Addressing Imbalanced Data for Health-related Travel Mode Choice

Akalin K. Berkhan[1*] and Guler S. Ilgin[1]

Department of Civil and Environmental Engineering, Pennsylvania State University, University Park, PA, 16802, US[1]

INTRODUCTION

This study addresses health-related travel data, encompassing information on individuals, households, and regions, to construct a model for predicting travel mode choices. Despite the richness of the dataset, consisting of 12,480 trips, the distribution of travel mode choices is imbalanced, with 97% opting for private transport and only 3% for public transport. Thus, the direct application of a Logistic Model[1,2], commonly used for mode choice modeling on this dataset is hindered by this imbalance. Because, during testing the model may erroneously allocate all instances to private transport, resulting in a misleading accuracy of 97%, while predicting 100% private transport and 0% public transport. This discrepancy necessitates the exploration of alternative data regularization, manipulation, or modeling methods.

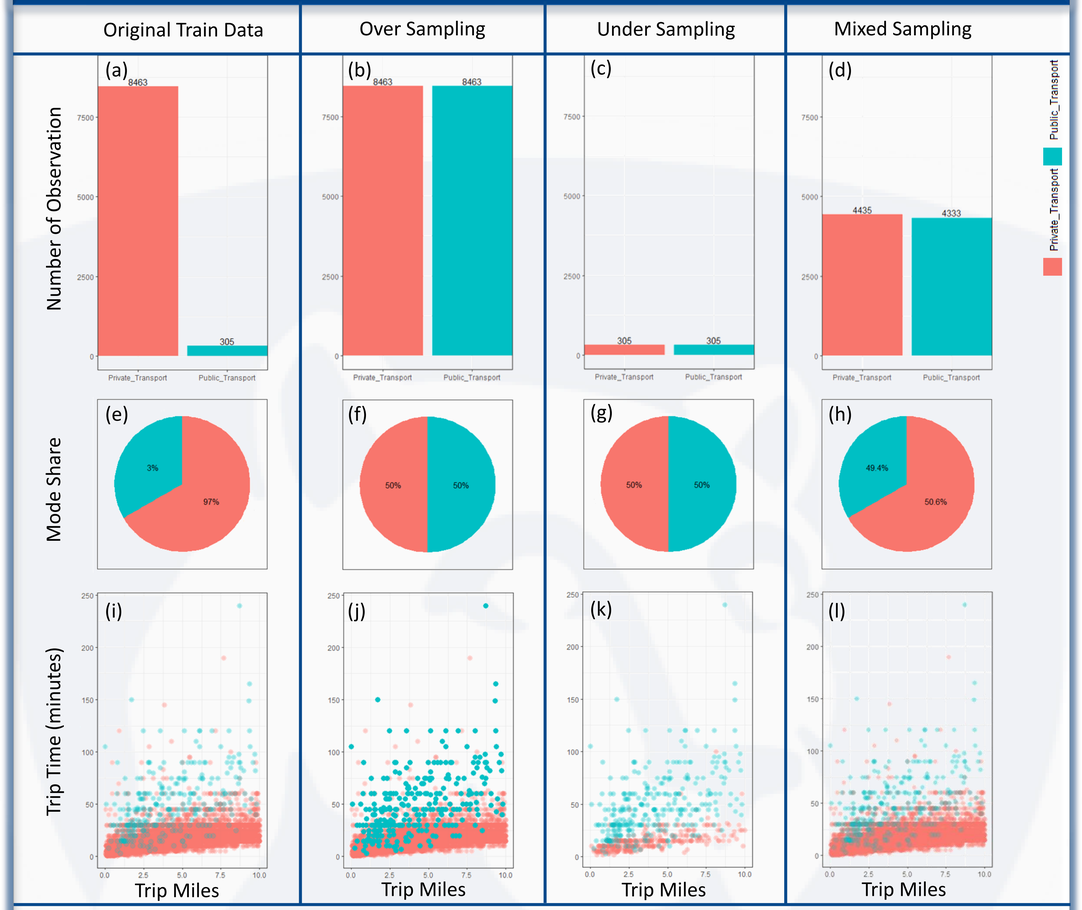
OBJECTIVE

The primary objective is to construct a model with interpretable prediction parameters that effectively captures individual mode choices, all while achieving enhanced accuracy compared to the imbalanced dataset. The rationale for evaluating the Logit Model over classification methods such as Support Vector Machine (SVM)[3] or Random Forest[4], commonly encountered in the literature, is the importance of parameter estimation methods for scenario analyses. Through these analyses, it becomes possible to assess shifts in mode choices arising from rises in fuel costs or increasing ticket prices.

MATERIAL AND METHOD

In this study, the dataset, comprising 19 independent variables (10 dummy and 9 numeric), is partitioned into 70% for training and 30% for testing. The training and test datasets exhibit a similar distribution concerning the dependent variable. Post-separation, models are developed and compared with synthetic datasets obtained through data manipulation methods aimed at rectifying training data set imbalance. Over sampling, under sampling, and mixed (over and under) sampling methods[5] are chosen over random sampling methods like Random Over Sampling Examples (ROSE)[6] and Synthetic Minority Oversampling Technique (SMOTE)[7].

ANALYSIS RESULTS



Variable Name	Variable Type	Description	Mean	Model Estimation: Original Train Data	Over Sampled	Under Sampled	Mixed Sampled
Constant				0.961	**4.270**	7.127	**3.636**
TripTimeMin	Numeric	Trip time in minutes	16.55	**0.079**	**0.107**	**0.160**	**0.101**
TripMiles	Numeric	Trip distance in miles	3.905	**-0.183**	**-0.251**	**-0.384**	**-0.207**
DwellTime	Numeric	Spent time in minutes arrival at destianation	73.41	**0.004**	**0.005**	**0.007**	**0.005**
PersonIncome	Numeric	Annual income in USD per person in hh	40102	0.001	**0.001**	0.001	**0.001**
VehiclePerDriver	Numeric	Number of cars per person in hh	1.092	**-2.308**	**-2.380**	**-2.469**	**-2.599**
GasPrice	Numeric	Gas price in USD per gallon	239.8	-0.005	**-0.019**	**-0.030**	**-0.021**
Age	Numeric	Age of the trip maker	59.35	**-0.014**	**-0.010**	**-0.042**	-0.005
PopulationDensity	Numeric	Persons per square mile	4559	0.001	0.001	0.001	**0.001**
ResidentAreaDensity	Numeric	Housing units per square mile	2090	**0.001**	**0.001**	**0.001**	**0.001**
TimeAftemoon	Dummy	1 if the trip took place in the afternoon, 0 otherwise	0.416	**-1.635**	0.121	0.707	0.323
TimeEvening	Dummy	1 if the trip took place in the evening, 0 otherwise	0.047	**-3.317**	**-2.582**	-4.622	**-2.531**
HomeOwnership	Dummy	1 if the hh has their own house, 0 otherwise	0.779	**-0.815**	**-1.196**	**-1.315**	**-1.313**
Child	Dummy	1 if there is an individual under 16 years old in hh, 0 otherwise	0.172	0.125	**-0.220**	**-1.188**	-0.172
Rail	Dummy	1 if the there is a rail system available, 0 otherwise	0.188	**0.753**	**0.695**	**1.240**	**0.869**
EduPS	Dummy	1 if trip maker's education level is less then high school, 0 otherwise	0.042	0.348	**1.617**	1.861	**1.911**
EduHS	Dummy	1 if trip maker has highschool graduate degree, 0 otherwise	0.179	0.021	**0.700**	0.644	**0.972**
EduUG	Dummy	1 if trip maker has undergraduate degree, 0 otherwise	0.285	-0.015	**0.761**	0.847	**0.942**
Worker	Dummy	1 if trip maker is working, 0 otherwise	0.384	0.017	**-0.190**	-0.208	-0.177
Driver	Dummy	1 if trip maker can drive, 0 otherwise	0.926	**-2.457**	**-2.179**	**-1.881**	**-2.296**
Model testing results			*Accuracy:*	*0.981*	*0.964*	*0.952*	*0.963*
			Sensitivity:	*0.550*	*0.907*	*0.922*	*0.899*
			Specificity:	*0.996*	*0.967*	*0.953*	*0.965*

MATERIAL AND METHOD *(continuing)*

The reason why random and unsupervised methods are not preferred lies in the nature of the dataset, which exclusively comprises positive values and dummy variables are constrained to assume only values of 0 or 1.

RESULTS

The resulting models undergo thorough evaluation using test data, considering accuracy, sensitivity, and specificity. Despite achieving the highest accuracy on the original training data, precision is notably low, indicating a significant deviation in predicting public transport trips from the actual situation. Consequently, the Logit Model obtained using oversampled data emerges as the most suitable, supported by the highest number of statistically significant model parameter estimates at a 90% confidence level.

CONCLUSION

This study successfully develops and evaluates a Binary Logit Model to predict individual travel mode choices, addressing the challenge of imbalanced data. Future research avenues may explore the development of similar models using random sampling techniques aligned with variable structures. Additionally, specific machine learning techniques adapted for unbalanced data in preference modeling could be developed. This work holds significant potential to inspire research across various fields that confront imbalanced datasets.

REFERENCES

1. Ben-Akiva, M. E. (1973). Structure of passenger travel demand models (Doctoral dissertation, Massachusetts Institute of Technology).
2. Ben-Akiva, M. E., & Lerman, S. R. (1985). Discrete choice analysis: theory and application to travel demand (Vol. 9). MIT press.
3. Zhang, Y., & Xie, Y. (2008). Travel mode choice modeling with support vector machines. *Transportation Research Record*, *2076*(1), 141-150.
4. Khoshgoftaar, T. M., Golawala, M., & Van Hulse, J. (2007, October). An empirical study of learning from imbalanced data using random forest. In 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007) (Vol. 2, pp. 310-317). IEEE.
5. Zhao, J., & Liu, N. (2019). Semi-supervised classification based mixed sampling for imbalanced data. Open Physics, 17(1), 975-983.
6. Lunardon, N., Menardi, G., & Torelli, N. (2013). R package'ROSE': Random Over-Sampling Examples.
7. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

Dr. K. Berkhan Akalin
kba5465@psu.edu*

PennState College of Engineering