NATIONAL RENEWABLE ENERGY LABORATORY

Predictive Analysis of Public Transportation Usage

Michael Moniger

GOAL

Given a wide variety of geographic and household features, I intend to build a predictive model to determine the likelihood that a given individual will utilize available public transportation.

DATA

The datasets, which can be found here, are from NREL's Transportation Secure Data Center (TSDC). Spanning fifteen US states between the years of 2001 and 2017, they consist of travel surveys (also called travel diaries) from over 300,000 individuals, collecting demographics and detailed travel behavior across each states, focusing primarily in urban areas. A small sample of the numerous features within the dataset include vehicle-acquisition decisions, parking choices, work schedules and flexibility, etc.

Several datasets include sub-samples of participants who utilized either wearable or vehicular GPS devices, the latter of which includes on-board diagnostic sensors to provide both drive cycle data and second-by-second vehicle speed profiles. Those participating in the wearable add-on part of the study were chosen either randomly or because their travel diaries reported use of public transit in day-to-day travel (depending on the state conducting the survey).

Additionally, I have applied for access to the TSDC's latitude and longitude spatial data to more precisely map departure and arrival locations. This can be combined with the EPA Smart Location Database (found here), which contains metrics like proximity to and density of transit stops, urban access, and walkability indices.

PROJECT PROGRESSION

Minimum Viable Product

Combine the datasets from all available states.
Build a predictive model to determine the likelihood that a given individual will utilize available public transportation using the California Household Travel Survey.
Determine which features contribute most to prediction.

Improvement 1

Extend this predictive model to include nationwide travel surveys.

Improvement 2

Overlay the NREL and EPA latitude and longitude spatial datasets to include these features in the model.