Travel Mode Choice Prediction

1. Goal

Understanding and predicting the travel mode choices of citizens are essential in transportation planning and policy-making in cities. An accurate machine learning model to predict the travel mode with limited available information would be helpful for the public and private sectors. Also, understanding which features are important is useful because it could make our cost of future data collection less expensive. We framed this problem as a classification task of each trip's travel mode in the latest national survey. We evaluated the prediction model by accuracy, precision, recall, and F-1 score.

2. Related Work

Hichem Omrani (2015) used Artificial Neural Net-MLP, Artificial Neural Net-RBF, Multinomial Logistic Regression, and Support Vector Machines to predict the travel mode in the city of Luxemburg. Julian Hagenauera and Marco Helbich (2017) used Multinomial Logistic Regression, Naive Bayes, Support Vector Machines, Artificial Neural Network, Classification Trees with Bagging, Classification Trees with Boosting, and Random Forest to predict the travel mode in the Netherlands.

These studies only used accuracy as a performance measure. We aimed to achieve the F-1 score that is sufficient for practical use in addition to accuracy. Also, the number of features in our dataset was much more prosperous than the datasets used in these studies. We aimed to understand the feature importances at a more granular level.

3. Dataset

The dataset for this project is National Household Travel Survey conducted by the Federal Highway Administration (FHWA) in 2017. It includes 923,572 daily non-commercial travels by all modes, including characteristics of the people traveling, their household, and their vehicles. The data were collected from a stratified random sample of households based on their address in all states in the US. In addition, the data was augmented by samples from 13 add-on areas; Arizona, California, Dallas Fort Worth TX, Des Moines IA, Georgia, Maryland, New York, North Carolina, South Carolina, Texas, Tulsa OK, Waterloo IA, and Wisconsin.

The data is separated into four files as below, and they can be merged with each file's primary key. There are 244 unique features in total. The major features are listed below.

1. Trip: mode of transportation, trip purpose, date and time, origin and destination, distance
2. Household: income, number of workers, housing type, neighborhood characteristics
3. Vehicle: make, model, age (year)
4. Person: gender, age, driver and worker status, annual miles

4. Data Preprocessing

Our target variable (TRPTRANS) consists of the 24 travel modes in total, including walk, bicycle, motorcycles, car, SUV, van, amtrak, subway, boats, airplanes. We grouped the modes into two labels; 1 for environmentally friendly modes (walk, bicycle, bus, amtrak, and subway) and 0 for other motorized vehicles. The target is highly skewed; environmentally friendly modes account for only 10% of the data records. Later in this report, we introduce several approaches to deal with data skewness.

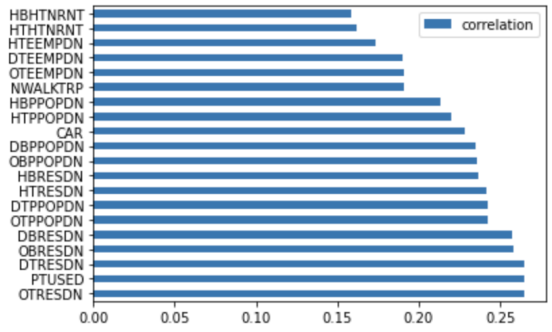
The dataset contains features that are the result of the travel mode choices, which cause target leakage in the prediction model training process. Also, the dataset includes features only used for identification. We removed those 75 features in total and kept 169 features for the further process.

5. Exploratory Data Analysis

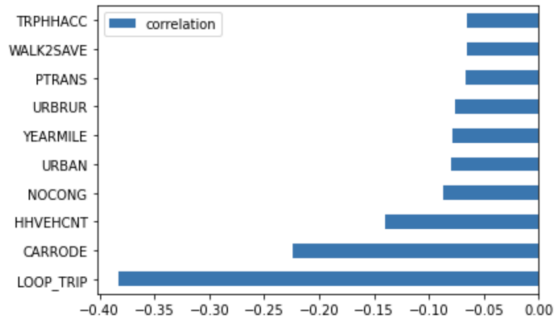
In order to understand the dataset even further, we tried to see the characteristics at the state level by anomaly detection, principal component analysis, and clustering.

5.1. Feature selection and preprocessing

We found that using all features for EDA made interpretability worse and clustering inaccurate. Therefore, we focused on hypothetically important features by computing the correlation with the target. 22 features with a correlation of more than or less than 0.15 were selected.



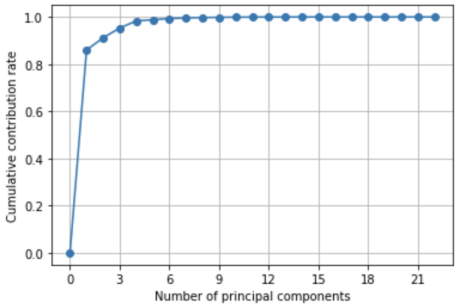
Features with Top Positive Correlation



Features with Top Negative Correlation

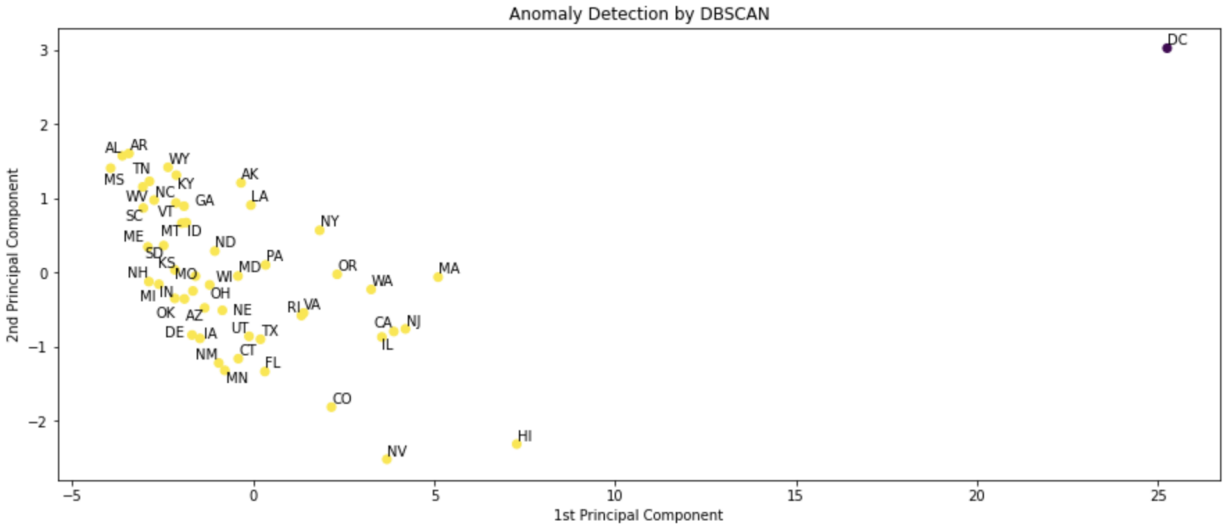
Then, we grouped the trips to the state level computing the average of each feature. We also applied StandardScaler so that all features have the same scale. The resulting table had 51 rows × 22 columns.

5.2. Anomaly detection and Principal Component Analysis



We applied DBSCAN with eps = 5 and min_samples = 3. For visualization, we also conducted Principal Component Analysis (PCA). The left chart shows the number of

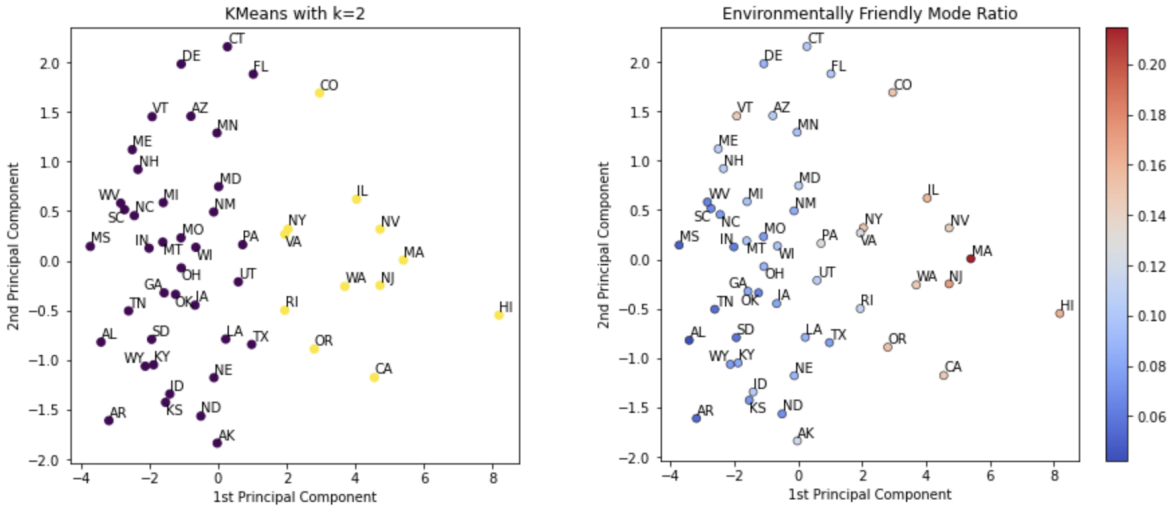
principal components and the cumulative contribution rate. The contribution from the first principal component is 86%, and the cumulative contribution from the first and second principal components is 91%. The below scatter plot is the result of DBSCAN with the first and second principal components.



Among them, only Washington, DC was not assigned to the cluster (purple). High population and housing density are the cause of being an anomaly. We excluded DC for the next step, clustering of states.

5.3. Clustering of states

We used KMeans with the average Silhouette Coefficient to determine the best number of clusters. The best number of clusters was two, with an average Silhouette Coefficient of 0.48. We conducted PCA again excluding DC, and visualized the corresponding clusters with the first and second principal components as below. The boundary between the two clusters is located at around 1.5 of the first principal component. We separately computed the mean value of the target in each state that is equivalent to the ratio of environmentally friendly modes to all trips. The higher value represents more dependence on environmentally friendly modes.



The same trend can be seen as clustering; states are roughly divided into two groups at 1.5 of the first principal component. We concluded that the states associated with the yellow clusters rely more on environmentally friendly modes. At the same time, we assumed that the wide variety of states would make it challenging to build the comprehensive prediction model with all states. Therefore, we decided to focus on New York State for modeling.

6. Modeling

The data for NY (923,572 records with 169 features) was split into the training set and the test set by 70%:30%. Categorical features were converted to dummy variables. We aimed to create the best performance model with Neural Network. Also, Random Forest was used for ensuring interpretability, and Bayesian Networks were used to estimate the causal relationship.

6.1. Neural Network

A neural network was built to improve the accuracy of our model, It was a 5 layered network built using Keras with a batch normalization after each layer, without which the accuracy actually decreased by 30%. To stop overfitting, used a dropout of 10-20% after every layer. After every layer a non-linear Relu:max(0,x) was used as an activation function. However, to get the final probability, we used the sigmoid function at the final layer. As the data was skewed, we used metrics- Accuracy, Recall, Precision and F1 Score, with the latter being the best judge of model as it takes both Recall and Precision into account. We used two loss functions-Binary Cross-entropy and F1 score too, which helped in improving the Accuracy and Recall.

$$Loss = -\{y_{target} \log y_{pred} + (1 - y_{target}) \log(1 - y_{pred})\} + 0.6(1 - F1)$$

The Neural Network was trained using ADAM algorithm, which takes into account both the exponential decay and momentum of past gradients.

$$m = \beta_1 m + (1 - \beta_1) g \qquad Lr: Learning\ Rate$$
$$v = \beta_2 v + (1 - \beta_2) g^2 \qquad \omega: Weight$$
$$\omega = \omega - Lr * \hat{m} / \left(\sqrt{\hat{v}} + \varepsilon\right) \qquad g: Gradient$$

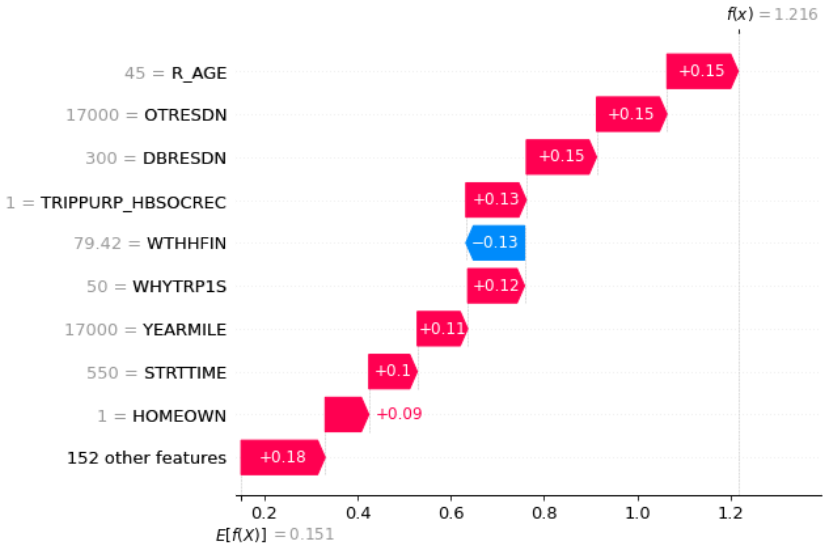
Interpretation using SHAP

We used a shapely value to get the interpretability of the Neural Network model, which is given by -

$$\phi_i(v) = \frac{1}{|N|!} \sum_R \left[v(P_i^R \cup \{i\}) - v(P_i^R) \right]$$

Φ: Shapley value
N: Number of player (feature)
P_i^R: Set of player with order
$V(P_i^R)$: Contribution of set of player with order
$V(P_i^R \cup \{i\})$: Contribution of set of player with order and player i

SHAP uses the average shapely value for a given batch in order to understand the importance of each feature.



Results from SHAP

The above image shows that for a batch with average age 45, the probability of using an environmentally friendly mode increases if the average age is 45.

6.2 Random Forest

We trained the baseline Random Forest model with GridSearchCV. The best parameters were n_estimators = 26, max_depth = 15, min_samples_split = 10, and max_leaf_nodes = 50. Additionally, we took two different sampling approaches to deal with the skewed data.

Downsampling

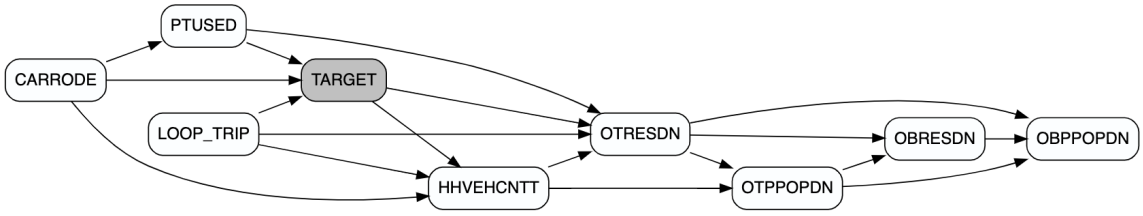
We randomly sampled the training data with the negative target to make the positive and negative targets equal. 12,782 records were selected out of 84,155 records with the negative target.

Oversampling

We synthesized new records by Synthetic Minority Oversampling Technique (SMOTE) with the imbalance-learn library. 58,880 records were added to the training set.

6.3 Bayesian Networks

Due to the computational limitations, eight features that have high correlations with the target were used for the Bayesian Networks model. We first learned the structure by HillClimbSearch with BicScore and estimated the parameters by BayesianEstimator. Interestingly, features related to the population and housing density (OBRESDN, OBPPOPDN, OTRESDN, OTPPOPDN) that we found important in the previous Neural Network model are estimated as the result of the target rather than the cause.



7. Results

7.1 Classification Performances

The below table shows the classification performances of all approaches. Neural Network with the optimized loss function was the best model, and we could achieve more than 0.8 of the F1 score.

Model	Accuracy	Precision	Recall	F1 Score
Neural Network (Baseline)	0.94	0.81	0.77	0.79
Neural Network (Loss Function)	0.95	0.83	0.81	0.82
Random Forest (Baseline)	0.91	0.87	0.47	0.61
Random Forest (Downsampling)	0.88	0.58	0.77	0.66
Random Forest (Oversampling)	0.91	0.87	0.47	0.61
Bayesian Networks	0.91	0.78	0.55	0.64

7.2 Feature Importances

The below table shows the top feature importances from different approaches. In addition to Neural Network and Random Forest, we created 2 clusters of trips by KMeans except for the

target variable and computed the difference of each cluster's feature average. We thought the more significant difference could be a proxy for feature importances.

Model	**Neural Network**	**Random Forest**	**KMeans (k=2)**
Method	SHAP	Impurity-based	Difference of clusters' averages
1	Age	Count of household vehicles	Loop trip
2	Housing density of origin	Loop trip	Housing density of origin
3	Housing density of destination	Housing density of destination	Count of public transit used in the past
4	Home-based trip	Frequency of personal vehicle used	Housing density of destination
5	Trip weight	Housing density of origin	Population density of origin

The density of housing and population in origin and destination were commonly critical in all models. Loop trip, which indicates the trip has the same origin and destination, was significantly used by Random Forest and Kmeans. Compared to Random Forest and KMeans, Neural Network captured feature importances uniquely. Neural Network utilized age, home-base trip flag, and trip weight, which is essential for travel mode choice but is not well captured by Random Forest and KMeans.

8. Conclusion

We could achieve high accuracy and F-1 score by Neural Network with the optimized loss function. We believe that this model's performance is sufficient for practical use, such as transportation planning and policy-making. Additionally, we could identify features that are commonly important and specific for a particular model. Based on this insight, we can optimize future data collection for travel mode choices.

Our possible future extension could be creating multiple classification models to predict the individual travel mode rather than the binary classification.