Walking and Health: Improving Communities

ORIE 4741 Final Project, Spring 2023
Team #33: Emily Mei (em664), Jonathan Yun (jly37)
12 May 2023

Executive Summary

Problem Statement

With a desire to improve both local health and communities, we spent this semester taking a closer look at the intersection between walkability and public health outcomes. In doing so, we wanted to discuss how walkability impacts health, and the strength of the relationship between the two.

In doing so, we decided to focus our analysis around three main questions:

I. *Can walkability be used as a predictor for various health outcomes? If so, is walkability a stronger predictor for certain health conditions over others?*
II. *Can we predict an area's walkability score if we are given health data?*
III. *Are there subfactors (e.g., race, income, etc.) of the Walkability Index that are stronger predictors of health outcomes than just the Walkability Index itself?*

Conclusions

After creating models to answer all three questions, we came to various conclusions, on a question-by-question basis:

I. *Only neighborhoods with high walkability are predicted to have good health outcomes; however, for neighborhoods with medium-low walkability, classification tends to predict higher prevalence of health issues, with varying confidence.*
II. *Based on the models we created, we can predict walkability with limited confidence, given health data about the neighborhood. In the best case, approx. ¾ of our predictions are within a score of 2 of the actual Walkability Index.*
III. *Other subfactors do not serve as stronger predictors of health outcomes than the Walkability Index itself; however, neither are particularly strong predictors of health outcomes.*

While, on the whole, we do not have enough confidence within these models to change how a business or government makes decisions (see "Model Confidence"), we believe that we have proven that a relationship between walkability and health exists; as such, to improve health outcomes, we strongly encourage the development and prioritization of walkable neighborhoods to improve America.

Technical Exposition

Part I. Dataset and Data Cleaning

For this project, we used a combination of the following: the Walkability Index Dataset from the EPA, the Local Data for Better Health Dataset from the CDC, and the Zip Code to Census Tract Relationship File from the U.S. Census Bureau. We will use walkability and health to refer to the two datasets.

While the health dataset contains relatively straightforward values for most measures (eg. % prevalence of diseases, access to insurance, etc.), we would like to go into further depth regarding the Walkability Index, which is a novel metric. In essence, it scores a neighborhood's walkability from 1 to 20, with the following benchmark values:

Neighborhood Type	Walkability Index
Suburban Residential (eg. Frisco, TX)	**8.5**
Historic Downtown (eg. Ithaca Commons)	**13.5**
High-Density Urban Center (eg. NYC)	**17.5**

Figure 1: Sample Walkability Scores

In terms of dataset size and complexity, we found approx. **220,000** census blocks with an assigned walkability score, and approx. **969,000** health measure-ZIP code pairings; while this did mean that we needed to use the relationship file to relate the census block (in the walkability data) and ZIP code (in the health data), we considered this to be adequately "large" for the purposes of analysis. After modifying the walkability dataset, we extracted the ZIP code and its corresponding walkability score and joined this with the health data to create a dataset with both health outcomes and walkability score, for approx. **29,990** ZIP codes across the U.S. (for reference, the U.S. has ~ 41,000 ZIP codes).

To go into the "messy" aspect of the data, we did run into a few issues. One was that the full census tract number in the walkability dataset was rounded, and thus had to be constructed manually using the state, county, and tract numbers. Another issue was that multiple census tracts correspond to the same ZIP code, and thus required aggregating the walkability scores corresponding to the same census tracts, which we did by averaging the scores. While we did lose some data while performing the joins described above, as some matches were missing and thus yielded N/A values that could not be used for analysis, this did not significantly affect the amount of usable data, as we retained **91%** of the data afterwards (we started with ~33,000 ZIP codes before processing).

Part II. Analytics and Modeling
Section 1: Tree Classification

To answer the first question we had — on whether it was possible to predict health outcomes based on walkability score alone — we first considered what measured outcome would sufficiently measure a positive / negative health outcome, and settled on whether an area had **elevated prevalence** of a health issue or not. As a binary true-false condition, we thus decided to create a classification model to answer this question, either as a form of logistic classification or using a tree-based model. To do this, we needed to perform some feature engineering to extract relevant data first, as described below.

Feature Engineering

In order to determine "elevated prevalence," we decided to use the Health Outcome Measure Definitions provided by the CDC with the PLACES dataset; this allowed us to find the national average of each health outcome or health predictor, and thus list each ZIP code with either a "True" or "False" label, depending on whether each measure was either more or less prevalent than the national average, respectively. Sample outputs after feature engineering are described below:

	index	ACCESS2	CANCER	ARTHRITIS	COPD	CHD	OBESITY	BPHIGH	MHLTH	PHLTH	LPA	NatWalkInd
0	1001	6.2	8.4	29.5	6.4	7.0	27.9	32.6	14.3	9.3	20.3	9.355556
1	1002	6.7	4.5	17.8	4.6	4.3	20.9	21.1	16.5	7.6	17.5	10.600000
2	1003	8.8	0.6	5.4	3.0	1.1	19.7	9.9	25.5	7.1	19.7	12.384615
3	1005	5.9	7.0	27.2	5.7	5.7	29.9	29.1	14.0	8.7	20.8	5.875000
5	1008	5.8	7.1	27.5	6.1	6.2	29.2	31.1	14.7	9.2	19.1	5.555556
...	...	...	...	...	...	...	...	...	...	...	...	...
32404	99923	9.6	8.4	26.0	5.8	6.6	28.0	36.4	9.8	9.0	17.9	1.833333
32405	99925	14.2	6.8	26.6	8.2	8.0	34.8	38.6	14.0	13.1	25.8	7.500000
32406	99926	15.4	5.6	24.0	8.0	7.6	37.5	37.0	16.3	14.0	27.9	6.000000
32407	99927	14.8	6.8	26.7	8.6	8.7	33.9	39.8	12.5	12.9	25.1	4.444444
32408	99929	13.7	7.5	26.2	7.6	7.7	33.3	36.6	12.6	11.4	23.3	6.166667

	index	ACCESS2	CANCER	ARTHRITIS	COPD	CHD	OBESITY	BPHIGH	MHLTH	PHLTH	LPA	NatWalkInd	NUM_ELEV
0	1001	False	True	True	False	True	False	False	False	False	False	9.355556	2-3
1	1002	False	False	False	False	False	False	False	False	False	False	10.600000	0-1
2	1003	True	False	False	False	False	False	False	True	False	False	12.384615	2-3
3	1005	False	True	True	False	True	False	False	False	False	False	5.875000	2-3
5	1008	False	True	True	False	True	False	False	False	False	False	5.555556	2-3
...	...	...	...	...	...	...	...	...	...	...	...	...	...
32404	99923	True	True	True	False	True	False	False	False	False	False	1.833333	4+
32405	99925	True	True	True	True	True	False	False	False	False	True	7.500000	4+
32406	99926	True	True	False	True	True	False	False	False	False	True	6.000000	4+
32407	99927	True	True	True	True	True	False	False	False	False	True	4.444444	4+
32408	99929	True	True	True	True	True	False	False	False	False	False	6.166667	4+

Figure 2: Raw Data (left), and Binary Extracted Features (right)

Model and Results

Following this, we thus decided on whether we wanted to classify using logistic regression or using a tree-based model; given that logistic regression would limit us to binary categorization, we decided on using a tree-based model, which would allow us to have multiple categories, and thus decided to create an ordinal prediction scheme on the number of measures that were above the national average (i.e., given a "True" label), as follows:

0 or 1 Elevated Measures	**2 or 3 Elevated Measures**	**4+ Elevated Measures**
Category 1	Category 2	Category 3

From a model creation standpoint, it should be noted that preliminary models returned trees that were more than 25 nodes deep and which had 200+ individual leaves; in order to prevent excessive bias and model overfitting, we decided to "prune" the tree and limited both tree depth and the allowed maximum number of leaves.

After fitting a model that would try to predict each category (see Figure 3), we found the following model would optimize the bias-variance tradeoff, with the following prediction regime based on walkability index:

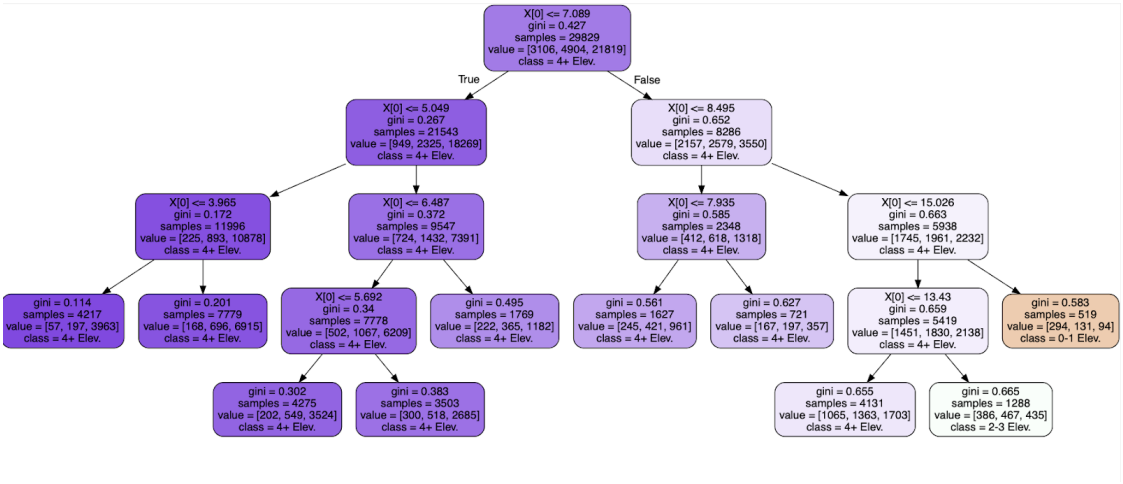


Figure 3a: Tree-Based Predictions, with Color Indicating Strength of Prediction

Walkability Index	**# of Predicted Elevated Measures**
Index > 15.026	0 or 1 (Orange)
15.026 > **Index** > 13.43	2 or 3 (White)
13.43 > **Index**	4 or more (Purple)

Figure 3b: Summary of Predictions

Following this, it becomes clear that we would only predict that areas with very high Walkability Indexes, such as historic downtowns or metropolitan city-centers with high public transit access, would have few elevated health measures, and that all others would immediately be predicted to have above-average prevalence of health issues, ranging from with weak confidence for more suburban areas (6.5 < Index < 13.5), to with high confidence for more rural areas (Index < 6.5). Due to the nature of the classification, highlights on specific health outcomes could not be ascertained (only the summed number of elevated measures); however, the relation between specific outcomes and walkability is further analyzed below, in Part II(b).

Section 2: Individual Features and PCA

To consider the second and third questions (on the subfactors of the walkability score, and on the prediction capabilities of health outcomes on walkability), we wanted to look at the impact of individual features, as well as how well we could explain the variance within the Walkability Indexes.

Principal Component Analysis

To start with explainability in variance, we first wanted to try Principal Component Analysis (PCA); this would allow us to visualize the spread within the health outcome data

on a human scale, and also determine whether there were clear clusters or trends with respect to walkability that we could capture with a full model. After running a PCA with three components, we found the following principal component plot, with the first three components capturing **>95%** of the variability within the health data (Figure 4a), and with the following elbow plot (Figure 4b):

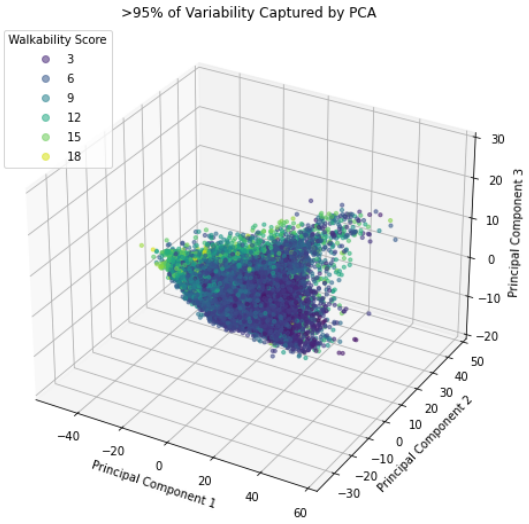


Figure 4a: Principal Components vs. Walkability

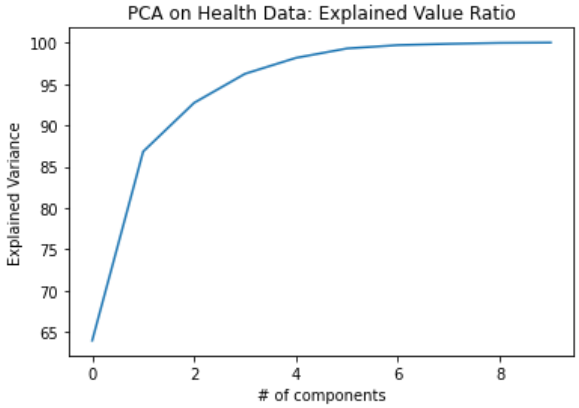
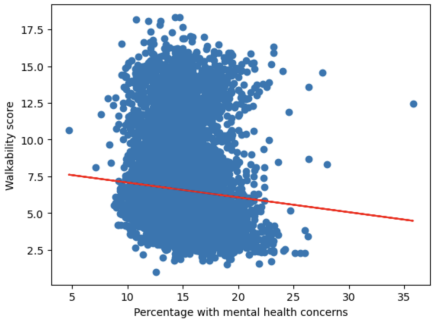


Figure 4b: Explained Value Ratio

While we did find that much of the variance in health data could be reduced to only a couple principal components (suggesting many features were redundant, or otherwise highly-correlated with one another), the PCA plot suggests that, while there is a clear trend in the walkability scores, there are no clear clusters or other connections; this points us both toward full regression, as well as a greater in-depth analysis of the individual features we can work with, as explained below.

Individual Features

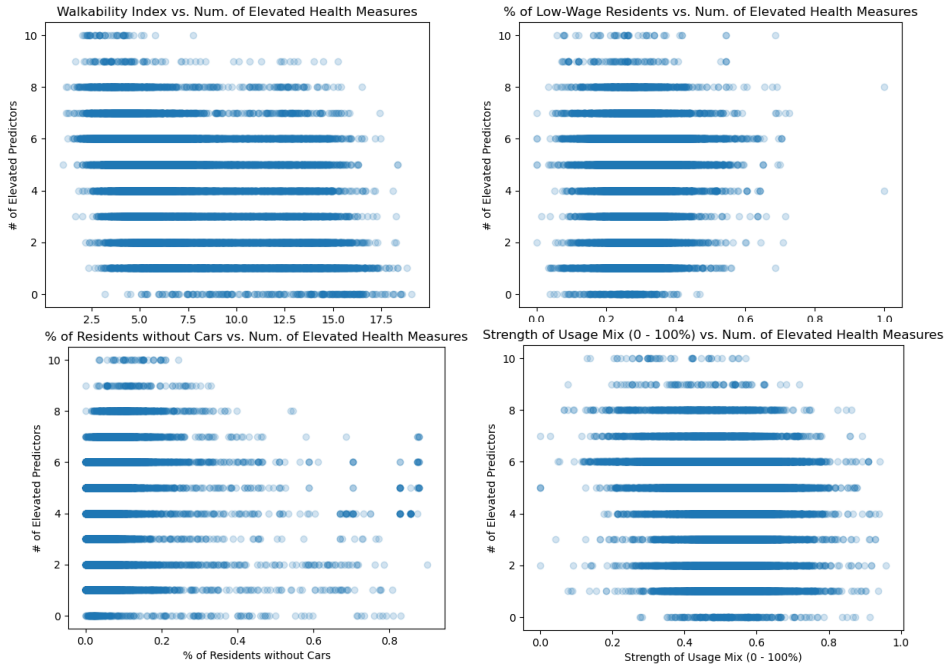
We considered whether individual features of the health data would be good predictors for walkability score. However, after fitting a linear regression model using each health feature, we found that there was not a strong correlation for any of the features. The graph on the right showing the predictor fit on mental health concerns, clearly demonstrates this.



Other Subfactors vs. Walkability

Finally, we wanted to look at some of the subfactors that made up the Walkability Index versus the composite Index itself, to both determine the strength of individual predictors on health outcomes and see if any individual predictors stood out as driving the correlation

between Walkability and Health. As such, we examined three subfactors: the percentage of low-wage earners (E_PctLowWage), the percentage of residents without cars (Pct_AO0), and the usage mix of buildings in each neighborhood (D2A_EPHHM), with the following plotted relationships against the number of Elevated Health Measures from Part II(a):



Figures 5(a)-(d): Subfactors Against Elevated Health Measures

From this, it is clear that while the walkability index itself has a (limited) negative correlation with elevated health measures in neighborhoods, many other indicators have correlations that are just as poor, or otherwise worse; this seems to indicate that they are not stronger predictors of health outcomes on their own. To verify, a simple linear regression was run predicting the Number of Elevated Health Measures: first, between the Walkability Index, and then, against the strongest subfactor, which is the Percentage of Low-Wage Earners and the Number of Elevated Health Measures. This resulted in the following R^2-adj. values:

Predictor	**R^2-adj.**
Walkability Index	0.23796
% of Low-Wage Earners	0.00331

Figure 6: Comparing R^2-adj. Between Subfactor and Composite Index

From this, it is clear that neither serve as strong predictors of health issues within a neighborhood, and thus conclude that walkability serves as the strongest composite predictor of health outcomes overall.

Section 3: Full Regression Models

To close the analytics section, we wanted to investigate whether we could predict walkability score using health data. To examine this relationship, we considered both a linear regression model and a non-linear Multi-layer Perceptron regression model.

First, we chose a subset of the health factors in the dataset to use in training our models. We decided to consider a combination of preventative health measures, health outcomes, health statistics and risk behavior. The health factors that we used were: access to health insurance, cancer, arthritis, chronic obstructive pulmonary disease (COPD), chronic heart disease (CHD), obesity, high blood pressure (HBP), mental and physical health concerns, and whether a person engaged in leisure time physical activity (LPA).

From this dataset, we did a 75-25 train/test split, and fitted both the linear regression and MLP regression model on the training data. The predicted walkability scores for the test data using these models are shown below, plotted against the true walkability score.

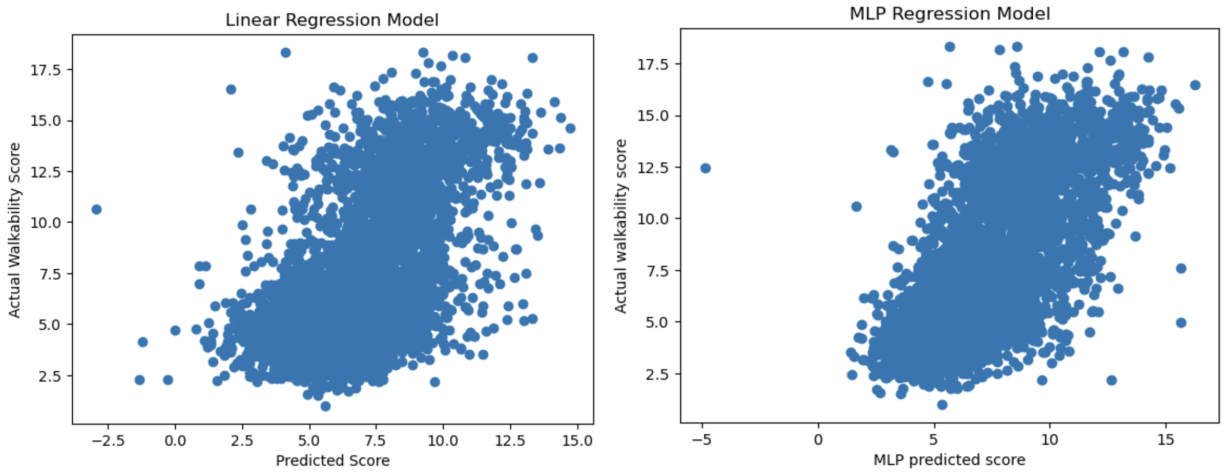


Figure 7: Predicted vs Actual Walkability Scores for Regression Models

As we can see from the figures above, the predicted walkability scores seem to be lower than the actual score for both models. We further analyzed the performance of the models by looking at their R-squared values and how much the predicted scores differed from the actual scores, which is summarized in the table below:

Metric	Linear Regression	MLP Regression
R-squared value	0.34	0.53
$\lvert$Prediction-Actual$\rvert \leq 0.5$	17.50%	24.52%
$\lvert$Prediction-Actual$\rvert \leq 1$	34.75%	**46.59%**
$\lvert$Prediction-Actual$\rvert \leq 2$	62.59%	**73.77%**

Figure 8: Linear vs Nonlinear Regression Model on various performance metrics

From the R-squared values, it is evident that there does not seem to exist a strong correlation between the health factors and walkability score, though the MLP does seem to "learn" a better relationship between them. We note that while neither model had predictions that were exactly the same as the actual score, around one-third to three quarters of the predictions were within 2 away from the true score, and consider this accuracy to be notable; *it seems possible to predict, with some confidence, the walkability of a neighborhood based on the health of its surrounding neighborhood*. It is clear, however, that predictions by the nonlinear MLP model are generally closer to the actual walkability score than predictions by the linear regression model, suggesting that the relationship between the health factors we considered and walkability score are more nuanced and complicated than a simple linear fit.

Other Considerations

Model Confidence

Overall, across all models that we have created, we find that our confidence in using these to answer questions for business / government purposes remain low. For example, considering the tree-based model, we find that the Gini Index for many predictors (including 0-1 elevated measures, or 2-3 elevated measures) is still high (> 0.5), and thus has a high rate of misclassification. Alternatively, for the linear regression and MLP, the low R^2-adj. value and high bounds around correct prediction limit our model confidence; it is hard to say we can empirically predict either walkability on health or vice-versa.

Algorithmic Fairness and Ethics

A possible concern with our project and its conclusions is the potential negative impact on how people perceive neighborhood quality. We have observed that our models tended to predict lower walkability scores, and that our tree classification indicated that walkability scores below a certain threshold corresponded to more health measures above the national average. If, for example, a health insurance company were to use these models to determine community health and walkability, this could inflate insurance costs. Resulting skewed perceptions could also negatively impact real estate values, investment and development decisions, or government appraisal, etc., thus affecting the quality of life within a community. Additionally, while we investigated underlying socioeconomic factors within the Walkability Index, it is possible we did not look far enough (health and socioeconomic concerns are highly correlated, especially protected attributes like gender, race, and income). As such, specific nuances may have been omitted or overlooked, especially when processing data (e.g., aggregation by ZIP code), which may result in overgeneralization of a geographic area or implicit correlation that we may have missed.