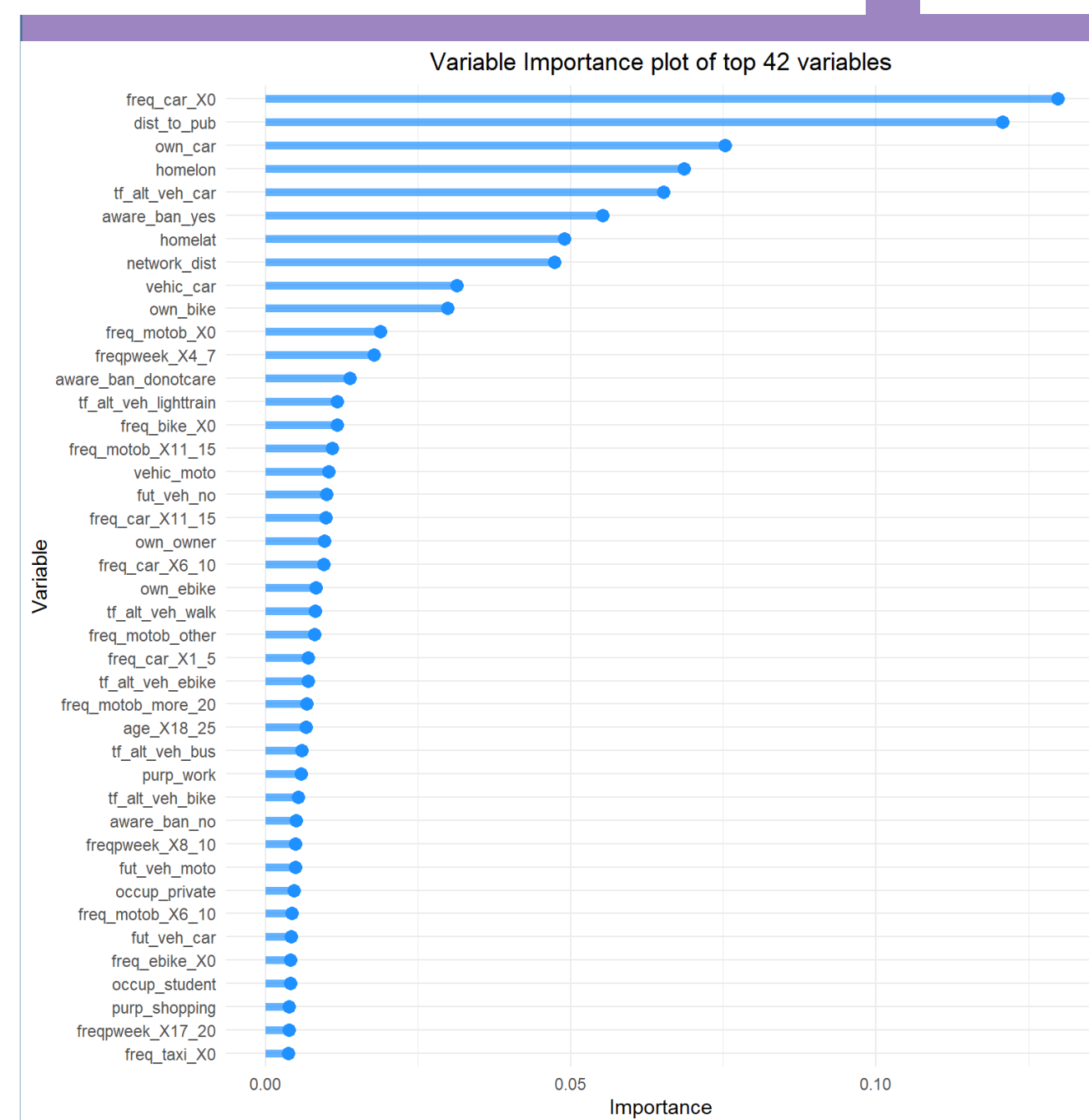
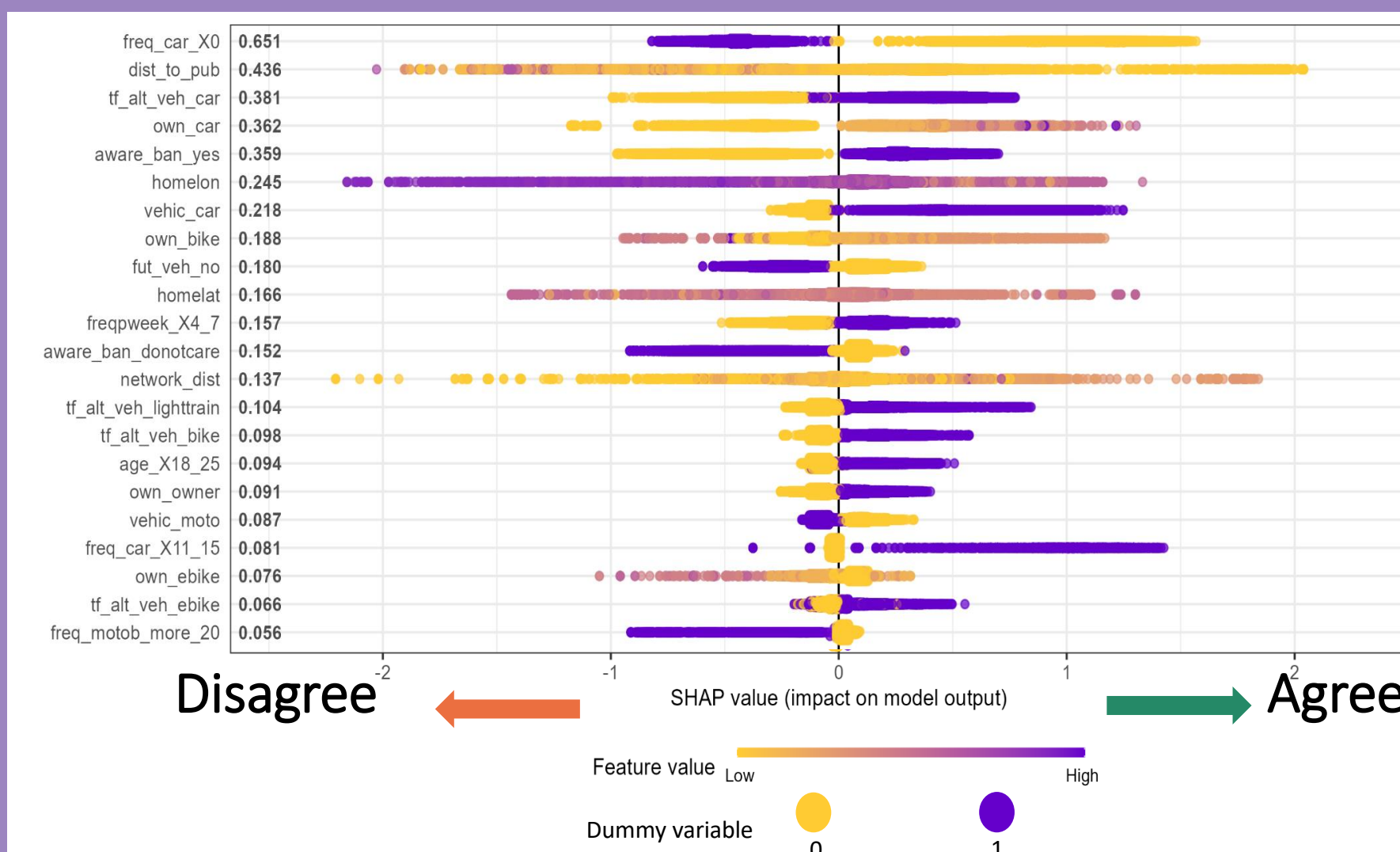
- **Demographics:** age, gender, location
- **Travel behavior:** origin, destination, transport mode, purpose
- **Attitudes toward a motorbike ban:** opinion, awareness, alternative vehicle

Objective 3:

Variable importance analysis

4. Model results, VIP, SHAP summary plot, Partial Dependence Plots

Metric	Value
Accuracy	0.878
Recall	0.853
Specificity	0.919
PPV	0.945
F-measure	0.897
ROC-AUC	0.953



6. Preliminary Model Interpretation

freq_car_X0, a “dummy variable” that has a value of 1 if a respondent does not travel using a car and 0 otherwise, is the most predictive variable in this model.

Respondents who do not use a car (*freq_car_X0* == 1) are associated with low SHAP values (which pushes the model towards classifying their response as “disagree”).

The second most important variable is the distance to public transport (*dist_to_pub*), a numeric variable.

It appears that larger distances to public transport are associated with the disapproval of the motorbike ban.

Respondents who would consider using cars and lighttrains (*tf_alt_veh_car*, *tf_alt_veh_lighttrain*) in case of a motorbike ban are more likely to agree with the motorbike ban.

Awareness of the ban (*aware_ban_yes*) increases the probability of agreeing with the motorbike ban.

See more model metrics, model interpretation techniques, and explore what-if scenarios at: bit.ly/utm_hn_model

7. Preliminary policy implications

Cars instead of active or public transport may be the most likely alternative means of transport if a motorbike ban were to be implemented. This is however undesirable.

Routing public means of transport close to where people live/work could increase their adoption as a means of transport. This can be guided by other important variables such as home location.

People would be open to other alternative means such as trams and bikes. It will be worth investing in these modes of travel.

The need for public sensitization and involvement in the formulation of policies is vital if they are to be widely accepted and adopted.

8. Future works/Ongoing works

Project's website: bit.ly/utm_hn_intro

British Academy project undertaking urban transport modelling in Hanoi [grant number UWB190190].

e.wanjau@leeds.ac.uk

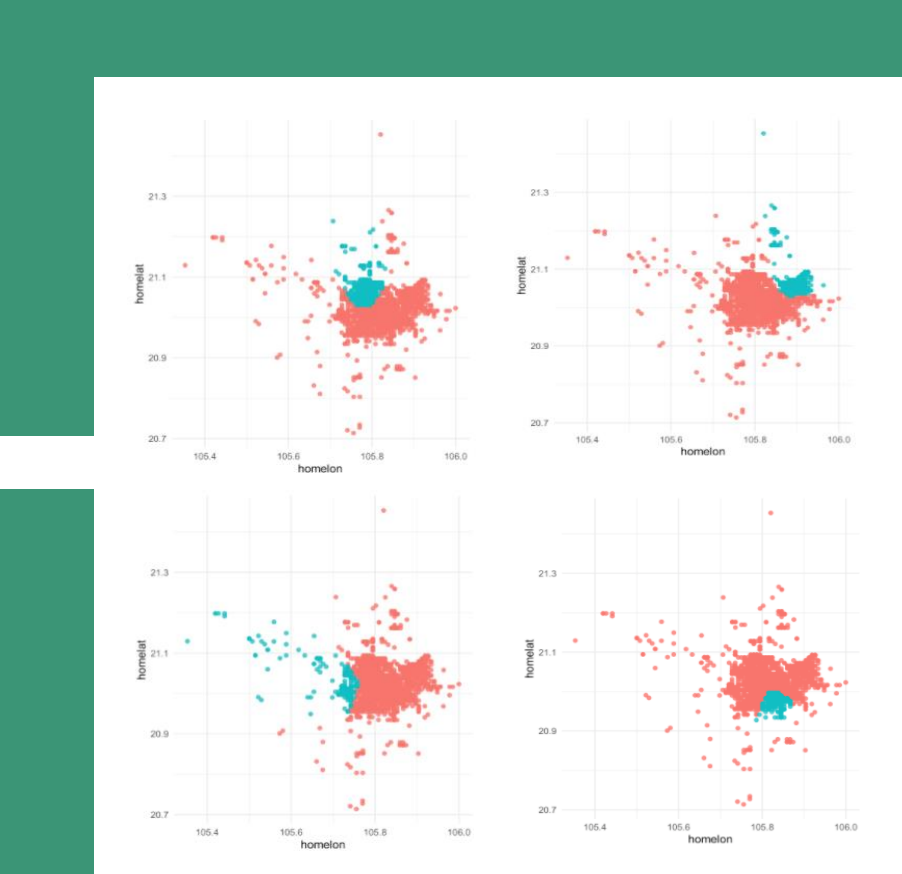


3. Modelling: Agree/Disagree with motorbike ban

Feature Engineering

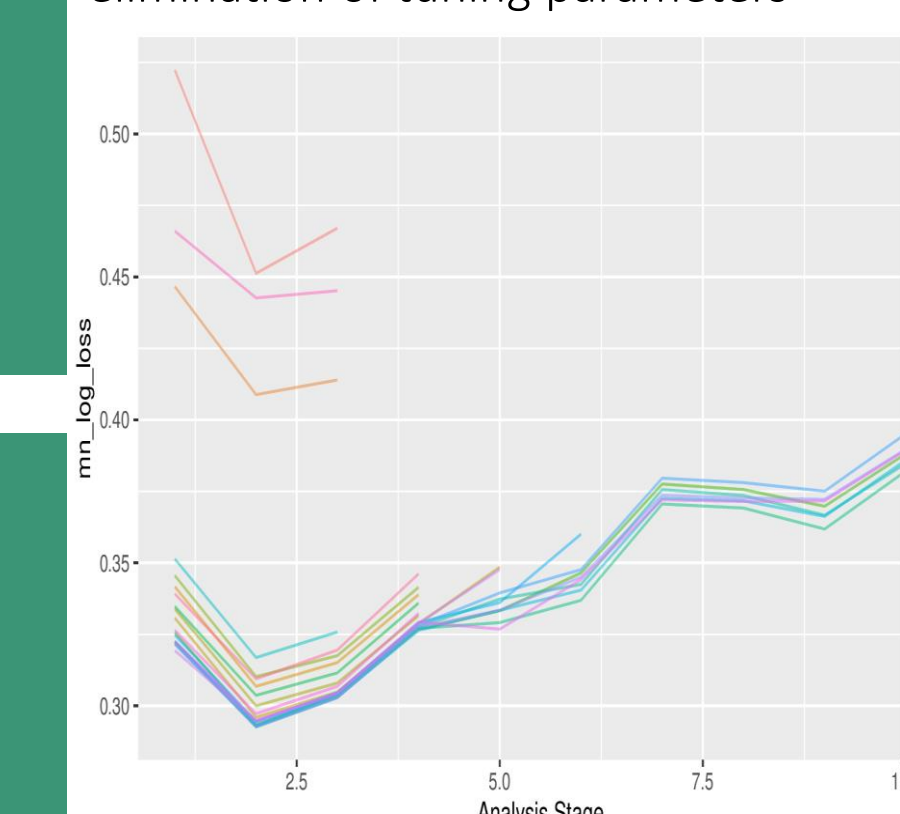
- Dummy variables
- Tokenization of text variables
- Near zero variance filtering
- Pooling infrequent observations
- Upsampling

Spatial k-fold CV

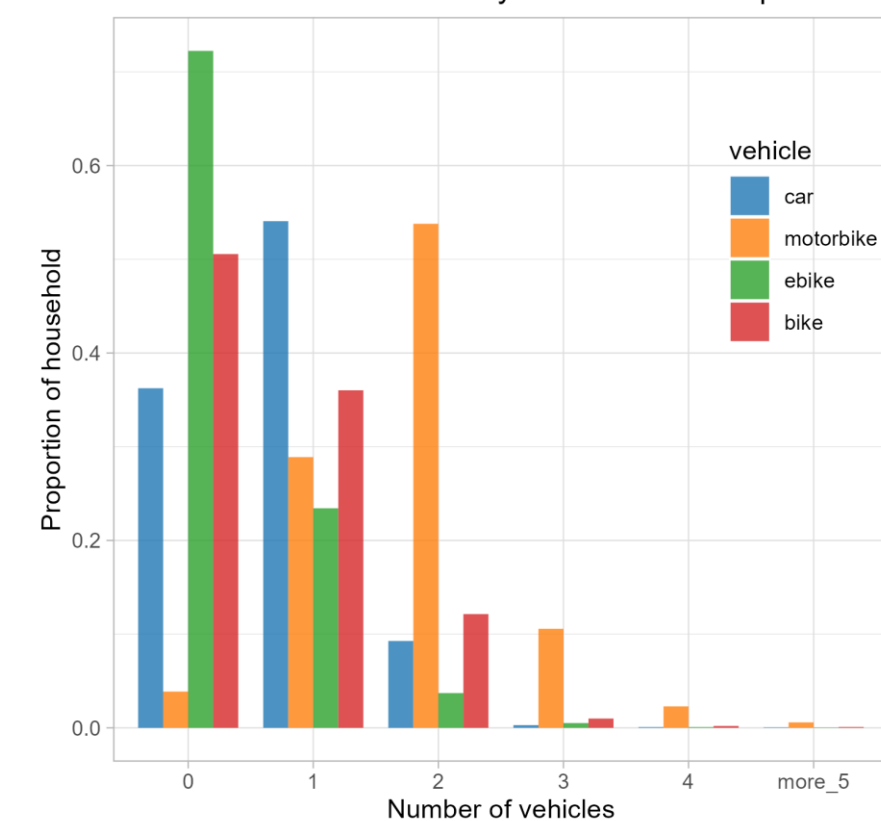


XGBoost Hyperparameter Tuning

ANOVA-based grid search: Incremental elimination of tuning parameters

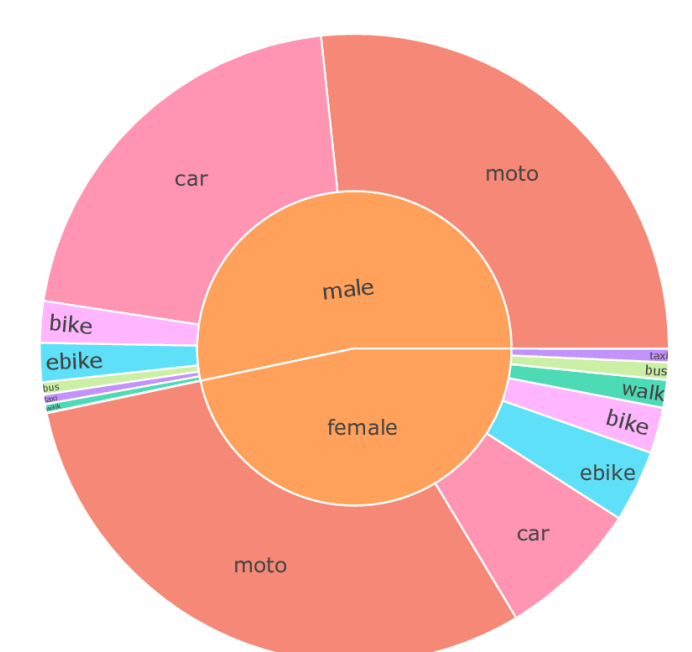


Distribution of family vehicle ownership

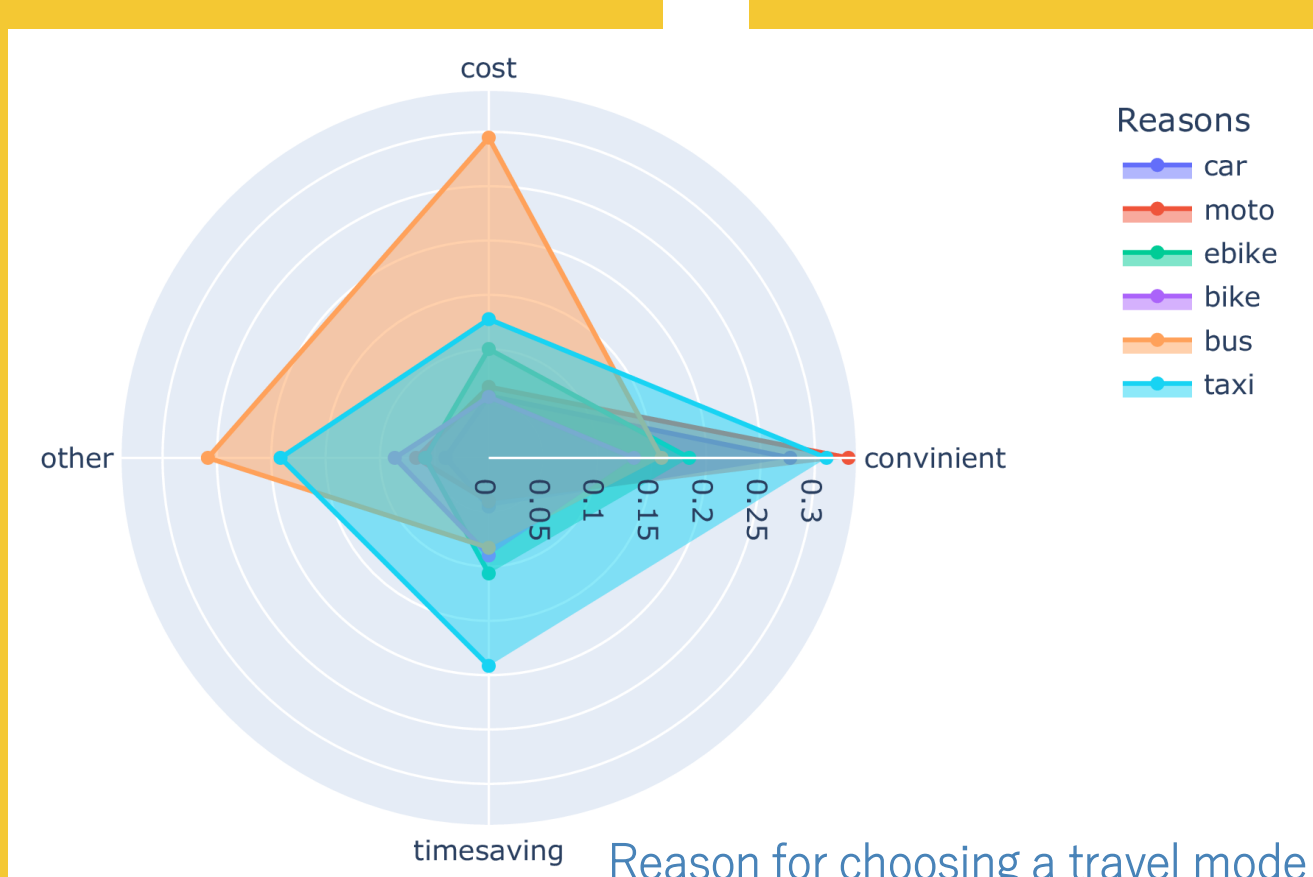


Awareness of ban vs opinion

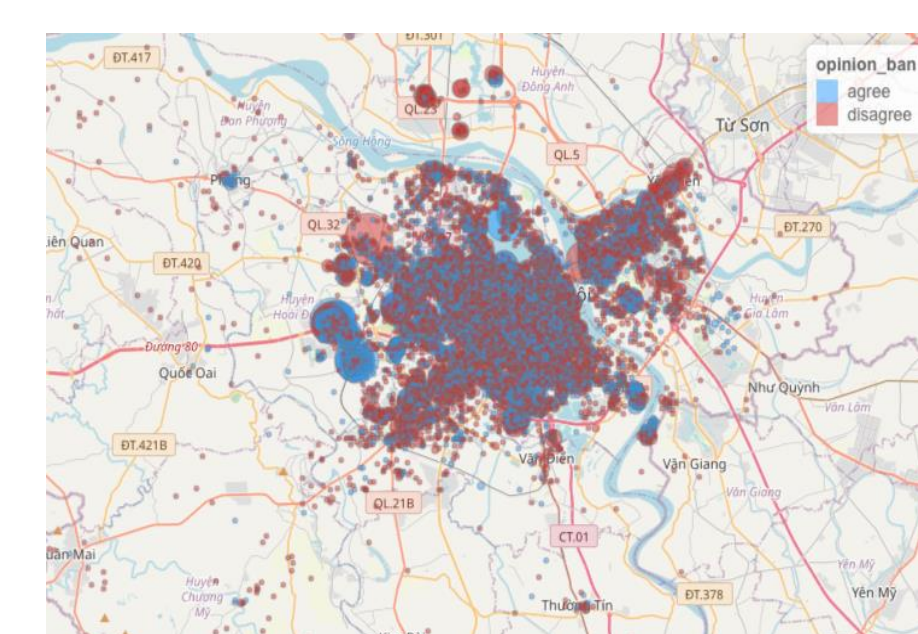
2. EDA



Gender vs mode for primary trip



Reason for choosing a travel mode



Opinion on ban based on home location