GU
Data Science
& Analytics

DSAN 5300: Statistical Learning

April 30, 2024 *https://github.com/anly512/5300-project-group-6*

Why Americans Travel

Marion Bauman
Georgetown University

Aaron Schwall
Georgetown University

Varun Patel
Georgetown University

Yuhan Cui
Georgetown University

Abstract

INSERT ABSTRACT HERE

Keywords: key, words.

1. Introduction

Americans are in constant transit, utilizing the national interstate and highway system to travel various distances. Whether commuting to work, taking vacations, visiting friends and family, transporting goods and services, or engaging in commerce, people are constantly moving. Understanding U.S. travel patterns provides essential insights into a wide range of fields including environmental protection resource allocation, urban planning, and economic trends. This research aims to understand the motivation behind American travel based on trip specifics. Specifically, we seek to predict the purpose of trips in order to understand the connection between travel details and purpose for travel.

1.1. Relevant Research

A wide variety of industries and stakeholders are interested in understanding travel habits and the usage of American transportation networks.

1.2.

2. Methods

2.1. Data

As a federally funded agency, the Department of Transportation's Federal Highway Administration seeks to understand how Americans are using the national highway system. Since 1969, the FHA has regularly conducted the National Highway Travel Survey (NHTS) in order to collect detailed data on the travel habits of Americans and their usage of federal transit networks (Federal Highway Administration 2022a). From January 2022 through January 2023, the FHA conducted the 2022 NHTS, receiving responses from 7,893 households (Federal Highway Administration 2022b). Survey data includes in-depth information regarding each household member's travel habits across the past 30 days, detailing individual demographic information; household-level demographic information; purpose, location, duration, and mode of transportation for recent travel; and vehicle information (Federal Highway Administration 2022a).

For our research, we chose to utilize the individual demographic information, called the `person` data, along with the trip specific data, called the `trip` data. These two data sources contained 85 predictors that were unique for each observation. We selected these table because 1) 85 predictors was more than sufficient for modeling, and 2) the household and vehicle level data was shared across different observation, creating potential obfuscation in modeling.

2.2. Preprocessing

To prepare the data for modeling, we conducting a thorough cleansing process that included creating human-readable variable names and removing missing or incomplete observations. The final data set was a joined combination between person-level data and trip-level data. After joining, all unique identifier variables and flag variables were removed from the data.

Next, a correlation analysis was conducted to prevent autocorrelation or multicollinearity between variables. An analysis revealed that three variables had correlations over 0.2 with the target variable and were somewhat related in content to the target variable (>0.2). In order to build a more robust model, these three variables, `why_trip`, `trip_purpose_old_schema`, and `reason_for_travel_to` were removed.

Our target variable had five classes for categorizing the purpose of trips. We assessed the balance of the classes in order to ensure our final model would be unbiased. Seven observations with no data for the target variables were omitted.

An analysis of the target variable support shows that the classes are fairly balanced. While there is a distinction between the class sizes, each class has a significantly large sample size to create a fairly unbiased model. Modeling will be assess using balanced metrics in order to verify that final results remain unbiased by class imbalance.

For all modeling, numerical predictors were scaled and centered to normalize predictive influence. Categorical predictors were one-hot encoded, generating binary nomial predictors for each class level. The final cleaned data set of `person` and `trip` data contained 31,050 observations, 71 predictors, and one target variable.

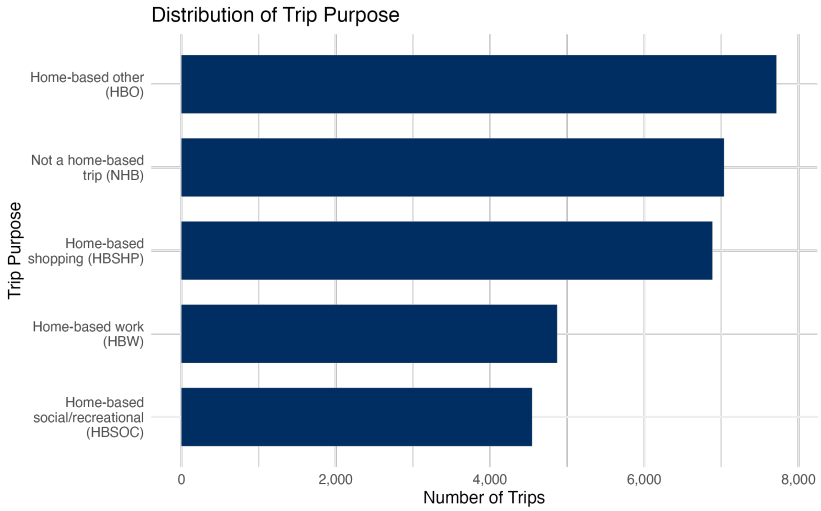

Figure 1: Support of trip purpose, the target variable for modeling.

2.3. Statistical Modeling

In order to predict trip purpose using NHTS data, we trained and assessed five different statistical learning model architectures. We aimed to establish a base line model and then test various methods in order to improve the predictive power of the statistical learning. Our goal was to achieve a high predictive accuracy and a high ROC AUC score, creating a model with strong predictive power and balanced results.

All statistical models were trained with 80% of the data, using k-fold cross validation for hypertuning parameters. Models were hypertuned using grid search techniques in order to find the most appropriate hyperparameters. Final results were assessed on the remaining 20% of the data.

Logistic Regression

In order to establish a base line predictive model, we first trained a logistic regression model on the training data set using Scikit-learn. The logistic model was trained on 22,356 observations of the data. We utilized 5-fold cross validation to hypertune the models. Hypertuning was utilized for regularization penalty type, regularization penalty value, and optimizer values. We utilized a one-vs-rest strategy to account for the multiclass nature of our data, meaning that each iterations trained 5 versions of the model comparing one target class level with all other classes.

Support Vector Machine

Next, a support vector machine was trained on the training data set to classify the data using Scikit-learn. The support vector classifier was hypertuned using 5-fold cross validation. The model was hypertuned for the $L2^2$ regularization penalty value and kernel type. All models were training with a one-vs-rest strategy for the multiclass problem. The final model was

trained to generate probabilistic predictions using Scikit-learn's `predict_proba()` function and a pairwise coupling strategy.

Neural Network

We designed and trained a neural network to predict trip purposes using `keras`. Extensive hypertuning using 5-fold cross validation included tuning for the optimizer's learning rate, hidden unit size, dropout regularization levels, and activation functions. The final model was a linear deep feed forward multilayer perceptron with 2 hidden layers. Both hidden layers have a hidden size of 64 and utilize sigmoidal activation. The output layer uses softmax to output predicted probabilities that the observation belongs in each of the five target classes. The model was trained with a patience level of one epoch and utilized 10% of the training data as a validation set to implement early stopping.

Random Forest

XGBoost

3. Results

What are the results?

4. Discussion

What do the results mean?

5. Conclusion

Conclusion here

Federal Highway Administration. 2022a. "2022 NextGen National Household Travel Survey Core Data." Washington, DC: U.S. Department of Transportation; Available online. http://nhts.ornl.gov.

———. 2022b. "2022 NHTS Frequently Asked Questions." *National Household Travel Survey*. U.S. Department of Transportation; Available online; Federal Highway Administration. https://nhts.ornl.gov/faq.