The University of Texas Rio Grande Valley

Center for Multidisciplinary Research Excellence in Cyber-Physical Infrastructure Systems (MECIS)

An Analysis of Consumer Preferences in Vehicle Selection Using Data-Driven Models

Darryl Amuzu, Fatemeh Nazari, Ph.D.

National Science Foundation

NSF Award No. 2112650

Abstract

A good understanding of the preferences of consumers in vehicle selection is key to facilitating efficient and sustainable transportation systems.
Policymakers equipped with this knowledge will be able to make decisions that cater to the specific needs of consumers. This study applies machine learning techniques (primarily random forest)on survey data from a diverse sample of residents in the state of California to predict with reasonable accuracy their vehicle preferences.

Introduction & Background

The survey data is comprised of responses pertaining to the economic and demographic background of a sample of Californian residents as well as some information about their current vehicle of choice such as:

- ❑ Vehicle Size (Compact, Midsize, SUV)
- ❑ Fuel Type (Gasoline, Diesel, Electric Vehicle)
- ❑ Vehicle Make (Ford, Toyota, Tesla)

The study aims to answer the following questions:

- ❑ Is it possible to accurately predict a consumer's choice of vehicle using the economic and demographic data?
- ❑ Does the economic and demographic data collected have any influence on a consumer's choice of fuel type? If it does, what are the most important features?

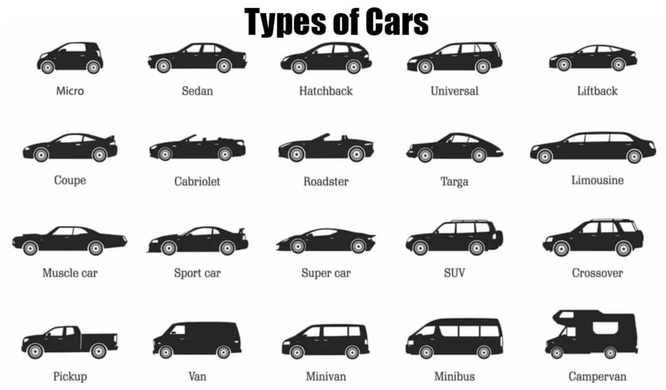

Figure 1: Various Types of Vehicles

Methodology

Workflow

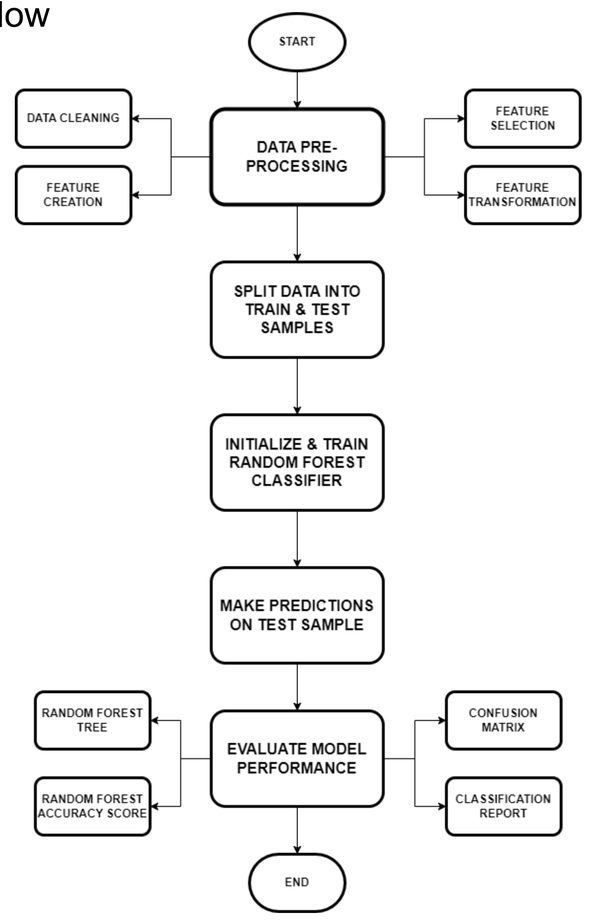

Figure 2: A Flowchart of the Methodology

Data and Results

- ➢ California Vehicle Survey Data (2019)
- ➢ Collected by the California Energy Commission
- ➢ Target Variable: Fuel Type
- ➢ 8042 samples in the dataset

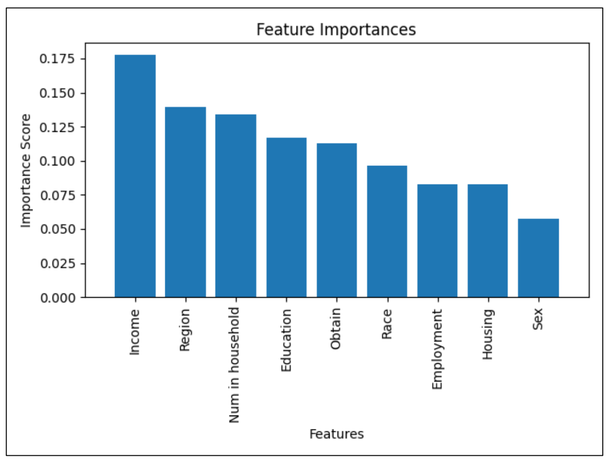

Figure 3: Feature Importance Graph

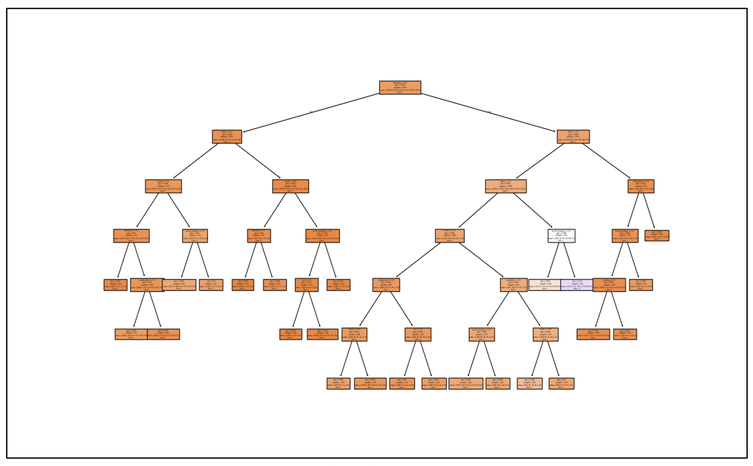

Figure 4: Decision Tree

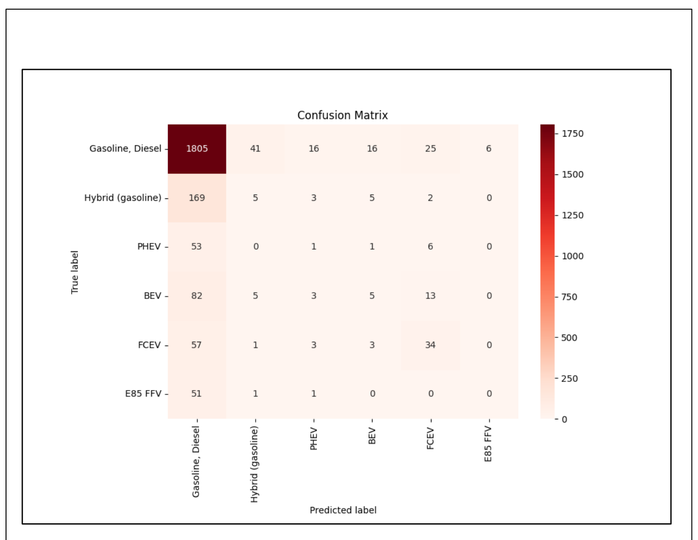

Figure 5: Confusion Matrix

Conclusions

- ❑ The random forest model was able to correct classify the samples according to their respective fuel types with an accuracy score of 0.785 (78.5%)
- ❑ 'income' and 'region' had the most influence on 'fuel type'.
- ❑ This model could be used to predict the vehicle selection preferences of consumers within the same study area, by feeding the model with the relevant information.

Random Forest Accuracy score: 0.7849150435142975

	precision	recall	f1-score	support
Gasoline, Diesel	0.83	0.95	0.88	1956
Hybrid (gasoline)	0.18	0.04	0.07	182
PHEV	0.00	0.00	0.00	60
BEV	0.11	0.05	0.07	74
FCEV	0.44	0.31	0.36	97
E85 FFV	0.00	0.00	0.00	44
accuracy			0.78	2413
macro avg	0.26	0.23	0.23	2413
weighted avg	0.71	0.78	0.74	2413

Figure 6: Classification Accuracy Report

Acknowledgments

We acknowledge the financial support provided by the NSF CREST Center for Multidisciplinary Research Excellence in Cyber-Physical Infrastructure Systems (MECIS) through NSF Award No. 2112650.

References

[1] "Transportation Secure Data Center." ([YEAR]). National Renewable Energy Laboratory. Accessed [10/24/2024]: www.nrel.gov/tsdc.
[2]Mienye, Domor & Sun, Yanxia. (2022). A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects. IEEE Access. Accessed [11/1/2024]