On The Rails (Predicting Public Transit)

Michael Moniger
mdmoniger@gmail.com | ln/michael-moniger | github.com/mdmoniger

NATIONAL RENEWABLE ENERGY LABORATORY

Background

According to the EPA, a typical passenger vehicle emits about 4.6 metric tons of carbon dioxide per year[1]. This amounts to annual emissions of over 1.1 billion metric tons of CO_2 emissions nationwide.

NREL's Transportation Secure Data Center provides public access to a variety of travel surveys and studies. Here, I focus on the California Travel Survey to explore behavior that may influence the adoption of more energy-efficient travel behavior, such as using public transit.

The Data

500,000 travel surveys with hundreds of geographic and demographic features; in addition, I engineered many features, including trip duration.

Objectives

- *Build a predictive model to determine the likelihood that a given trip will utilize available public transit.*
- *Determine which features most strongly predict whether an individual trip will be serviced by public transit.*
- *For the most important features, determine the magnitude and direction by which each is related to mode of transportation.*

Methods

After adjusting for a massive class imbalance, I trained and compared three boosting algorithms, then grid searched over each to determine the hyperparameters that provided the highest F1 score.

I then extracted feature importances from the best model and plotted partial dependence plots to visualize the relationship between feature and target variables.

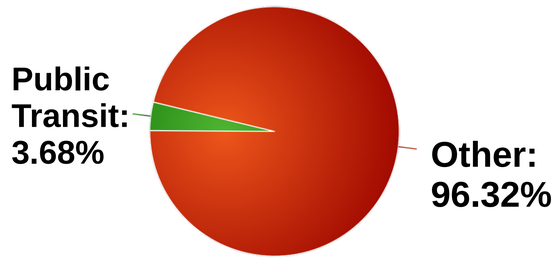

Figure 1. Class Imbalance

Results

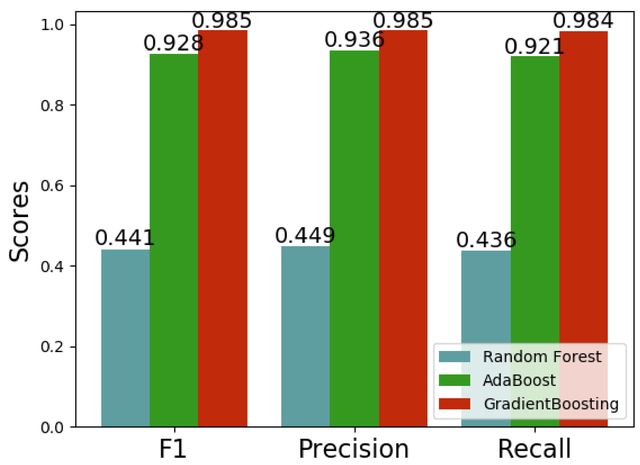

Figure 2. Model Scores

In terms of all three scoring metrics, GradientBoosting algorithm outperformed both Random Forest and AdaBoost.

Results

Table 1. Feature Importances

Feature	Score
Departure time	0.392
Arrival time	0.264
Trip duration	0.120
Activity duration	0.090
Activity count	0.072
Air trip distance	0.036
Trip distance	0.026

Each score represents the proportion of information gain attributed to its respective feature.

Discussion

Partial dependence plots visually represent the "shape" of the relationship between feature and target variables.

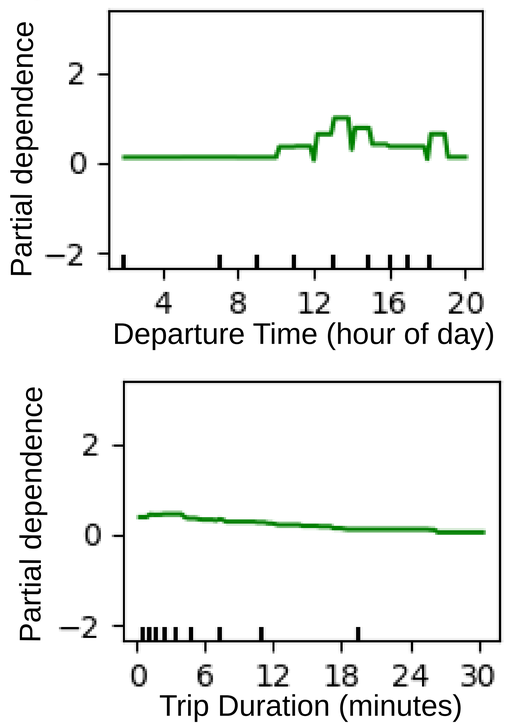

Figure 3. Partial Dependence Plots

Conclusion

Since partial dependence plots indicate stronger effect on the target variable when the y-axis is farther from 0, one can see that the most important features are time-dependent. The first plot shows that public transit is more widely used between the hours of 12:00 and 3:00 PM, while the second indicates more use for shorter trips. Therefore, to reduce carbon emissions, we should focus primarily on incentivizing and streamlining short intracity systems, such as bus routes and light rail services.

Stack

python™ · SciPy · scikit learn · NumPy · matplotlib · ANACONDA · amazon web services

References

1. Sources of Greenhouse Gas Emissions. EPA. https://www.epa.gov/ghgemissions/sources-greenhouse-gas-emissions

2. Transportation Secure Data Center. National Renewable Energy. https://www.nrel.gov/transportation/secure-transportation-data/