CS6140 Final Project Group 1: Walkability Index

James Fan, James Florez, Dominic Cauteruccio, Sara Spasojevic
Northeastern University, fan.ja, florez.ja, cauteruccio.d, spasojevic.s@northeastern.edu

Abstract - **This report is a look at the EPA's Smart Location Database and the calculated National Walkability Index. Current research suggests this current calculation has limitations, so this report looks to see if there are other factors besides the ones used to calculate Walkability that might do a better job of measuring the walkability of an area. The report finds that there are several important features, including D5ar, D5ae, D5br, and D5be, and determined that using a random forest or MLP model did the best job of predicting walkability in this study.**

Index Terms - Machine Learning, Multi-Layer Perceptron, Random Forest, Walkability

Introduction

The walkability index of a neighborhood is supposed to be a measure of how easy it is for a resident to live their daily life of shopping, dining, going to work, and recreation by walking instead of driving. The concept of walkability is gaining popularity across the world as people realize the health benefits of walking and the growing inconvenience of commuting. In the United States, the Environmental Protection Agency calculates the walkability index using the following formula and gives a score in the range of 0 to 20 (low to high).

$$\text{Walkability Index} = (w/3) + (x/3) + (y/6) + (z/6) \quad (1)$$

w = ranked score for intersection density
x = ranked score for proximity to transit stops
y = ranked score for employment mix
z = ranked score for employment and household mix

Current research has noted that this formula is insufficient for capturing all the variables that influence walkability, which is something we want to explore in this project. The EPA-provided dataset for walkability includes data on many other environmental and socioeconomic factors, and we built ML models that use these variables instead of the ones in the formula above to try to predict the walkability index of an area [1].

Literature Review

A literature review of 132 papers covering walkability was published in 2021 that analyzed the influence of built environment attributes on walkability. Their key finding is that there are many different definitions of walkability that use a variety of attributes and weights. The most used attributes are "intersection density, residential density and land use mix" while "attributes related to streetscape design were much less identified". They recommend improving walkability analysis by addressing "the impact of safety and security in walkability", including "more streetscape attributes (pedestrian facility and comfort as well as streetscape design features)", and evaluating "street network connectivity and accessibility by considering the real pedestrian network (including footpaths, pedestrian crossings, bridges, and tunnels) rather than the street network" [2].

A measure of walkability that has gained popularity is the Walk Score which is used on websites such as Zillow to inform housing decisions. Walk Score uses distance to common amenities, population density, block length, and intersection density to measure the walkability of an area. The shorter a walk is, the more weight it gets in the score. The final score is on a scale of 0-100 (low to high). Walk Score has shown to be a reasonable estimate of walkability, especially in urban areas. However, notable limitations include the lack of micro-scale elements such as sidewalk conditions, street trees, or crime rate. A study of perceived walkability by residents in Omaha, NE confirms that walkability aligns with residents' perception of walkability in urban areas but differs in more recreational areas and in suburban strip mall areas that feature a lot of retail [3].

Data Preprocessing

The dataset relied on census data, where each sample corresponded with a particular CBG (Census Block Group) - which generally contains from 600-3000 people. Each CBG entry had a variety of information - from the number of employed people in the area and types of work, number of workers (people that reside in the area that are employed), incomes, traffic that the area generates, infrastructure, size of the area, interaction with the surrounding CBGs, vehicle ownership, etc.

The largest part of the preprocessing step was getting to understand all the variables in the Walkability Index dataset and choosing which ones to use in this project based on a variety of factors. Using the technical documentation of the data, the group was able to determine the columns D2b_E8MixA, D2a_EpHHm, and D4a were used in the current Walkability Index [4]. Since this project aimed to look at other factors besides those currently used in the Walkability Score, these columns were discarded.

The next step was to pare down the remaining variables into a final group to use for models. The dataset is broken into groups, with all variables in the group heavily correlated with each other. Given that our goal was to explore how environmental and socioeconomic factors will predict the walkability score, the Design, Transit Access, and Walkability groups were discarded. The variables selected from the rest of the groups were Ac_Total, Ac_Water, Ac_Land, TotPop, CountHU, HH, P_wrkAge, Pct_AO0, Pct_AO1, Pct_AO2p, Workers, R_LowWagWk, R_MedWageWk, R_HiWageWk, TotEmp, E5_Ret, E5_Off, E5_Ind, E5_Svc, E5_Ent, E_LowWageWk, E_MedWageWk, E_HiWageWk, D1a, D1b, D1c, D1d, D2c_TrpMx2, D2r_WrkEmp, D5ar, D5ae, D5br, and D5be, NatWalkInd.

After picking the relevant variables and gaining some more understanding of what the data looks like, we have realized that it is a very clean dataset with little need for clean-up and handling of missing values. For household-related variables (CountHU and HH), null values were dropped, and for all other numeric variables, nulls were imputed using the median. The variables were finally standardized using z-scores in preparation for use in the models.

Exploratory Data Analysis And Feature Engineering

Our EDA analysis consisted of performing univariate and covariate analysis on the dataset which helped us gain an understanding of features' correlations and how they affect the walkability index (NatWalkInd). Our research led us to an understanding that the percentage representation of certain variables was more valuable to predicting the score than raw numbers, so we used percentages where available and created our own when not. We engineered features such as "Pop_Acre" (number of people per acre in that CBG), and "Occupied" (percentage of occupied households) among others; consolidated the number of autos owned into two categories (one or no cars, and two or more cars) after realizing the trends, and opted to use only the employment types relevant to the walkability index prediction (retail, service, entertainment). We kept track of the improving correlation matrix utilizing heatmaps.

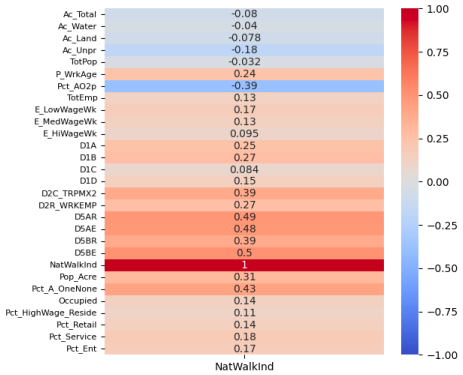

FIGURE I
EXPLORATORY DATA ANALYSIS CORRELATION MATRIX

Model Training, Validation, And Regularization

Before finding the best parameters and training our models, we created 3 functions that would aid us in this process

- model_train_and_eval_single: Given a model and data, the function trains the model, plots the coefficients, and evaluates the performance with MSE and R^2.
- model_train_and_eval: This function handles k-folds (or 80-20 split if k_fold = False), standardization, PCA, and makes calls to the previous function to produce the overall results for us.

The strategy for tuning the model parameters was to identify the important ones that will stay fixed, and then iterate through sets of options for other ones. For each combination identify the training and validation error and plot them to identify any trends and pick the best performing model. High training error is associated with high bias as bias-variance trade-off refers to finding a model with the right complexity, while also minimizing both the train and test error. This is done on an 80-20 split, and after finding the winning model, our two functions are utilized to finally train and evaluate that model with enabled K-fold and standardization parameters.

Further augmenting the models, regularization is done through hyperparameters set before the initialization of each model and caught in the model and evaluation function. This allows us to adjust various ways the model adheres to the data to prevent under or over-fitting. Depending on the model, the parameters can be set either in the model flags itself or in the model evaluation function.

Linear And Ridge Regression

Due to the simplicity of these models, and lack of any specific parameters to fine-tune, both were initialized using sklearn library and then passed to our function for fitting, evaluating, and plotting. In ridge regression, alpha was set to 1.1, as a regularization parameter.

Random Forest

Our fixed values for this model were min_samples_leaf (10) and max_features (10). The number of maximum features used would be almost split in half. If the number of samples in a node is less than min samples leaf, the splitting will stop and this node will become the leaf node. Both parameters gave the best results consistently, and they were kept, while a search through sets of other parameters was performed. The result that minimizes training and validation errors the most had 30 estimators and depth None (can be as deep as needed), though we opted to go with 10 estimators for the winning model, as R-squared score was close to the 4th decimal space, but time to compute was much shorter.

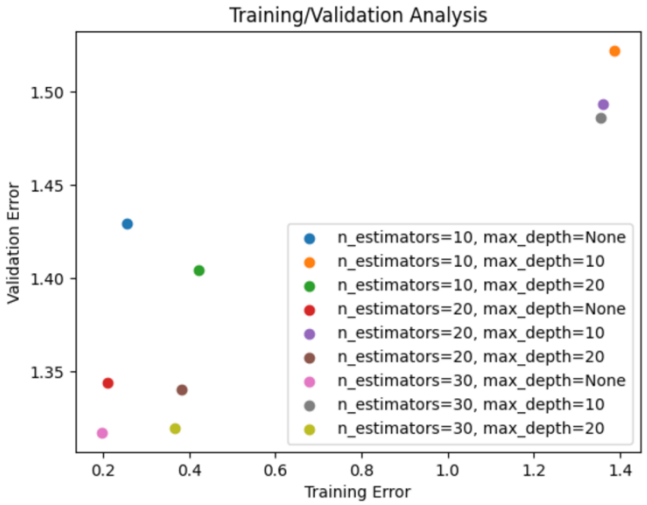

FIGURE II
BIAS-VARIANCE TRAINING/VALIDATION ANALYSIS FOR RANDOM FOREST MODELS

SUPPORT VECTOR MACHINE

Similar approach was taken with SVM, where parameters kernel (rbf) and max_iter (1000) were kept, as they gave the best results (even though the model rarely converged no matter how differently the parameters were tuned). C and epsilon parameters were picked from a set of options and the best model according to the graph had C as 1 and epsilon as 0.1. This model was unstable with no pattern, as different configurations led to drastically different results. Ultimately, we picked the winning model with the highest R-squared and lowest MSE score to be the one with C - 20 and epsilon 1.0.

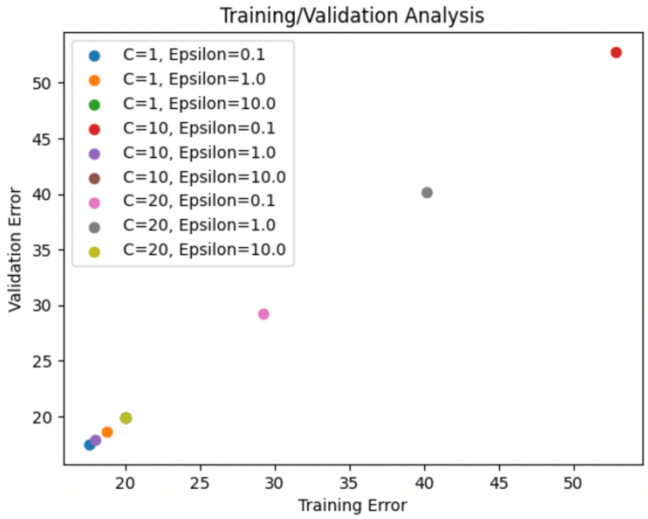

FIGURE III
BIAS-VARIANCE TRAINING/VALIDATION ANALYSIS FOR SUPPORT VECTOR MACHINE MODELS

MULTI-LAYER PERCEPTRON

Another model that benefited from this approach was MLP, where fixed parameters were alpha (0.1), learning_rate_init (0.001) and max_iter (300). Then, we went through a set of neurons and hidden layers to see which ones gave us the best result. Just by reading the graph, 2 layers and 20 neurons gave us the smallest training and validation errors. As we tested different options, we found that having 1 layer with 20 neurons performed very close to this one yet is it a simpler model that takes less time to compute. Therefore, we decided to move on with the simpler one, keeping in mind that trade-off.

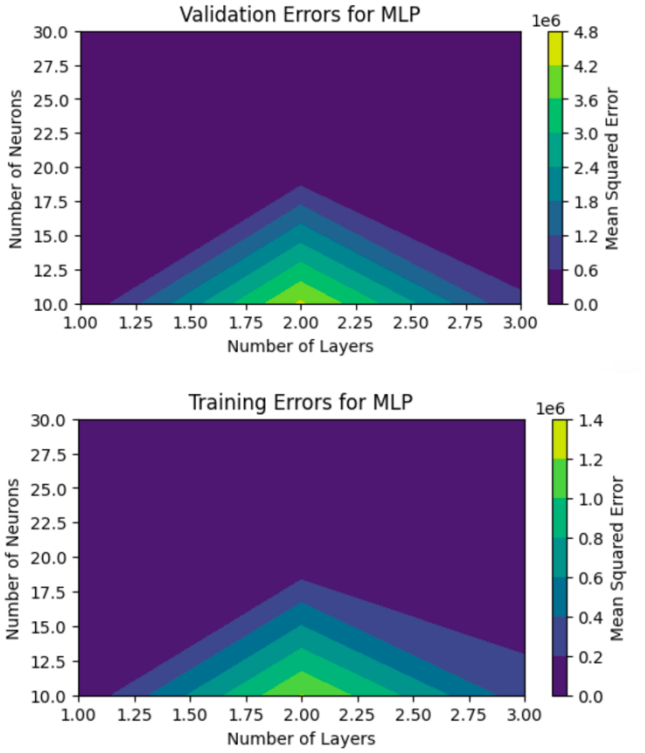

FIGURE IV & FIGURE V
BIAS-VARIANCE TRAINING/VALIDATION ANALYSIS FOR MULTI-LAYER PERCEPTRON MODELS

PRINCIPAL COMPONENT ANALYSIS

The Principal Component Analysis implemented in this project was SKlearn's decomposition.PCA. Using a training dataset with an 80/20 split, the eigenvalues were calculated and the explained variance of each was plotted on a Scree Plot. From there, it was determined that 6 Principal Components would be used re-running the best performing models with these principal components. Then, consistent with the rest of the project, k-fold validation was run to see how PCA performed across 5 different iterations with an 80/20 split on training and testing data.

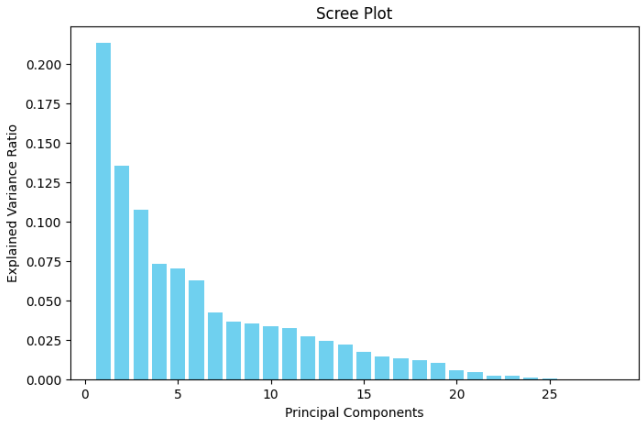

FIGURE VI
SCREE PLOT FROM PRINCIPAL COMPONENT ANALYSIS

RESULTS

Our analysis has shown that random forest and multi-layer perceptron models are the most effective at predicting the national walkability index by utilizing the other variables present in the EPA's dataset. They achieve the highest R^2 scores and lowest mean squared error values. For each model, k-fold cross-validation was run to perform 5 iterations with 80/20 splits.

Analyzing the impact of each feature on the model performance shows that the most important features are D5ar, D5ae, D5br, and D5be.

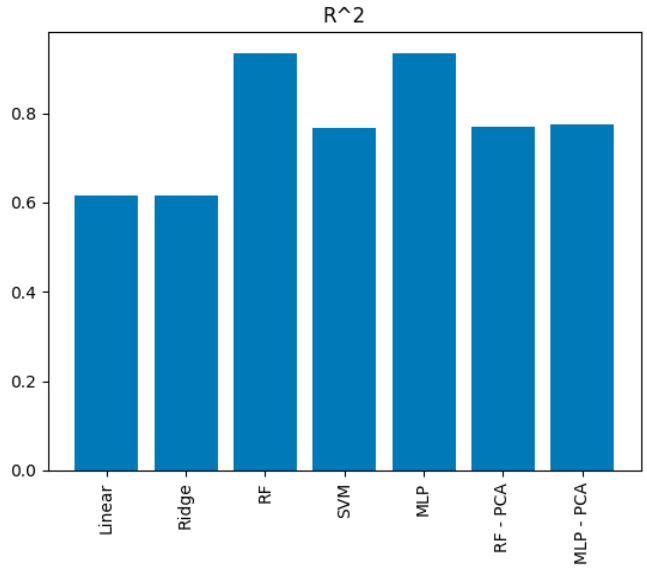

FIGURE VII
RESULTS FOR R-SQUARED FOR EACH OF THE MODELS

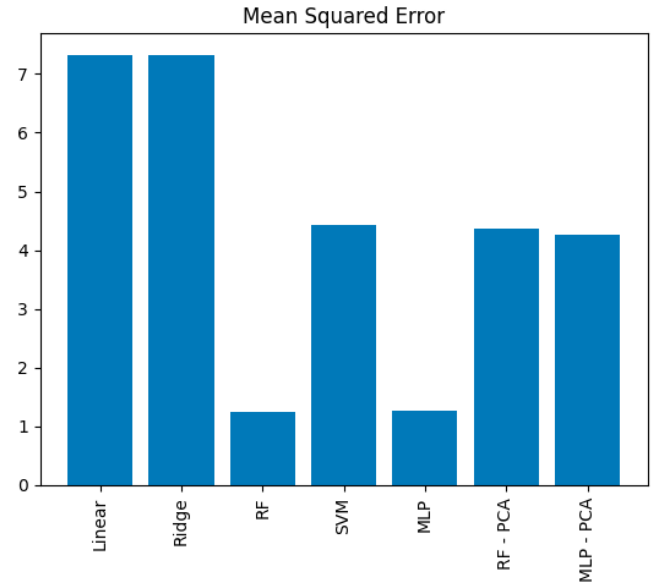

FIGURE VIII
RESULTS FOR THE MSE FOR EACH OF THE MODELS

LINEAR REGRESSION

The linear regression model implemented produced an average MSE of 7.34 and an R^2 of .613. The model emphasized two main variables, TotEmp and E_HiWageWK, and had two others with non-zero coefficients, E_MedWageWk and E_LowWageWk.

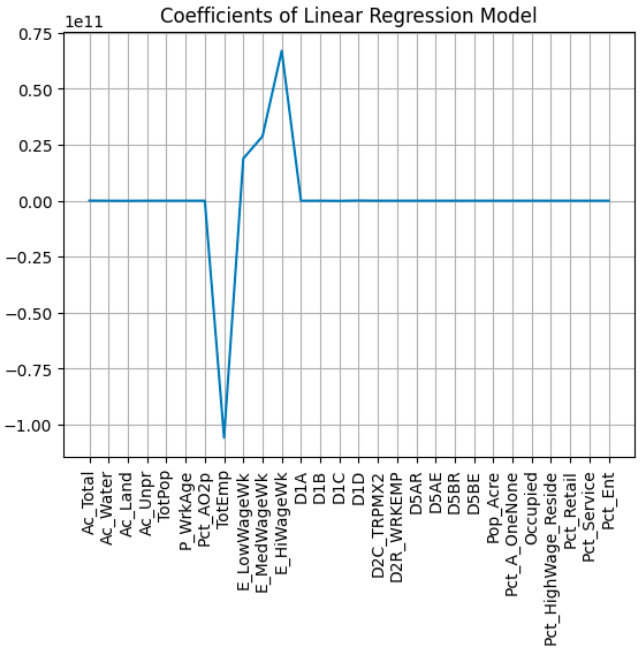

FIGURE IX
COEFFICIENTS AND FEATURES WITH MOST IMPACT FOR LINEAR REGRESSION

For ablation studies, there is little to take away from linear regression given that it's a simple model. The ablation

studies done here will show how taking away different features affects the model. The first variation done was taking away the engineered features in the data, which resulted in worse MSE and R^2. Given that our project's goal is to explore environmental and socioeconomic factors' effect on walkability score, the goal of the next two variations is to take away each of these groups one at a time and see how the model reacts. When taking away the environmental factors, the model's MSE (7.4) and R^2 (.61) return to values close to the base model.

Taking away the socioeconomic factors is a big step to show the importance of these features in our model. We chose to remove everything that has to do with wages, and social status, and keep only the environmental features. As expected, the score dropped significantly. This shows that environmental factors do have some predicting power, however, the importance lies in the socioeconomic ones.

TABLE I

COMPARISON OF LINEAR REGRESSION MODEL VARIATIONS

Model Variation	MSE	R^2
Base Model	7.34	.61
No Engineered Features	7.77	.59
No Environmental Factors	7.4	.61
No Socioeconomic Factors	10.91	.43

RIDGE REGRESSION

The Ridge Regression model implemented produced an MSE of 7.32 and an R^2 of .613. The model utilized a more diverse range of variables, with D5BE, D5BR, and D5AR being the most heavily weighted.

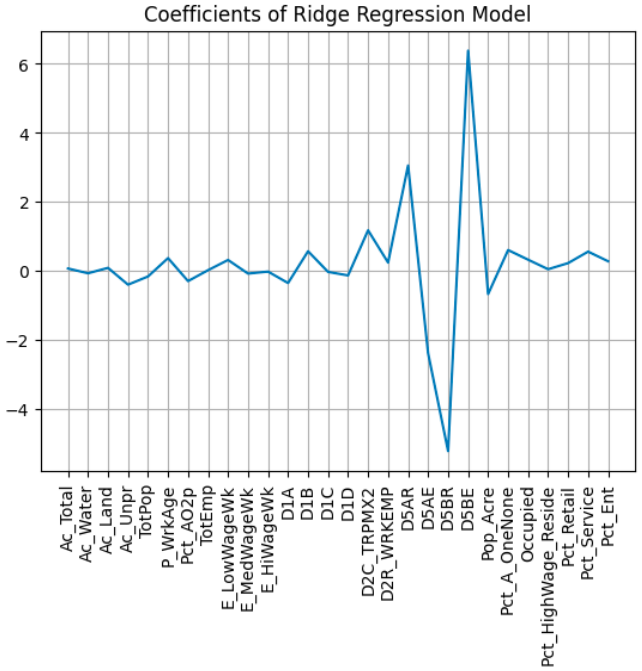

FIGURE X

COEFFICIENTS AND FEATURES WITH MOST IMPACT FOR RIDGE REGRESSION

For ablation studies, the two general approaches were to test different levels of regularization and to remove variables being passed in. Two levels of regularization were tested, one very close to 0 at .1 and one high at 100. With regularization at .1, the MSE was 7.32 and the R^2 was .616, showing not much change from baseline. With regularization at 100, the MSE and R^2 were virtually the same again. This approach shows the resistance of this model to changes in the regularization parameter.

If the dataset is not very noisy or if there are not many highly correlated features, our model may not be that sensitive to changes in alpha. Ridge regression is more effective when dealing with multicollinearity. Generally, if the model is already well-behaved, regularization may not have a substantial effect.

The next approach was to take away some of the variables with close-to-zero coefficients. When leaving 18 variables, there still was virtually no change in performance, showing an MSE of 7.32 and R^2 of .614.

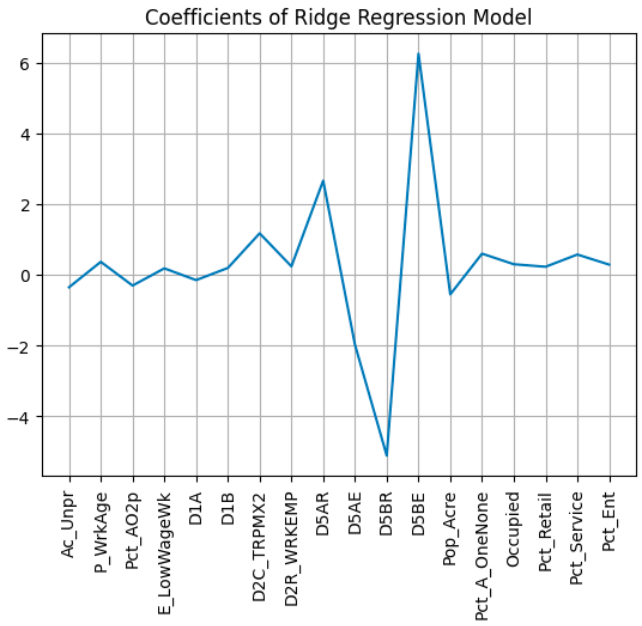

FIGURE XI

COEFFICIENTS AND FEATURES WITH MOST IMPACT FOR RIDGE REGRESSION WITH LOW IMPACT FEATURES TAKEN OUT

Pruning it down even further to just 6 features, the scores dropped, but not by a lot. It is very interesting to see that these 6 features carry most of the prediction power, which can also be seen from the graph below.

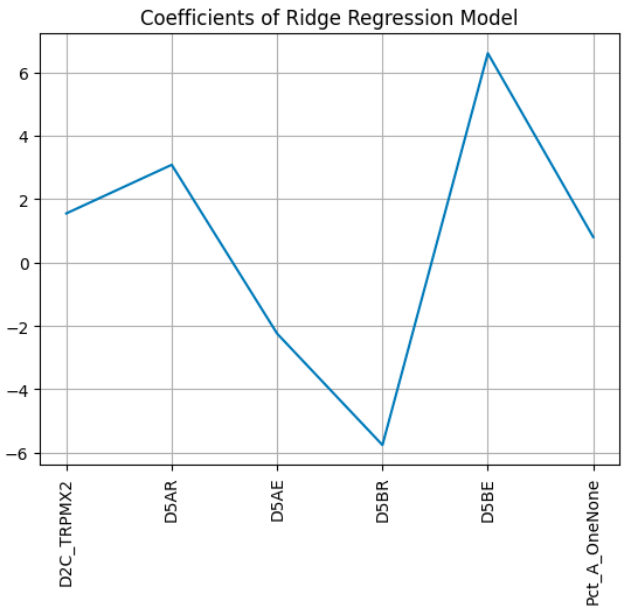

FIGURE XII
TOP 6 COEFFICIENTS AND FEATURES WITH MOST IMPACT FOR RIDGE REGRESSION ONLY

TABLE II
COMPARISON OF RIDGE REGRESSION MODEL VARIATIONS

Model Variation	MSE	R^2
Base Model	7.32	.613
Regularization = .1	7.32	.616
Regularization = 100	7.32	.616
Top 18 Features	7.32	.614
Top 6 Features	8.11	.574

Because of the simplicity of linear regression models, we chose to move away from them and explore other ones.

RANDOM FOREST

The base Random Forest model implemented used the parameters that were influenced by the bias/variance tradeoff testing. The model produced a mean MSE of 1.29 and mean R^2 of 0.932.

The ablation studies involved looking into adjusting model parameters and taking out certain groups of features.

Changing the tree depth to 10 did indeed drop the score, which indicates that regularization is too high. It overfits to the training set and generalizing to unseen data drops in comparison to our baseline model. Adjusting the number of trees to 1 essentially makes this a single decision tree. Doing this results in a worse MSE and R^2. Changing the min_samples_leaf to the default improves the time of the model but does not significantly change performance. Finally, the max_features will be changed to the default setting, which raises run time from about 10 seconds to about 40 seconds and results in an MSE of 1.35 and an R^2 of .930.

Next, three different groups of features were taken away to see how they affected scores. Taking away environmental variables doesn't change the score that much and indicates these variables are not that crucial in predicting walkability scores. Removing the combined environmental-socioeconomic factors tells a different story, however, dropping MSE to 4.26 and R^2 to .771. Finally, removing the engineered features results in an MSE of 1.52 and a R^2 of .920.

TABLE III
COMPARISON OF RIDGE REGRESSION MODEL VARIATIONS

Model Variation	MSE	R^2
Base Model	1.25	.932
Tree Depth = 10	1.40	.930
Number of Trees = 1	1.88	.901
Min Leaf Samples = default	1.35	.929
Max Features	1.36	.929
No Environmental Features	1.32	.931
No Env. or Socioeconomic	4.36	.771
No Engineered Features	1.52	.920

SUPPORT VECTOR MACHINE

The base model parameters in the SVM produced a MSE of 4.43 and a R^2 of .767. The ablation studies performed on the SVM model mainly focused on adjusting the parameters of the model to see how the score changes. Overall, none of the tweaking here produced results that were competitive with some of the other models looked at. Below you can see how tweaking the C, the Epsilon, and trying different kernels affects the SVM model.

TABLE IV
COMPARISON OF SUPPORT VECTOR MACHINE MODEL VARIATIONS

Model Variation	MSE	R^2
Base Model	4.43	.767
C = 7	6.21	.674
Epsilon = .1	6.47	.660
Linear Kernel	330.92	-16.43
Poly Kernel	208,172.09	-10,921.69

MULTI-LAYER PERCEPTRON

The MLP model proved to be one of the best of the bunch in estimating walkability. The layer, 20-neuron model produced an MSE of 1.27 and a R^2 of .933.

For ablation studies, we looked both at adjusting the parameters of the model and in removing features. Changing the Learning rate to .1 slightly worsens the score. When removing the alpha, the model slightly improves, but time increases. Even though the score with these parameters is better by 0.0005, this is a tradeoff we are taking for time purposes. The reason why the model acts this way could be because our model might not be sensitive to changes in the regularization parameter. Reducing the model to 2 neurons over-simplifies the model and greatly worsens performance.

For feature ablation, we have seen with other models that removing only basic environmental features did not change any model that much, therefore we will move towards ablating all features that have anything to do with the environment - including the number of jobs/people in that area. When doing that, the number significantly dropped, once again showing the importance of features D5ar, D5ae, D5br, and D5be. Next, the engineered features

were removed, which produced results that showed minor drops in scores. Finally, removing the socioeconomic features shows results consistent with other models as this dropped the score as well.

TABLE V
COMPARISON OF MULTI-LAYER PERCEPTRON MODEL VARIATIONS

Model Variation	MSE	R^2
Base Model	1.27	.933
Learning Rate = .1	1.86	.902
Minimum Alpha = .0001	1.26	.934
Neurons per Layer = 2	19.06	-1.498
No Environmental Features	4.29	.775
No Engineered Features	1.63	.914
No Socioeconomic Features	4.12	.784

CONCLUSION AND FUTURE WORK

Overall, this project involved looking into the Smart Location Database to determine if other factors impacted walkability besides the ones currently used in the National Walkability Index. We found several important features, including D5ar, D5ae, D5br, and D5be, and determined that using a random forest or MLP model did the best job of predicting walkability in this study.

The largest limitation of this project is that in our dataset we are using the EPA-created national walkability index as the target variable that our models are trained on and tested against. However, in the literature review, we learned that the standardized formula that creates this score has been criticized for its simplicity and inability to capture all relevant factors in what makes a location walkable. This analysis could be improved if the dataset was augmented with columns containing other types of walkability scores, such as the Walk Score or others that consider streetscape attributes. The models could then be used to compare the different types of walkability scores and analyze the variance in feature importance to their respective scores.

REFERENCES

[1] Walkability Dataset, 2021, U.S. Environmental Protection Agency, https://catalog.data.gov/dataset/walkability-index1

[2] Fernando Fonseca, Paulo J. G. Ribeiro, Elisa Conticelli, et al (2022) Built environment attributes and their influence on walkability, International Journal of Sustainable Transportation, 16:7, 660-679, DOI: 10.1080/15568318.2021.1914793

[3] Bradley Bereitschaft (2018) Walk Score® versus residents' perceptions of walkability in Omaha, NE, Journal of Urbanism: International Research on Placemaking and Urban Sustainability, 11:4, 412-435, DOI: 10.1080/17549175.2018.1484795

[4] Jim Chapman, Eric H. Fox, William Bachman, et al (2021) Smart Location Database Technical Documentation and User Guide, U.S. Environmental Protection Agency, https://www.epa.gov/system/files/documents/2023-10/epa_sld_3.0_technicaldocumentationuserguide_may2021_0.pdf