DSAN 5300: Statistical Learning

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

Why Americans Travel: Predicting Trip Purpose from the National Household Travel Survey

Marion Bauman
Georgetown University

Aaron Schwall
Georgetown University

Varun Patel
Georgetown University

Yuhan Cui
Georgetown University

Abstract

In this study, we examine data from the National Highway Travel Survey (NHTS), which is conducted yearly by the Federal Highway Administration (FHWA) to gather information about household travel habits (Federal Highway Administration 2022a). We examine data from the 2022 survey and train five models in an attempt to accurately predict the trip purpose. The models trained include logistic regression, support vector machine (SVM), nueral network, random forest, and XGBoost. We find that XGBoost performs the best with a test accuracy over 82 percent and a ROC AUC of 0.97. We then discuss the implications of these results and possible future applications of these findings.

Keywords: NHTS, R, XGBoost, SVM, Random Forest, Neural Network, Logistic Regression.

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.

A wide variety of industries and stakeholders are interested in understanding travel habits and the usage of American transportation networks, particularly after the disruption of the COVID-19 pandemic that began in 2020. Research applications include: pedestrian studies, bicycle studies, environmental impact, energy consumption, health, public policty, transit planning, understanding emerging travel modes, and identifying special population groups (Federal Highway Administration 2023). Lu and Giuliano studied the travel habits of different income and race groups across the pandemic, finding that lower income and ethnic minority groups continued to travel for work and shopping purposes. Their model using trip purpose indicated that higher-income and White populations were able to limit travel according to government restrictions, likely due to their higher resources and increased ability to work from home (Lu and Giuliano 2023). Paul's dissertation studies the vehicle-sharing its equity impacts, finding that trip purpose is a significant predictor for whether private vehicles are being shared (Paul 2023).

While extensive research has been conducted into the implications of American travel habits on equity, environment, policy, and more, little research has been done on predicting *why* a person traveled based on trip-level data, without knowing the final destination. Our research aims to fill this gap by understanding the purpose of travel without geospatial information. This modeling task importantly protects individual privacy because it includes no GPS data on travel. Using travel data such as trip length, trip timing, and demographics to predict trip purpose can be useful for understanding why Americans are in transit at a given moment. This could be useful in guiding policy decisions, informing traffic patterns, improving public transportation routes, and address socioeconomic disparities in transit.

2. Methods

2.1. Data

As a federally funded agency, the Department of Transportation's Federal Highway Administration (FHWA) seeks to understand how Americans are using the national highway system. Since 1969, the FHWA 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 FHWA 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 using `tidyverse` 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.

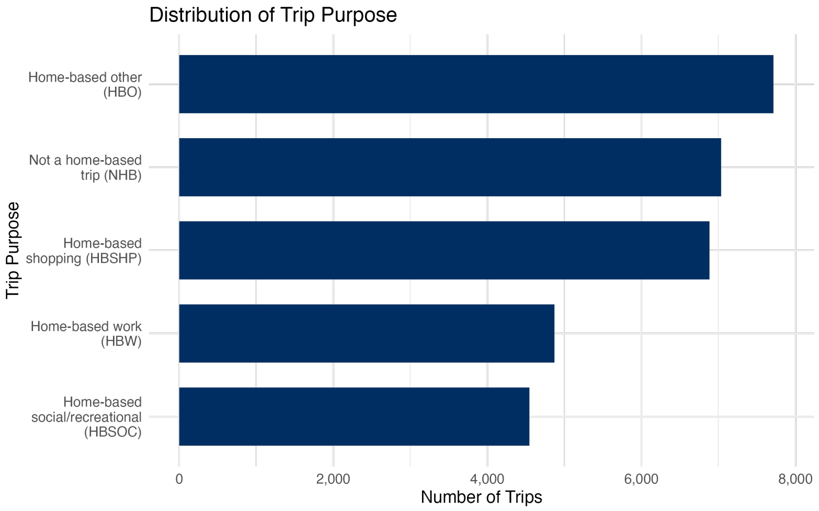


Figure 1: Support of trip purpose, the target variable for modeling.

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.

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. 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. 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. After hypertuning, we found that the best model used no regularization, so the dropout level was 0. 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 two epochs and utilized 10% of the training data as a validation set to implement early stopping. The loss function was categorical crossentropy and the optimizer was Adam.

Random Forest

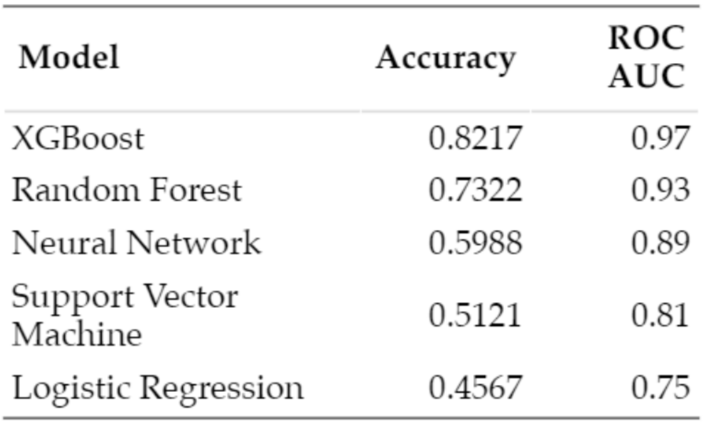
A random forest classifier was utilized to predict trip purpose from the observations. We trained a random forest classifier using 5-fold cross validation to find the best hyperparameters. Hyperparameter search included the maximum depth of branches, minimum number of samples requires to split a leaf, and the number of trees (estimators) in the forest. The best model was selected as the model with the highest accuracy.

XGBoost

Finally, an XGBoost classifier was trained to predict trip purpose from our training data. The XGBoost model was hypertuned for maximum tree depth, learning rate of the optimizer, number of trees (estimators), the minimum loss reduction required to split a leaf (also called gamma), the subsampling ratio for preventing overfitting, and the subsample size of the columns used by each tree. XGBoost was hypertuned using 3-fold cross validation and a random grid search.

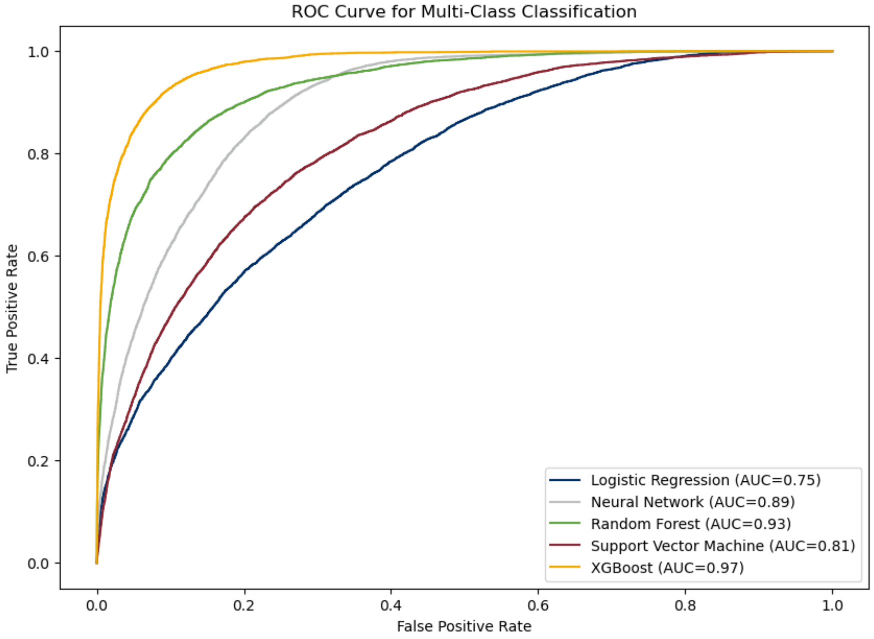
3. Results

We observe that the XGBoost machine learning algorithm performed the best with a test accuracy of 82.17% followed by a Random Forest model which had an accuracy of 73.22%. A neural network model was also fit to the data but it only resulted in an accuracy of 59.88%, followed by Support Vector Machines and Logistic Regression which had accuracy rates of 51.21% and 45.67%.

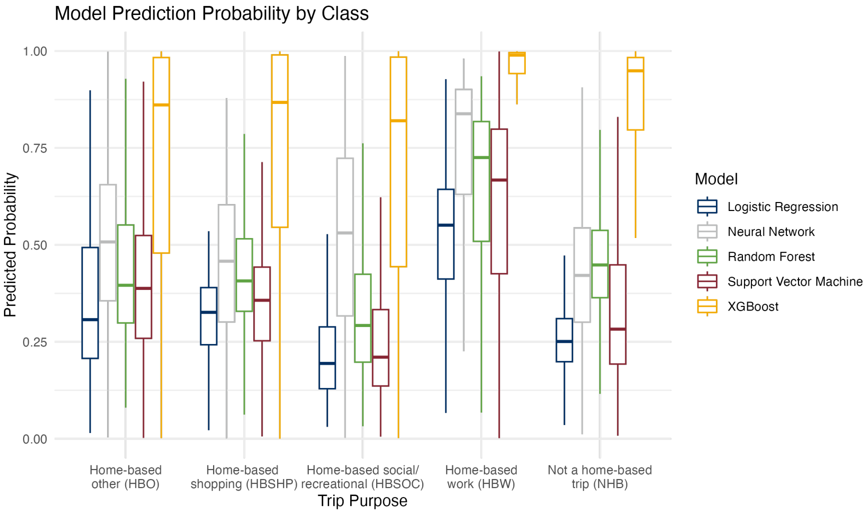


Model	**Accuracy**	**ROC AUC**
XGBoost	0.8217	0.97
Random Forest	0.7322	0.93
Neural Network	0.5988	0.89
Support Vector Machine	0.5121	0.81
Logistic Regression	0.4567	0.75

We also observed the ROC AUC curve to be the greatest for XGBoost model-0.97, followed by Random Forest which had the ROC AUC curve of 0.93. Overall, XGBoost model performed the best out of all the five models that were trained on the data.



We also observe the predicted probability for different classes for different models that were trained. As observed, the mean predicted probability for each class for the XGBoost model appears to be between 0.75 and 1.00 which displays the high likelihood that the data point belongs to that particular class. Almost all other algorithms perform much poorly compared to XGBoost



4. Discussion

Among all the models considered, XGBoost stands out as the best performer, followed closely by Random Forest. Both XGBoost and Random Forest are ensemble methods that have the ability to achieve low variance and bias by leveraging weak estimators to build strong classifiers. However, XGBoost exhibits certain advantages over Random Forest. XGBoost incorporates a gradient boosting framework that enables it to handle more complex relationships within the data. This allows XGBoost to learn complex patterns and interactions within the data and achieves a better predictive performance.

Neural networks also have the potential to capture complex patterns within the data. However, they come with a trade-off in terms of their computational intensity. Neural networks require significant computational resources and time to train, especially when dealing with large datasets or complex architectures.

SVM and logistic regression, while being widely used classification algorithms, did not perform as well in this context. The complexity of the underlying data seems to pose a challenge for these models. SVM relies on finding optimal hyperplanes to separate data points, and logistic regression assumes a linear relationship between the features and the target variable. These assumptions may not hold true for the given dataset based on the result.

One notable advantage of XGBoost and Random Forest is their interpretability, or XAI (explainable artificial intelligence). These models provide insights into feature importance, allowing us to understand the factors driving the estimation of travel purposes. From the analysis, it is evident that the "reason_for_travel" and "time_at_destination" features hold the most significant importance in estimating the travel purpose. This aligns with our intuition as these features likely contain patterns that strongly correlate with determining the travel purpose. That being said, the features were checked against trip_purpose for multicollinearity and none was found. It is important to note that while the feature names trip_purpose and reason_for_trip sound similar, they are measuring different things. For example, "Buy meals" appears frequently in the reason_for_travel feature, but it appears in home based shopping, social and recreational, and not home based shopping. In contrast, interpreting neural network models and SVMs can be challenging due to their more complex internal workings and lack of feature importance measures.

5. Conclusion

Based on this study, we can see that tree-based methods, especially XGBoost, have the best results for predicting trip purpose from NHTS data in 2022. The most important features in this model include the reason for travel, worker status, the trip destination, and whether the trip occurred over the weekend. Our best test accuracy from the XGBoost model was 82.17%, which is nearly double the best test accuracy achieved by the baseline logistic regression model. Our ability to predict trip purpose with high accuracy and ROC AUC suggests that trip purpose is explainable by trip characteristics and demographic information.

There are many potential future applications that can be based on the findings in this study. One possible future application would be to move from using survey data to observed data. Similar data could be collected by taxi or ride sharing companies like Uber or Lyft, and the predicted trip purpose could be used to to better determine how many cars will be needed in certain areas. Similar trip characteristics and demographic information can be collected by many other types of businesses and organizations from hotel chains to environmental researchers. Being able to accurately predict the purposes of trips will have significant impact for hotel chains by allowing them to predict where more capacity will be needed and how specific hotels can be expected to perform. Similarly, accurate trip purpose predictions will allow climate researchers to determine where to target efforts for future green house gas emission reductions. For example, places and trip modalities that are mostly used for work are unlikely to be easily changed, while trip modalities that are used mostly for recreation may be more easily targeted for reductions.

Lastly, there are several ways in which the models examined in this study may be enhanced in the future. Our full dataset contained over 60 different features to be used by our models. By applying a feature selection or dimensionality reduction algorithm like PCA or t-SNE we could remove the features with the lowest importance to our final models. This may allow the models to perform better. One area of future research would be to examine how adding past NHTS survey data effects the predictions. This would give us a better idea of how reliable our predictions may be in the future based on similar survey data.

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.

———. 2023. "National Household Travel Survey Compendium of Uses." Published by Federal Highway Administration [Online]. https://nhts.ornl.gov/assets/2023_compendium.pdf.

Lu, Yougeng, and Genevieve Giuliano. 2023. "Understanding Mobility Change in Response to COVID-19: A Los Angeles Case Study." *Travel Behaviour and Society* 31: 189–201. https://doi.org/https://doi.org/10.1016/j.tbs.2022.11.011.

Paul, Julene. 2023. "Sharing in and Sharing Out: The Equity Implications of Informal Vehicle-Sharing." UCLA.