Predicting Potential Use of Shared E-scooters through Machine Learning Techniques- a Birmingham Case Study

Mostafa Jafarzadehfadaki, Graduate Research Assistant and **Dr. Virginia Sisiopiku**, Professor

TREND Lab, Department of Civil, Construction, and Environmental Engineering, University of Alabama at Birmingham (UAB)

Introduction

Shared micro-mobility, including e-scooter sharing, has drawn research attention in recent years. However, there is still a need to understand facilitators and barriers to adoption of shared e-scooters within different urban contexts.

Study Objective

The aim of the study is to (a) gain preliminary insights into socio-demographic characteristics, travel preferences and mode choice factors of travelers in Birmingham, AL, and (b) use these factors to predict potential use of shared e-scooters using machine learning techniques.

Data and Method

✓ **Data Collection:**
- A questionnaire survey was developed using the Qualtrics platform to survey travelers' preferences and attitudes toward micromobility use in March 2022. Participants were >18 years old and lived in Birmingham.

✓ **Sample size:**
- 281 valid responses were obtained. Based on literature, the sample size was deemed an acceptable sample size for Birmingham (162,000 population over 18 years old), with a 90% confidence level (Census 2020).

✓ **Dependent variables:**
- Various sociodemographic characteristics of survey participants
- Travel modal choice factors (e.g., cost, time, reliability, comfort, safety)

✓ **Independent variable: (Imbalanced Dataset)**
- E-scooter use (Users N=19; and non-users N=262)

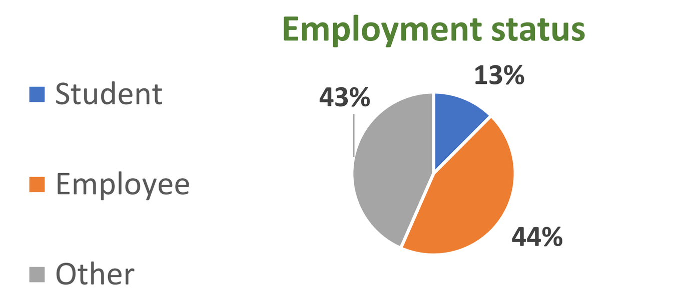
✓ **Statistical model:**
- Logistic regression (LR)

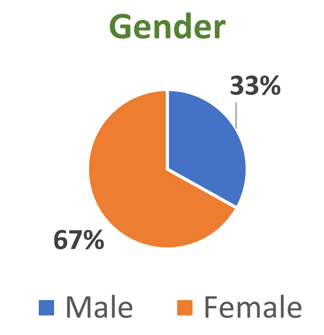
✓ **Machine learning models:**
- Random forest (RF)
- Decision trees (DT)
- Support vector machine (SVM)

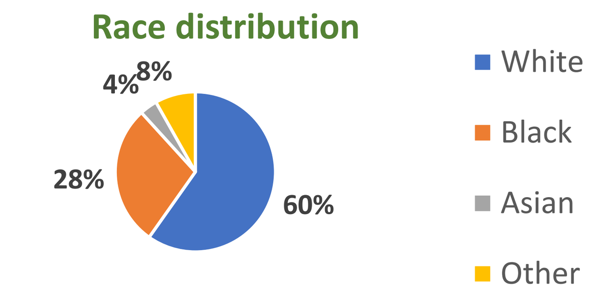
✓ **Resampling methods:**
- Random oversampling (RS)
- SMOTETomek (ST)
- SMOTE

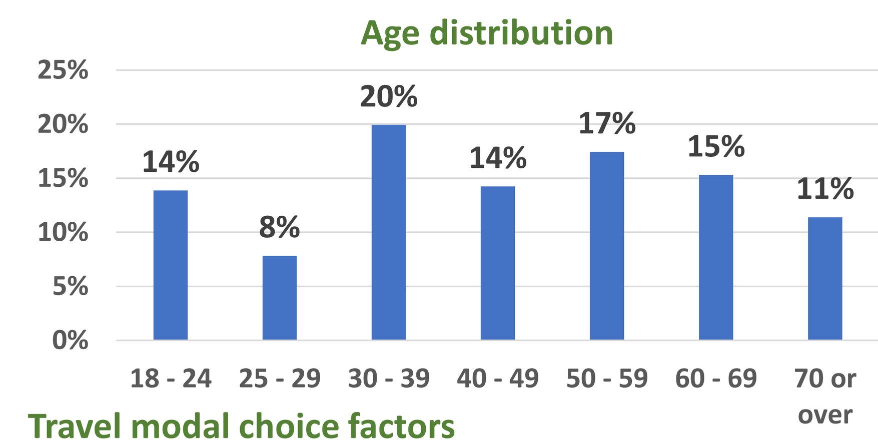
✓ **Feature importance methods:**
- Gini importance (mean decrease impurity)
- Permutation-based importance

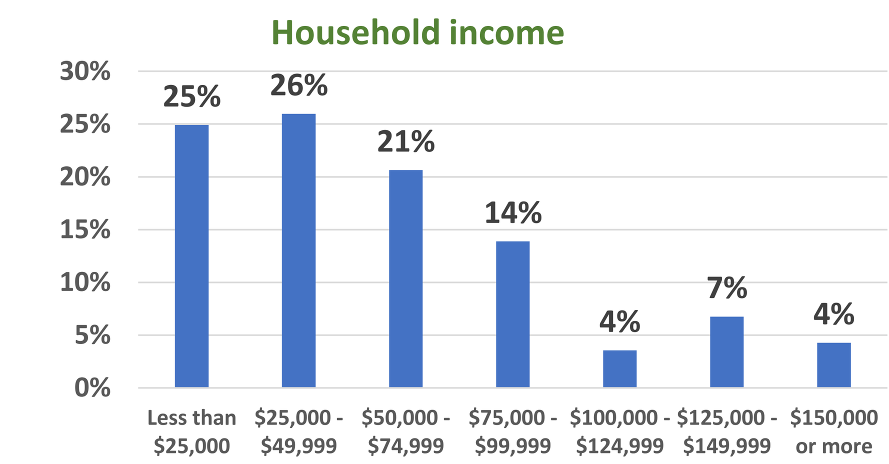
Descriptive Analysis Results











Travel modal choice factors

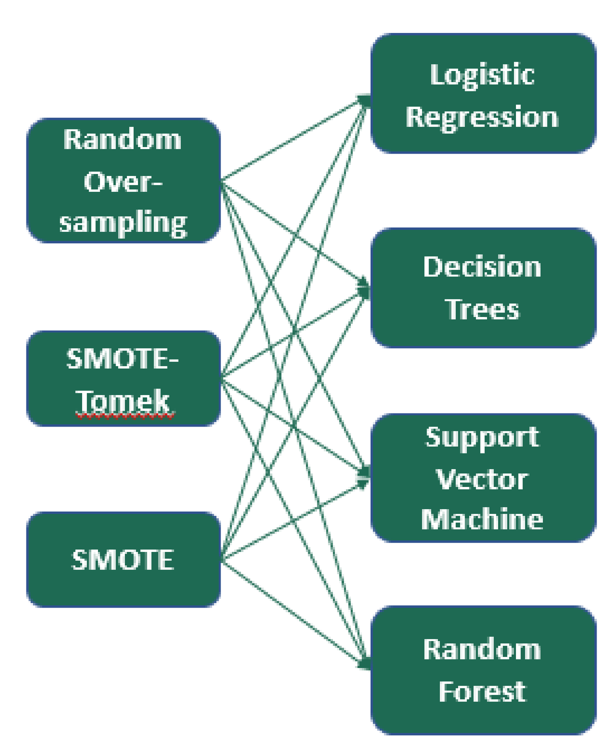
Variables	Mean	STD Dev.
Cost	3.65	1.22
Time	3.81	1.04
Reliability	**4.21**	0.97
Comfort	3.83	0.95
Safety	**4.32**	0.94
Environmental impacts	3.15	1.22

Personal attitudes and preferences

Statements	Mean	STD Dev.
Living without car	1.78	1.10
Using public transit	1.97	1.18
Traveling with non-motorized modes	2.27	1.24
Ability to use new technologies	**4.11**	1.10

Resampling Methods & Prediction Models Results

Resampling Methods

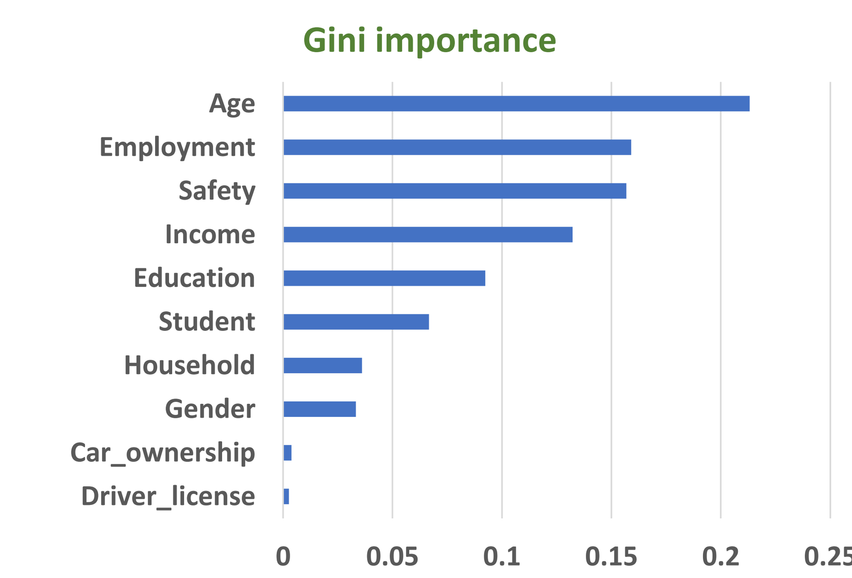


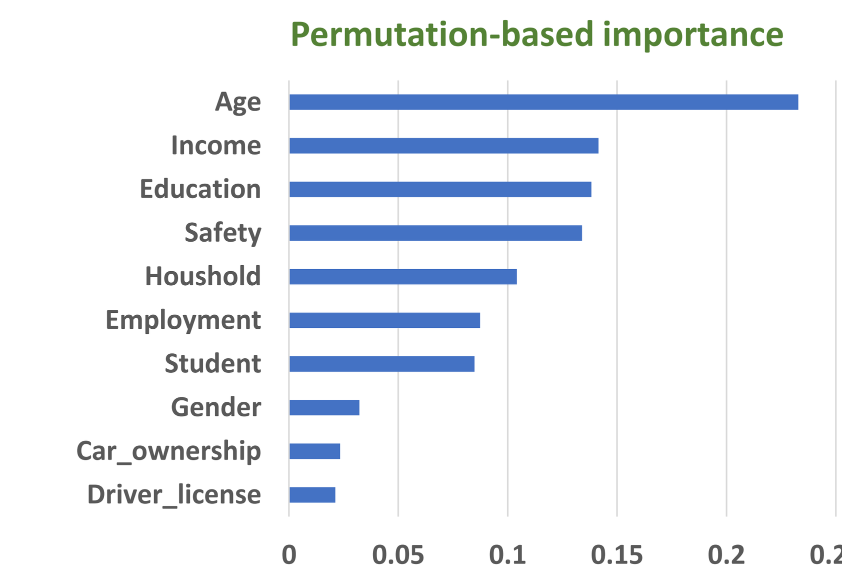
Comparison of Prediction Models Results

Model	Resampling methods	Precision	Recall	F1-score	ROC-AUC	Accuracy
Logistic Regression	RS	0.22	0.77	0.87	0.80	0.77
	ST	0.21	0.83	0.33	0.80	0.76
	SMOTE	0.17	0.67	0.28	0.71	0.75
Decision Tree	RS	0.50	1.00	0.67	0.96	0.92
	ST	0.28	0.83	0.42	0.83	0.83
	SMOTE	0.42	0.83	0.56	0.71	0.75
Support Vector Machine	RS	0.23	0.83	0.36	0.81	0.78
	ST	0.18	0.67	0.29	0.72	0.76
	SMOTE	0.17	0.67	0.27	0.71	0.74
Random Forest	**RS**	**0.67**	**1.00**	**0.80**	**0.98**	**0.96**
	ST	0.30	0.83	0.50	0.86	0.88
	SMOTE	0.42	0.83	0.56	0.86	0.92

- ➢ The dataset is imbalanced, i.e., the number of samples for e-scooter users and non-users is significantly different (e.g., 6 % of e-scooter users of the total sample).
- ➢ Random oversampling, SMOTETomek, and SMOTE were employed to oversample the minatory class to examine which classifier yields best prediction results.
- ➢ Multi-collinearity tests were performed by Spearman correlation. The dataset combines binary and categorical variables (ordinal and nominal scale).
- ➢ performance metrics were derived from the confusion matrix, such as accuracy, precision, recall, and F1-score. The **random forest with random oversampling** was the most accurate model due to the higher accuracy and ROC_AUC score.

Feature Importance Results





- ➢ Feature Importance techniques calculate a score for all the input features for a particular model. The scores represent the importance of each feature for predicting e-scooter use. A higher score indicates that the feature will have a more significant impact on the prediction model.

Model Assessment

- ➢ The prediction performance of each method, considering the resampling strategy, was evaluated based on the performance measures such as K-fold cross-validation (K=5) and the Matthews correlation coefficient, used in machine learning as a measure of the quality of binary and multiclass classifications.

Summary and Conclusions

- ➢ In this study, an online survey was designed and distributed in order to collect data about shared e-scooter use in Birmingham. Out of 281 valid responses, 19 participants used e-scooter.
- ➢ Given the small sample size, a variety of statistical and machine learning methods were used to predict shared e-scooter user adoption and select the best model fit.
- ➢ The findings showed that **age**, **income**, **employment**, and **safety** are the **most influential factors in using shared e-scooter in Birmingham**.
- ➢ The feature importance results showed that **gender**, **car ownership**, and **possession of a driver's license** are **not significantly associated with shared e-scooter use**, **contradicting reports from earlier studies.**

Acknowledgment

❑ This work was supported by the US DOT through STRIDE (Project D3)

STRIDE Southeastern Transportation Research, Innovation, Development and Education Center