Final Project Stage 5: Predicting Neighborhood Walkability

PPOL 5204

Katharyn Loweth

05.05.2024

Background

Where an individual chooses to live in a metropolitan area is a complex decision. It depends not just on the housing unit itself but the community in which it is a part of. When contemplating a neighborhood, people may ask themselves “How close is the grocery store? How long will it take me to get to work? Is there a metro/bus stop nearby? Does it feel safe to walk around this area?”. A neighborhood’s walkability, or how easy it is for someone to live and walk/bike/use public transportation around the area, is an important aspect for both individuals and urban development.

Multiple studies have shown the positive benefits of walkability. Individuals who live in more walkable communities are more likely to walk to a destination (like a store), walk for leisure, and walk or bike to to work, all of which can support their individual health and reduce environmental pollution (Thomas & Rourk Reyes, 2021; Watson et al., 2020; Kim et al., 2020). Additionally, walkable communities have been found to have positive effects on individual’s social health because they are more likely to have more social interactions with others and develop a sense of community with their neighbors (Thomas & Rourk Reyes, 2021; Carson et al., 2023). Making areas more walkable can also have positive economic effects on the area. For example, in New York City, new bike and bus lanes increased retail sales and reduced commercial vacancy in these locations (Litman, 2023).

While walkability has several positive impacts on the local community, it is also important to consider which sociodemographic groups live in these walkable areas. Conderino et al. (2021) studied the relationship between walkability and demographic factors in cities across the United States, and found an inverse relationship between income and neighborhood walkability. They found that neighborhoods with higher average incomes were more likely to have low walkability scores (2021). Conderino et al. (2021) also conducted a subgroup analysis based on the neighborhood’s majority racial group and found this pattern in all racial groups except for African American-majority neighborhoods, who were more likely to live in walkable communities as their income increased.

In this research project, I am examining what demographics and transportation habits can predict walkability, specifically the Environmental Protection Agency (EPA)’s National Walkability Index. Initially published in 2017 and updated in 2021, the EPA created the National Walkability Index (NWI) to consistently compare employment and built environment variables for every census block group (CBG) in the United States (Chapman et al., 2021). The EPA strictly defines walkability as a measure of local infrastructure and calculates it using 3 infrastructure metrics: street intersection density, proximity to transit stops, and diversity of the land uses. Notably, this is a different walkability metric than used by Conderino et al. (2021), who used a commercial metric for walkability (WalkScore) that considers distance to different amenity types in its calculation (“Walk Score Methodology”, n.d.).

This project would contribute to the growing knowledge on the relationship between a community’s walkability and the people who live in the community. If a model based only on demographic and transportation habits is able to accurately predict the EPA’s calculated walkability of the CBG, then it would provide further evidence of a relationship between walkability and community population characteristics.

Research Question

What demographic features and transportation habits predict neighborhood walkability?

Data Preprocessing

The data for this project is primarily coming from two sources: 1) the Environmental Protection Agency (EPA)'s Smart Location Database, and (2) the U.S. Census Bureau's American Community Survey (ACS), an annual survey on American households and workforce. Both of these datasets are available at the Census Block Group (CBG) level and utilize a 12-digit GEOID as the key indicator. The Smart Location Database and the ACS Datasets contain 220,741 observations. Other data used in this project includes the 2020 Decennial Census Population.

From the **Smart Location Database**, I am focusing on the following variables:

- **National Walkability Index (NWI)**: In the dataset, the NWI is a float variable on a scale of 1-20.
- **TotalPop**: The estimated total population in the CBG. Notably this is the ACS 2018 estimate of the CBG population, which is based on the 2010 Decennial Census. The total population is an integer variable
- **Percent of Workers considered low-wage workers**: This variables provide context on the sociodemographics of the area and who is living within the CBG. The percentage is a float variable on a scale of 0 to 1.
- **AutoOwn variables**: There are 6 AutoOwn variables in the dataset that provide context of the number of people with cars in the CBG. For the analysis I will be using the percent version of them (float variable) that are on a scale of 0 to 1.
  - PCT_AO0: Percentage of households in the CBG who do not own a car.
  - PCT_AO1: Percentage of households in the CBG who own 1 car.
  - PCT_AO2: Percentage of households in the CBG who own 2+ cars.

From the **ACS**, I am utilizing the 2022 5-year estimates. For some of the features I am provided the median value, whereas for others I am provided the total count of people who selected each answer. I am focusing on the specific demographic features and responses related to commuting habits such as:

- **Median Age:** In the ACS dataset, Median Age is saved as a string variable type.

- **Median Household Income:** In the ACS dataset, Median Age is saved as a string variable type.
- **Racial/Ethnicity Demographics**: The racial demographics options I am including are White, Black, Asian, American Indian/Alaskan Native, Hawaiian/Pacific Islander, Other race, and Two or More Races. I am also including information about the Hispanic/Latino population. In the ACS, these variables are integers because they represent counts.
- **Commute Methods**: The commute methods options include car, public transportation, walking, and bike/motorcycle. In the ACS, these variables are integers because they represent counts.
- **Commute Duration**: The commute duration options include less than 10 minutes, 10-14 minutes, 15-19 minutes, 20-24 minutes, 25-29 minutes, 30-34 minutes, 35-44 minutes, 45-59 minutes, and more than 60. In the ACS, these variables are integers because they represent counts.
- **Education Attainment**. Similar to the other demographic characteristics, I added in information related to percent of adults over the age of 25 who have achieved no education, HS diploma, Associate's Degree, Bachelor's Degree, Master's Degree, and Doctoral Degree for each census block group.

As stated previously, the total population variable in the Smart Location Dataset is based on the 2018 ACS 5-year estimates from the 2010 decennial census. Therefore I added the **2020 Decennial Census population** data to provide a more recent estimate of the Census Block Group's (CBG) population. I also calculated the difference between these two population estimates to discern CBG that had significant change in the time period.

Preprocessing:

Data cleaning has primarily consisted of:

- Filtering the EPA Smart Location Dataset so that it only contains CBGs within the 50 most populous metropolitan areas in the United States. I did this by utilizing the U.S. Census' list of Metropolitan statistical areas, sorted by population, and filtered the first 50 based on the top MSA ID variable (which is included in the Smart Location Dataset). This filtering method cut the dataset roughly in half to around 110,000 observations.
- Reformatting the Census ACS datasets (each topic was its own dataset) so that they were prepared to merge with NWI_Top50. This included isolating the CBG GEOID as its own variable and renaming column headings.
- Joining the NWI_Top50 and the ACS datasets on the CBG GEOID variable
- Transforming demographic and commute integer variables into percentages based on the total_count variables so that they are standardized. This will help later on in the analysis.

- Transforming the NWI Index into a categorical variable based on the EPA's guidance (Chapman et al., 2021). By categorizing the variable into four categories will help with the analysis later on.
    - NWI score of 1.00-5.75: least walkable community
    - NWI score of 5.76 - 10.00: below average walkable community
    - NWI score of 10.51 - 15.25: above average walkable community
    - NWI score of 15.26 - 20.00: most walkable community.

Initial Data Exploration/Data Descriptives

Data Cleaning Issues

The biggest issue within the EPA dataset is that for the majority of observations, the CBG key identifier variable was corrupted by Excel. Because CBG's GEOID are 12-digit variables, Excel wants to convert them into scientific notation. It is unclear how the issue occurred, but in the dataset on the Data.gov website, only CBGs whose state has a state GEOID with a value less than 10 have an accurate GEOID; otherwise the number's specific value has been lost (for example: 123456789012 has become 123450000000). I reconstructed the IDs based on the U.S. Census Bureau's guidance on the formulation of CBG GEOIDs and using the Smart Location Dataset's other GEOID variables – for state, county, census tract, and census block group ("Understanding Georgraphic Identifiers", n.d.). This process salvaged many of the IDs, but still reduced the dataset as part of the join. After joining the datasets together, the total sample reduced to approximately 70,000.

Other issues in the data include missing values or values imputed in such a way that makes it difficult for analysis. For example, missing values in the ACS datasets were initially represented by a "-" and were not being picked up as missing in exploratory data analysis. I replaced all missing values in the dataset with the universal NaN for standardization. The Median HH income dataset is particularly full of odd codings such as "250,000+" as the maximum value in the data. However, the maximum number of missing values for any column is less than 1,000, which is small compared to the overall size of the dataset.

Missing Values

To deal with missing values in the dataset, I imputed values for the "Median_HH_Income", "Median_Male_Age", and "Median_Female_Age" and dropped observations for the missing values for the other variables. For these 3 columns, I determined the median value for each metropolitan area ("CBSA_Name") and assigned accordingly. I chose not to input for the other variables because they are mostly the percentage variables, which I created in the previous stage, that are highly related. It would be difficult to imput these values, and the number of observations with these missing values is less than 2% of the dataset. After dealing with the missing values, the dataset includes **68,941 observations.**

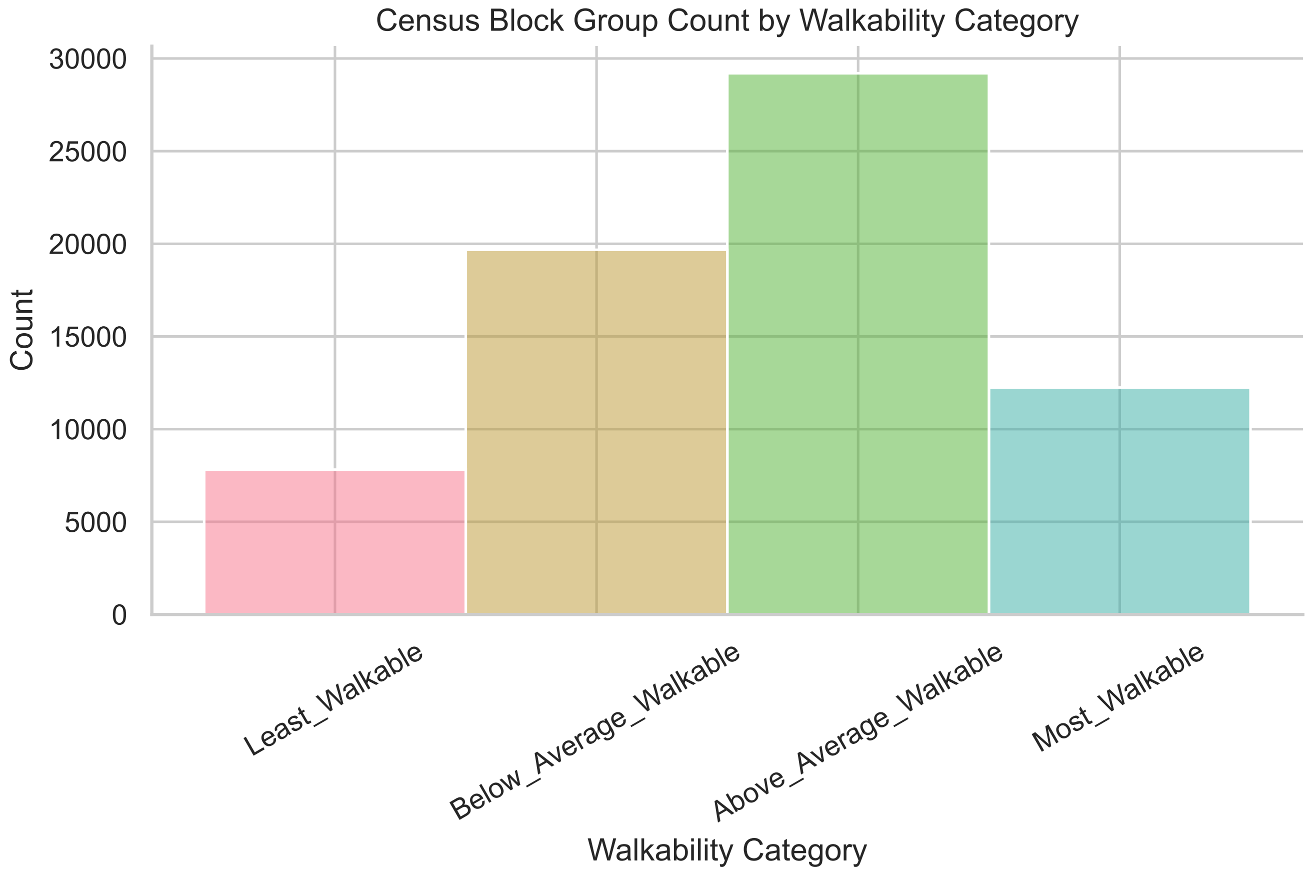
After removing missing values, I reviewed the descriptive statistics for key variables to understand how they were distributed throughout the dataset.

Table 1: Descriptive Statistics of Key Variables

Variable	Description	Count	Mean	Std.Dev	Median
NatWalkInd	NWI score	68941	11.32	4.02	12
NWI_category	NWI score grouped into the EPA's categories of walkability	Least Walkable: 7816 Below Average: 19678 Above Average: 29203 Most Walkable: 12244	N/A	N/A	N/A
White_perc	Percent of CBG population that identifies as white	68941	.605	.295	.68
Black_perc	Percent of CBG population that identifies as black	68941	.18	.26	.06
R_PCTLOWWAGE	Percent of low wage workers in CBG	68941	.23	.05	.22
PCT_AO2p	Percent of households that own 2+ cars	68941	.53	.24	.56
N_car_perc	Percent of CBG population who report commuting to work via car	68941	.84	.22	.93
N_pubtrans_perc	Percent of CBG population who report commuting to work via public transportation	68941	.1	.18	.02

As seen in the graph below, there is class imbalance in the target feature, and there are similarities across classes (i.e., both Least Walkable and Below Average Walkable represent less-walkable communities). This imbalance and overlap may be something that impacts the accuracy of our model.

```
<Figure size 1800x1200 with 0 Axes>
```



Creating New Variables

As part of exploratory data analysis, I created pairplots between features and specifically with “NatWalkInd” to discern any interaction between variables and non-linear relationships between features and the continuous version of the target variable. Based on these plots, I created 10 interaction variables and 9 squared variables as part of the dataset to capture these aspects in my analysis. For example, I created variable “Pct_AO0_sq”, a squared version of “Pct_AO0” because the scatterplot of the feature versus “NatWalkInd” had a quadratic appearance. The pairplots and scatterplots are included in the Stage_3 python notebook.

Factorizing Dependent Variable

Other data cleaning steps that I took include the standardization of all continuous variables using StandardScaler(), and transforming “CBSA_Name” variable, which contains the names of all 50 metropolitan areas included in the dataset, into dummy variables. Additionally, I converted the “NWI_category” variable into a numeric categorical variable named “NWI_cat”. This categorical variable is on a scale of 0-3 with the following code. I did this because it will support the better supporting the modeling process.

- 0 = “least walkable”
- 1 =“below average walkable”
- 2 = “above average walkable”
- 3 = “most walkable”

Modeling and Analysis

The dataset used for analysis initially included 113 features. The identifier variables and the total count variables were removed from the datasets, resulting in the feature dataset including 105 features. Four modeling techniques were used: Decision Tree, K-Nearest Neighbor (KNN), Random Forest, and XGBoost. These methods were utilized because they are nonparametrics models that can capture non-linear relationships between the model features and target variable. Additionally, they are all equipped to perform multiclass classification.

Train-Validation-Test Split

The NWI dataset is split into 3 groups: training, validation, and test. The training dataset consists of 70% of the clean dataset (roughly 48,000 observations), and both the validation and test set consist of 15% of the dataset (roughly 10,300 observations each).

Results

Models

For each of these model types, I fitted the training data on (1) the model with default hyperparameters, and (2) with tuned hyperparameters determined using a gridsearchCV. For all of the models, I used a 5-fold cross validation. For the four models described below I did not weight the target feature class imbalance.

Decision Tree

For tuning my decision tree model, I used a gridsearchCV to select the optimal **criterion** (“gini”, “entropy”), **max_depth** (None, 5, 10, 15, 20, 25, 30), **min_samples_split** (50, 100, 200, 300,500,700,900), and **min_samples_leaf** (1, 2, 5, 10) combination. With a cv=5, the grid search tested 392 candidates, for a total of 1960 fits. Of the candidates, the following model had the best average accuracy score.

best_tree_model = DecisionTreeClassifier(criterion=‘entropy’, max_depth = 10, min_samples_leaf = 5, min_samples_split = 500, random_state=10)

With tuning, the model’s performance with the training dataset improved by about 20%, with the CV average accuracy rate increasing from 0.42 to 0.52.

K-Nearest Neighbors

For tuning my KNN model, I used a GridSearchCV to consider the optimal combination of **n_neighbors** (in range(5, 160, 5)), **weights** ('uniform', 'distance'), and **metric** ('euclidean', 'manhattan', 'minkowski'). Because my computer was struggling with the computation power, I reduced the CV from 5 to 3 for both the KNN and later random forest tuning process. Therefore, the grid search tested 186 potential models 3 times. The model with the highest average accuracy score is below:

best_knn_model = KNeighborsClassifier(n_neighbors = 65, metric = "manhattan", weights = "distance")

With tuning, the performance improves slightly, the average CV accuracy rate increases from 0.47 to 0.53.

Random Forest

For tuning my random forest model, I used a GridSearchCV to determine the optimal combination of **criterion** ("gini", "entropy"), **n_estimators** (50, 100, 150, 200, 250), **max_depth** (10, 15, 20, 25), **min_samples_split** (50, 100, 200, 300,500), and **min_samples_leaf** (1, 2, 5, 10). With a CV = 3, the grid search fitted 3 folds for 800 potential models, resulting in 2400 fits. The model with the highest average accuracy score is below.

best_rf_model = RandomForestClassifier(criterion= 'gini', max_depth= 25, min_samples_leaf= 5, min_samples_split= 50, n_estimators= 250)

Compared to the decision tree and KNN models, the tuned and untuned random forest models have approximately the same accuracy rate with the training data (0.55). Therefore, at least with the training data, it is unclear the impact that the tuning has on the model performance.

XGBoost

For tuning my XGBoost model, I used a GridSearchCV to determine the optimal combination of **n_estimators** (50, 100, 200), **learning_rate** (0.01, 0.1, 0.3), **max_depth** (3,5,7,10), and **subsample** (0.5, 0.8. 1.0). With a CV = 5, the grid search fitted 5 folds for each combination. The model with the highest average accuracy score is below.

best_xgb_model = XGBClassifier(use_label_encoder=False, eval_metric='logloss', learning_rate=0.1, max_depth = 5, n_estimators = 200, subsample = 0.8)

Compared to the decision tree and KNN models, the tuned and untuned XGBoost models have approximately the same accuracy rate with the training data (0.55). Therefore, at least with the training data, it is unclear the impact that the tuning has on the model performance.

Comparing Tuned Model Results

	Model	Training Accuracy	Test Accuracy	Precision	Recall	F1 Score
0	Decision Tree	0.538	0.526	0.522	0.526	0.486
1	KNN	1.000	0.537	0.544	0.537	0.506
2	Random Forest	0.687	0.555	0.571	0.555	0.518
3	XGBoost	0.658	0.571	0.569	0.571	0.549

All 4 tuned models struggled to classify communities into the correct NWI category. Each of the models test accuracy, precision, and recall scores are approximately the same (range between 0.522 and 0.571). These metrics suggest that these models were able to correctly classify less than 60% of the observations in the unseen test dataset, and across the different classes, only corrected classified about half of the actual class observation into its appropriate class. The F1 Score values for all of the models is between 0.486 and 0.549, which suggests that the precision and recall scores are balanced but the performance is still low.

Overall, the model metrics indicate that the models have significant bias in their predictions. The models range in their variance. The KNN model, with a training accuracy of 1, overfits the training data and therefore its results are less generalizable for the unseen test dataset. The three other models (decision trees, random forest, and XGBoost) appear to have lower variance because its training and test accuracies are closer together, but it is also important to recognize that even in the training of the model, the models had a relatively low accuracy (less than 0.70).

This suggests that there are unobserved factors not captured in our models that may be impacting NWI. However, based on the knowledge of how the NWI is calculated using infrastructure measures as opposed to social and demographic measures, this makes sense.

Reviewing Model Results by Class

Because the target feature contains multiple categories, I want to review how the model's performance varies by class (NWI categories).

As a reminder, the categories are as follows:

- NWI Category 0 = Least Walkable Community
- NWI Category 1 = Below Average Walkable
- NWI Category 2 = Above Average Walkable
- NWI Category 3 = Most Walkable

Class Precision

	Decision Tree	KNN	Random Forest	XGBoost
NWI Cat 0	0.604697	0.572932	0.698031	0.618598

	Decision Tree	KNN	Random Forest	XGBoost
NWI Cat 1	0.517544	0.484474	0.532990	0.564758
NWI Cat 2	0.528004	0.566520	0.552457	0.572394
NWI Cat 3	0.463576	0.570370	0.595349	0.536814

These results show that the models' precision across categories is approximately the same: the precision scores range from 0.46 to 0.69. All of the models have the highest precision with NWI category 0, meaning that of those predicted to be category 0, 57.3-69% of them are actually category 0. For categories 1-3, the models consistently have precision scores less than 0.6. Overall, Random Forest appears to have the highest precision across the different categories.

Class Recall

	Decision Tree	KNN	Random Forest	XGBoost
NWI Cat 0	0.269398	0.332171	0.278117	0.400174
NWI Cat 1	0.479350	0.623223	0.560596	0.560934
NWI Cat 2	0.796502	0.704065	0.796502	0.769475
NWI Cat 3	0.114255	0.125680	0.139282	0.218172

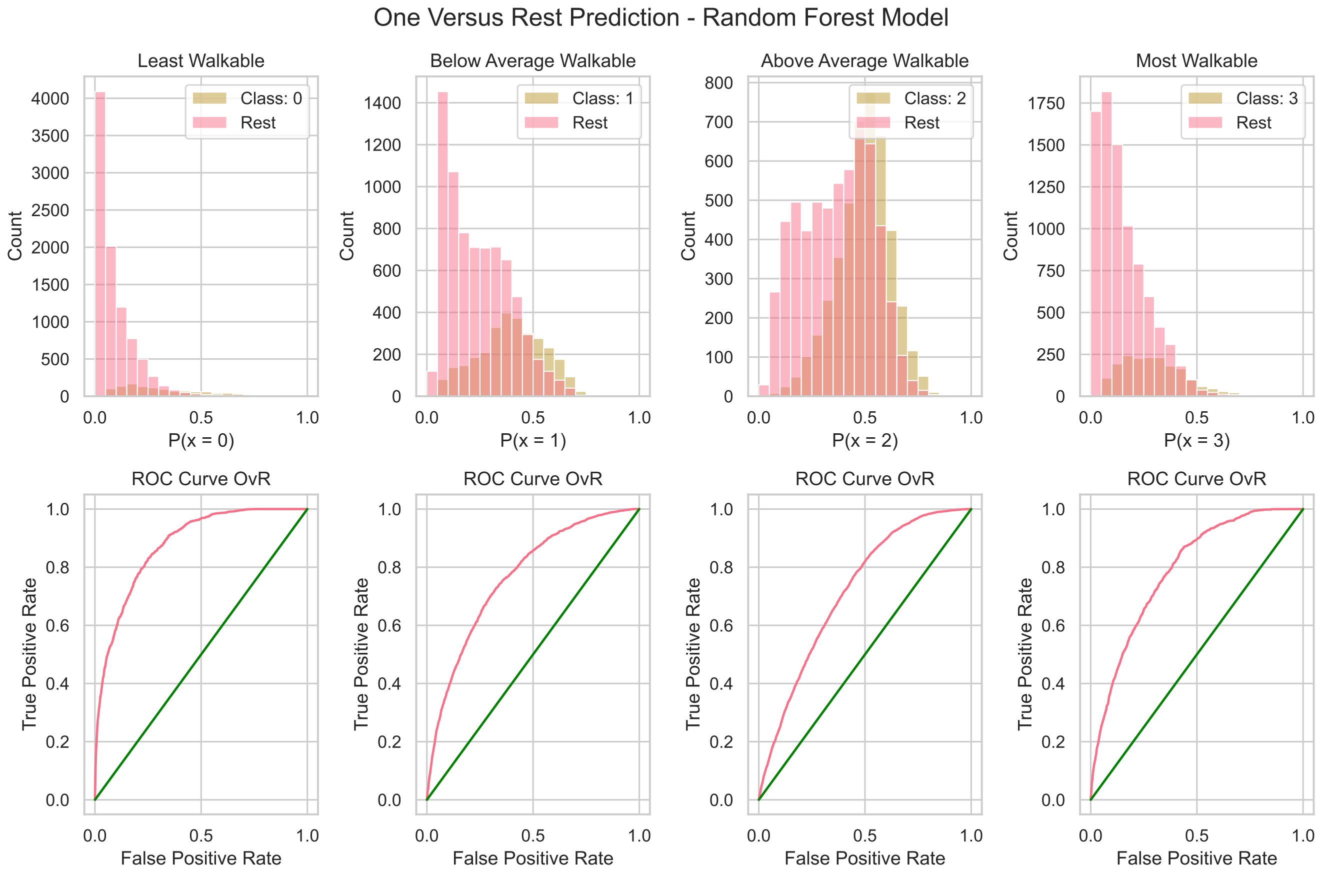
These results show that the models vary significantly in their predictions by class. All of the models are highly sensitive to category 2, with recall scores between 0.7 and 0.8, but have very low sensitivity to category 3 (recall scores between 0.11 and 0.21). This means that at most 1 in 5 of the NWI category 3 observations are being categorized as such. This suggests that our models are significantly overpredicting categories 1 and 2 and underpredicting categories 0 and 3. In our dataset, NWI categories 1 and 2 are the most common and NWI categories 0 and 3 are the least common, which may explain the differences in recall scores. Based on these results, the XGBoost model is the most sensitive model to the different classes.

Class ROC-AUC Scores (One versus Rest)

While the Precision and Recall scores suggest ways the model could perform better in the multiclass classification, the ROC-AUC scores for one versus rest suggest that the model is able to distinguish between classes when it considers the classification as a binary classification issue. In this calculation for One versus Rest, which iterates through each class to treat the selected class as the positive class and the rest as the negative class, we see that the models are much better than random at differentiating the selected class from the rest. The ROC AUC calculations are at least 0.686, with most being greater than 0.75. These results for NWI category 0 and 3, in contrast to the recall scores above, suggest that when viewed separately from the other classes the model is can distinguish these observations well from the others.

	DT	KNN	RF	XGB
NWI Cat 0	0.848	0.868	0.876	0.891
NWI Cat 1	0.737	0.754	0.767	0.786
NWI Cat 2	0.686	0.705	0.717	0.732
NWI Cat 3	0.760	0.778	0.788	0.797

The graphs below, which depicts the ROC-AUC results for the Random Forest Model, demonstrate the model's ability to distinguish that class versus the rest and how it varies based on the probability of the class in the distribution. The ROC-AUC results suggest that may be the class imbalance and the similarity between classes that may be impacting the models' performance.



Comparing Models that Account for Target Feature Imbalance with Original Models

Considering the known class imbalance in the target feature, I decided to see how accounting for this in the training would effect the model's performance. For this, I only focused on two models, Random Forest and XGBoost, because of their robustness and their ability to account for weights in the training process.

Balanced Random Forest (from imbalanced-learn package)

A balanced random forest differs from a classical random forest by the fact that it will draw a bootstrap sample from the minority class and sample with replacement the same number of samples from the majority class. This results in an undersampling of the majority class.

To tune my balanced random forest, I used a RandomizedSearchCV to test 20 different combinations of **n_estimators**, **max_depth**, **min_samples_split**, **min_samples_leaf**, **bootstrap** (True/False), **class_weight**, and **sampling_strategy**. With a CV = 5, the model with the highest average accuracy score is below.

best_Bal_rf_model = BalancedRandomForestClassifier(random_state=42, replacement = True, sampling_strategy = "all", n_estimators = 134, max_leaf_nodes = 41, max_features = .7999, max_depth = 6, class_weight = "balanced_subsample", bootstrap = True)

XGBoost with Class weighting

Compared to random forest, there is not a different package for the weighted version of the classifier. Instead, to add target class weights to the XGBoost, I computed class weights based on the training data distribution, and used them as part of the fitting process. I also used a RandomizedSearchCV to test 15 different hyperparameter configurations that considered the weights as part of the fitting process. Based on this, the model with optimal hyperparameters with weights is below.

best_xgb_2 = XGBClassifier(use_label_encoder=False, eval_metric='logloss', learning_rate=0.1, max_depth = 10, n_estimators = 200, subsample = 1)

Based on the results, the models that account for the class imbalance overall perform worse than the models that do not. The balanced random forest model has lower test and training accuracy than the model that does not weight the target classes. The weighted XGBoost appears to be overfitting on the training data because of its high training accuracy but then has lower test accuracy than the unweighted model. The weighted models also have lower precision and recall scores.

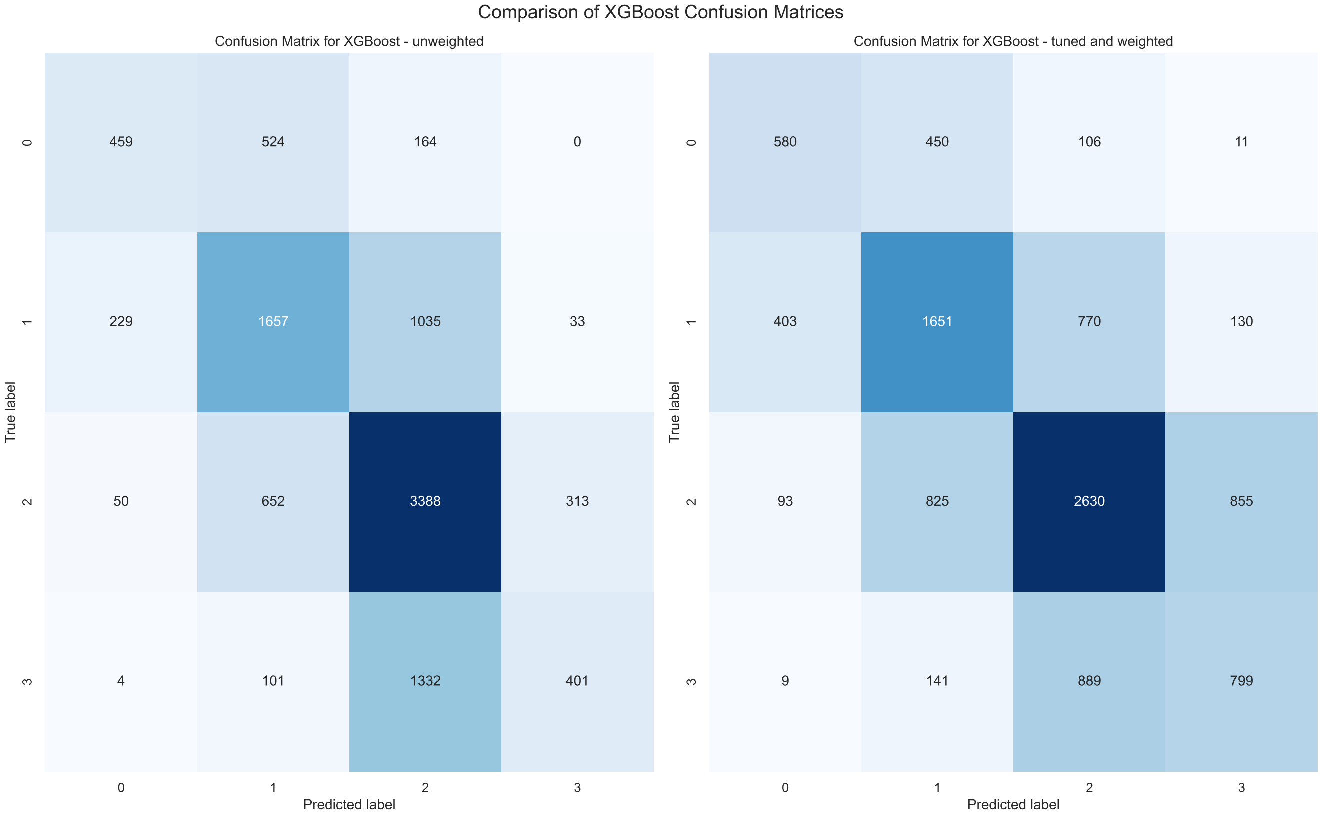
	Model	Training Accuracy	Test Accuracy	Precision	Recall	F1 Score
0	Random Forest - Unweighted	0.687	0.555	0.571	0.555	0.518
1	Random Forest - Weighted	0.454	0.431	0.481	0.431	0.428
2	XGBoost - Unweighted	0.658	0.571	0.569	0.571	0.549
3	XGBoost - Weighted	0.971	0.547	0.547	0.547	0.547

However, in reviewing the models' predictions of the test data compared the actual class distribution, we can see how that the unweighted versions of the models are significantly underpredicting classes 0 and 3, the minority classes, in favor of classes 1 and 2. The weighted versions mimic the distribution of the test data more closely.

	Least Walkable	Below Average Walkable	Above Average Walkable	Most Walkable
True Values	1147	2954	4403	1838
RF_unweighted	457	3107	6348	430
RF_weighted	2224	2842	2335	2941
XGB_unweighted	742	2934	5919	747
XGB_weighted	1085	3067	4395	1795

Additionally, looking more closely at how the models are classifying the different records, we see that the key areas of misclassifications are occurring between classes 2 and 3. In the left confusion matrix, which is for the unweighted XGBoost model, we see that a significant majority of the Class 3 observations are being predicted as Class 2. In contrast, the right confusion matrix, which is for the weighted XGBoost model, while more of the Class 3 observations are being correctly classified, the model is more likely to misclassify Class 2 records as either Class 1 or 3.

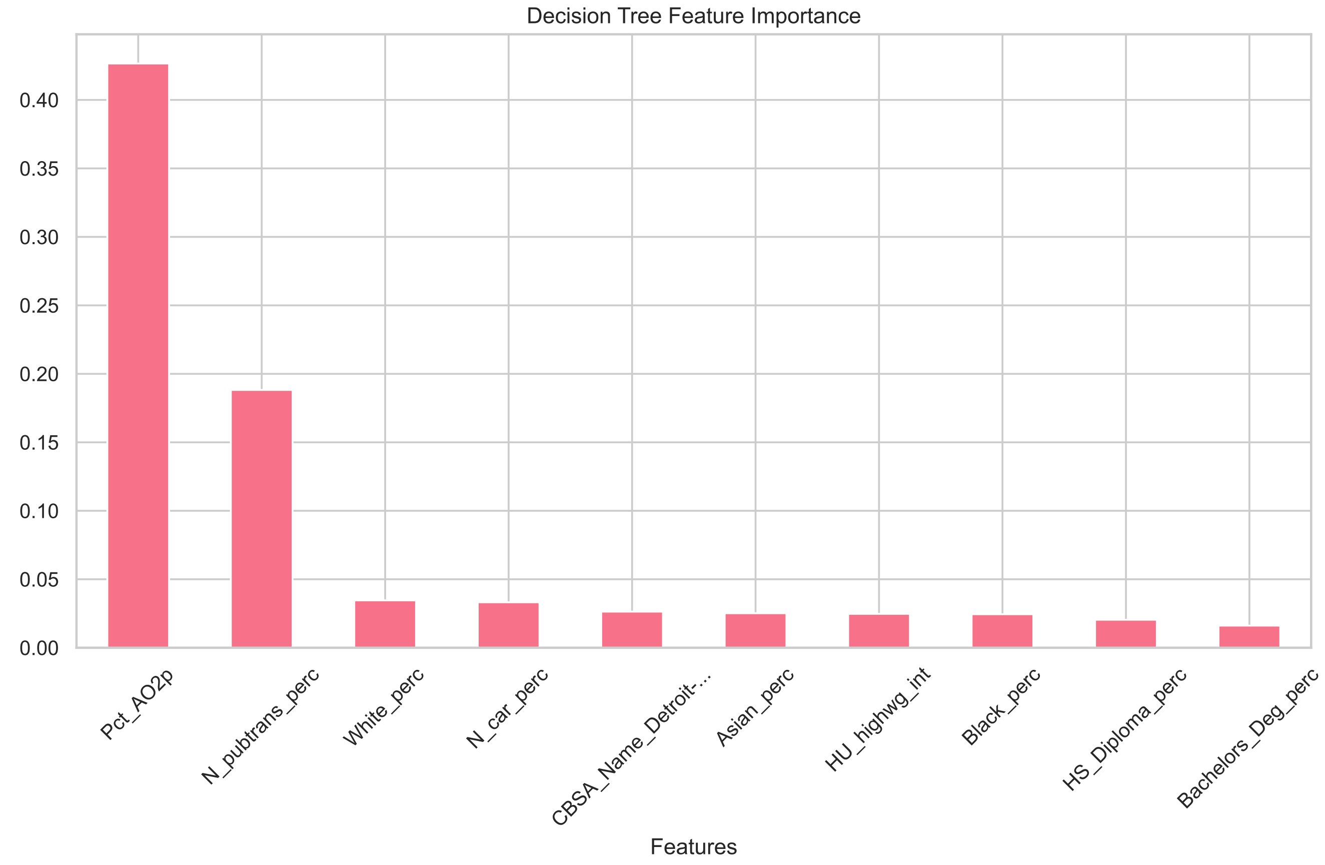
Thus, accounting for the class imbalance in the training does not result in better model results.



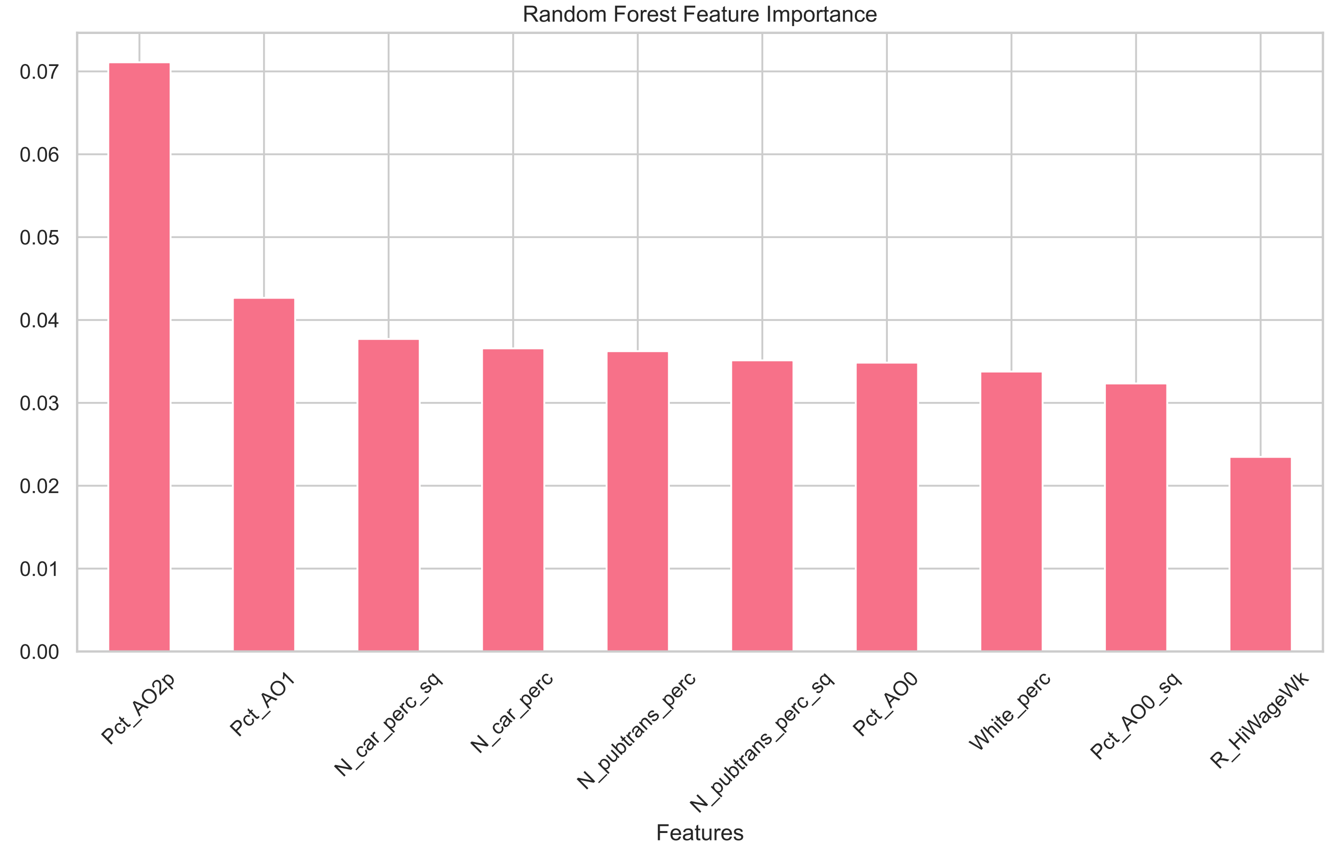
Feature Importance Across Models

For the Decision Tree, Random Forest, and XGBoost unweighted models, I can discern the features considered to be important in differentiating between the classes.

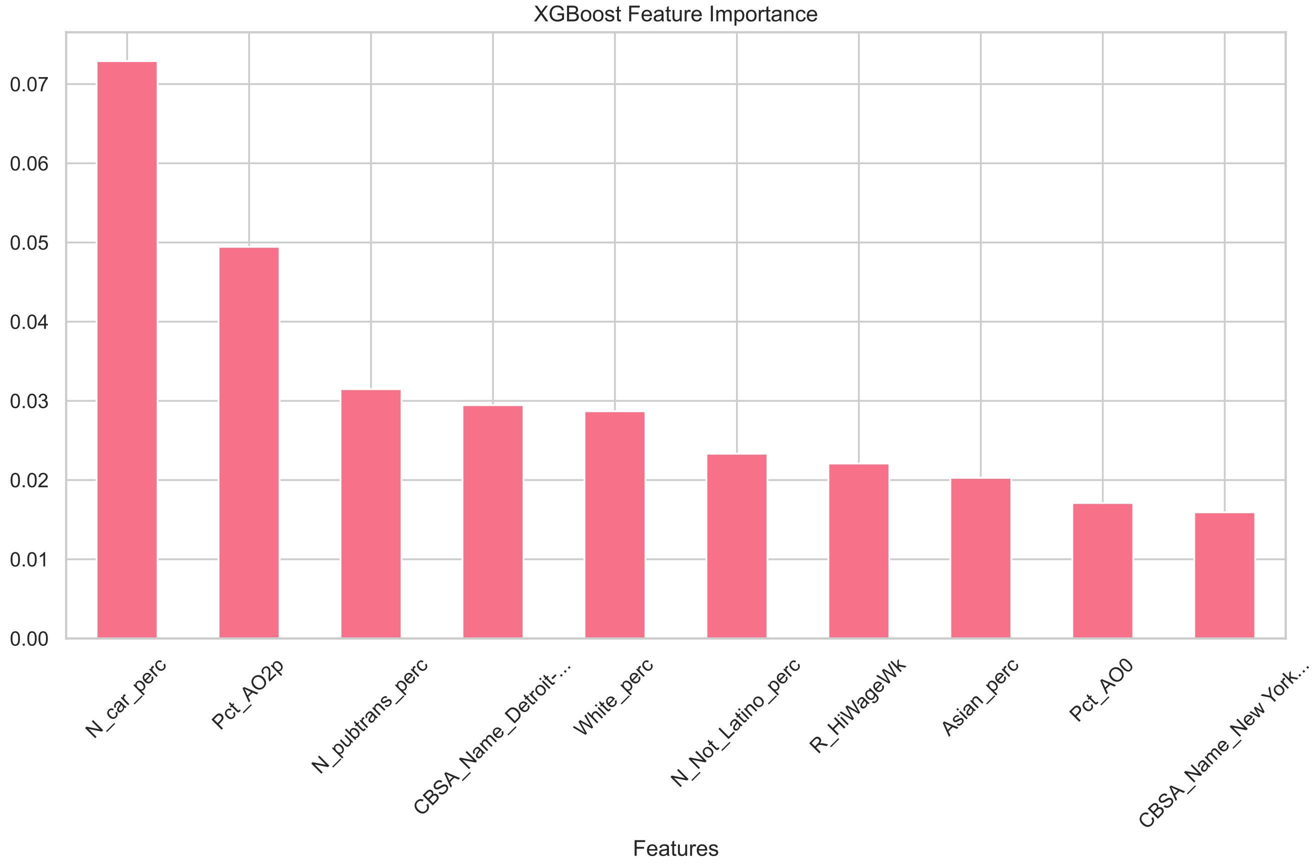
Decision Trees



Random Forest



XGBoost



A few features appear relevant across the models. For all models, the percent of Households with at least two cars (Pct_AO2p) and percent of people who indicate commuting using public transit (N_pubtrans_perc) are some of the most relevant features across models. This makes sense considering the context of the analysis. Other variables that appear important are percent of population that is white (white_perc). Of the metropolitan area indicators, only one of them appears relevant in a few of the models (Detroit) but it is unclear why.

While the models struggle to accurately classify the different locations, the features that appear important do suggest that commuting habits and select socio-economic indicators like number of households with two or more cars (which suggest having the income to do so) may provide context on walkability in a community.

Reviewing Placebo Model Results

To confirm that there is not a mechanical issue occurring in my models, I did complete a placebo test for all models where I fitted the models only using the 4 infrastructure features that the EPA says that it uses to calculate the NWI. Based on the results below, it indicates that these features are successful in correctly distinguishing between the four NWI categories.

	Model	Training Accuracy	Test Accuracy	Precision	Recall	F1 Score
0	Decision Tree	1.000	0.959	0.959	0.959	0.959
1	KNN	0.928	0.899	0.899	0.899	0.899
2	Random Forest	1.000	0.976	0.976	0.976	0.976
3	XGBoost	1.000	0.982	0.982	0.982	0.982

Discussion and Reflections

Based on the results of the models and feature importance graphs above, I am able to adequately address the research question with caveats. Transportation habits such as driving to work versus taking public transit, and having access to multiple cars have some predictive qualities in determining neighborhood walkability. Additionally, select demographic features, such as percent of population that is white or number of people with high wages, had some predictive capabilities. However, these features are all in context of models that have an average rate around 50%. Therefore, while I can identify the important features within these models, overall trying to predict walkability based only on demographic features and transportation habits does not provide optimal results.

I am mostly satisfied with my results, but see areas for improvements. The EPA's National Walkability Index is a relatively new metric that has not been applied in many contexts. As stated previously, the EPA based the NWI on 3 infrastructure metrics: street intersection density, proximity to transit stops, and diversity of the land uses. Therefore, in reviewing the results of my models based only on social and transportation features compared to the placebo models that are based only on the identified infrastructure metrics, I can see that the walkability can be accurately predicted by the infrastructure features but not as much by the social context of the neighborhood. I view it in a positive light that sociodemographic factors have difficulty predicting walkability, because it demonstrates that the EPA has created a metric of walkability infrastructure that is not highly correlated with demographic characteristics. It may be a helpful metric for local governments interested in improving walkability because it provides a measure of infrastructure rather than of presence of particular sociodemographic groups. However, further investigation is needed to confirm whether the NWI metric can be predicted by other neighborhood factors.

Regarding my methodological choices, based on the models' confusion between NWI categories 2 and 3, I think that collapsing the 4 NWI categories into 2 so that the model only classified areas as less walkable or more walkable likely would have improved the predictive capabilities of the model. However, I wanted to stick with the EPA's defined categories for this project and the divisions of the NWI scores across the four categories is not equal.

Reflections

I felt the learning curve at numerous points while working on this project. At every stage, I would feel like I would propose something, and then quickly after submission we would learn in class a technique that would work better for my project. For example, in Stage 2 I proposed only creating Decision Trees and KNN models, but added Random Forest and XGBoost in subsequent stages because I thought (correctly) that they will perform better with this task. The biggest mental shift occurred in considering this project from a machine learning perspective. Initially, my stage 1 considered this more topic from a statistical learning perspective because I wanted to understand the impact of walkability (X) on commuting habits (Y). But from a machine learning perspective, we are more interested in how the model performs overall in predicting Y rather than a certain feature's impact on Y. With this mental shift I think focusing on walkability as my target feature made this project more successful and also helped me learn more about neighborhood walkability.

In reviewing the earlier stages of my project, one thing that I assumed based on the literature is that there would be a noticeable predictive relationship between a community's sociodemographic factors and commuting habits and its walkability. However, based on the results of the model, a more apt main research question would have been "Can demographic features and transportation habits predict neighborhood walkability? If yes, which features are most important in the prediction?". This would have signaled that the research project was more of a feasibility study exploring whether there was a relationship. While I was able to discern the important features in my models, I cannot say that the models based on these features can highly predict walkability.

Considering the learning objectives, while I know the goal is always to optimize the predictions of the ML model, I also think it is important to consider that there will be times that the model will not perform well because of the data left out of the model. I also think with my approach it makes sense that the model did not perform very well. However, despite the models' limitations, the consistency in the feature importance across the models shows that some sociodemographic features and commuting habits may be important to consider in future studies.

Future directions of the research topic could be to continue adding in more features or weights to the model to improve the estimation, or as new iterations of the EPA NWI are released, to compare the changes in scores over time to see if there are social factors related to the changes. I could also use the models to compare the EPA NWI to a more common walkability metric like WalkScore to see how different metrics of walkability relate to demographic features and commuting habits.

Works Cited

Carson, J. R., Conway, T. L., Perez, L. G., Frank, L. D., Saelens, B. E., Cain, K. L., & Sallis, J. F. (2023). Neighborhood walkability, neighborhood social health, and self-selection among U.S. adults. *Health & Place, 82*, 103036. 10.1016/j.healthplace.2023.103036

Chapman, J., Fox, E. H., Bachman, W., Frank, L. D., Thomas, J., & Rourk Reyes, A. (2021). *Smart Location Database: Technical Documentation and User Guide.* https://www.epa.gov/system/files/

documents/2023-10/epa_sld_3.0_technicaldocumentationuserguide_may2021_0.pdf

Conderino, S. E., Feldman, J. M., Spoer, B., Gourevitch, M. N., & Thorpe, L. E. (2021). Social and Economic Differences in Neighborhood Walkability Across 500 U.S. Cities. *American Journal of Preventive Medicine, 61*(3), 394-401. 10.1016/j.amepre.2021.03.014

Kim, E. J., Kim, J., & Kim, H. (2020). Does Environmental Walkability Matter? The Role of Walkable Environment in Active Commuting. *International Journal of Environmental Research and Public Health, 17*(4)10.3390/ijerph17041261

Litman, T. (2023). *Economic Value of Walkability.* Victoria Transport Policy Institute. https://www.vtpi.org/walkability.pdf

Smart Location Database. (2023, August 30,). Data.gov. Retrieved January 27, 2024, from https://catalog.data.gov/dataset/smart-location-database1

Smart Location Mapping. EPA.gov. Retrieved Jan 27, 2024, from https://www.epa.gov/smartgrowth/smart-location-mapping

Thomas, J., & Zeller, L. (2021). *National Walkability Index: Methodology and User Guide.* Environmental Protection Agency. https://www.epa.gov/sites/default/files/2021-06/documents/national_walkability_index_methodology_and_user_guide_june2021.pdf

U.S. Census Bureau.*American Community Survey Data.* Census.gov. Retrieved January 27, 2024, from https://www.census.gov/programs-surveys/acs/data.html

Understanding Geographic Identifiers (GEOIDs). U.S. Census. Retrieved March 2, 2024, from https://www.census.gov/programs-surveys/geography/guidance/geo-identifiers.html

Walk Score Methodology. Walk Score. Retrieved May 5, 2024, from https://www.walkscore.com/methodology.shtml.

Watson, K. B., Whitfield, G. P., Thomas, J. V., Berrigan, D., Fulton, J. E., & Carlson, S. A. (2020). Associations between the National Walkability Index and walking among US Adults — National Health Interview Survey, 2015. *Preventive Medicine, 137*, 106122. 10.1016/j.ypmed.2020.106122