UNDERSTANDING HOW RESIDENTIAL PROPERTIES IN MADISON, WISCONSIN ARE ASSESSED

Ahsan Fawwaz (asyraf@wisc.edu) | Faris Hazim (mohamedzaimi@wisc.edu) | Imran Iskander (biniskanderg@wisc.edu) | Nick Elias (nelias@wisc.edu) | Tyler Kelly (tpkelly@wisc.edu)

[Project 16]

Introduction

Many of us are contemplating settling in Madison after graduation. Thus, we decided to analyze data on the dynamics of property valuation in Wisconsin's capital city. Two critical questions guided our investigation:

1) Can tax amounts be classified based on a residential property's assessed value?

2) Can we predict the market value of a house using available features?

We employed various machine learning techniques, including classification and regression. Our findings reveal that tax categories can indeed be accurately determined using the assessed value of residential properties. We also discovered that mainly physical features influence the market value of houses here.

Data

Our analysis uses three 2022 datasets from the City of Madison Open Data portal: *Assessor Property Information* for market values and property features, *Property Tax Roll* for tax data, and *Metro Transit by Stop* for bus stop proximity.

We merged these datasets, focusing on the variables relevant to residential properties. Our final dataset consists of 72000 residential properties and 38 columns.

Preliminary Analysis

It is essential first to have a general overview of the distribution of house values in Madison.

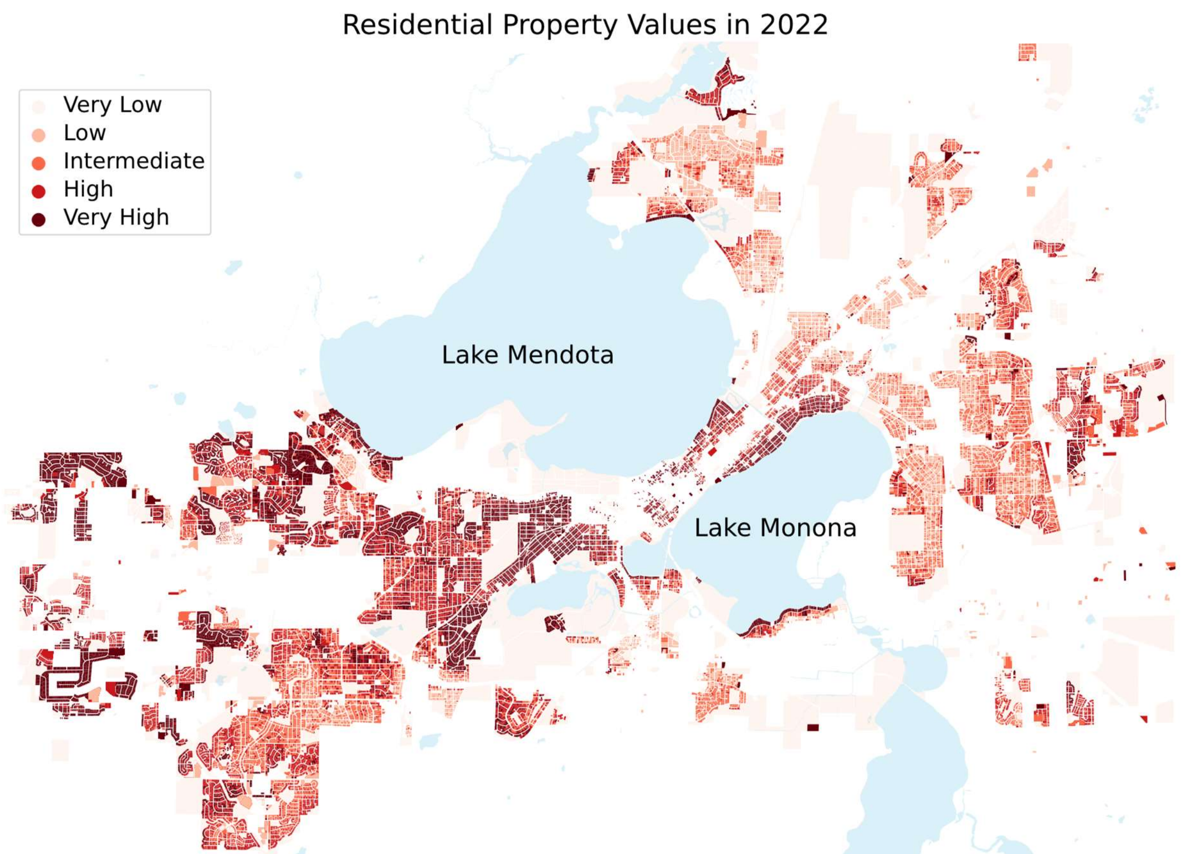

Figure 1

Figure 1 depicts that houses in west Madison are more expensive than those in the east. We assume that location is a crucial factor. Living in the west puts residents closer to the university, shopping malls, and major businesses, including Epic. Compared to the east, these people have better social and professional opportunities.

Analysis 1: Classifying Tax Amounts

In purchasing a house, we also have to pay property tax annually. For our first analysis, we sought to build a multi-class classification model that can categorize tax amounts based on the assessed value of residential properties.

We constructed our model by:

1. Splitting into 80% training, 10% validation, and 10% test data.
2. Applying standard scaling based on training data.
3. Binning the tax amounts into `Very Low,` `Low,` `Intermediate,` `High,` and `Very High` by its five quintiles. These quintiles are based on training data to prevent leakage. Additionally, quintiles make it so that the proportion of each target label is equal.
4. Utilizing grid search for hyperparameter tuning of the four classifiers we intend to compare.

Table 1 shows the accuracy scores of each model on validation data:

Decision Tree	k-NN	Logistic Regression	SVM
0.94	0.93	0.92	0.93

Table 1

We learned that our entropy-based decision tree classifier with a maximum depth of 3 was our best model, with an accuracy of 0.94. Hence, we chose this classifier as our final model. Interestingly, this algorithm also scored 0.94 on test data.

Figure 2 illustrates how the residential property taxes are classified under this model. We can confidently say that by looking at house values, we can accurately classify taxes.

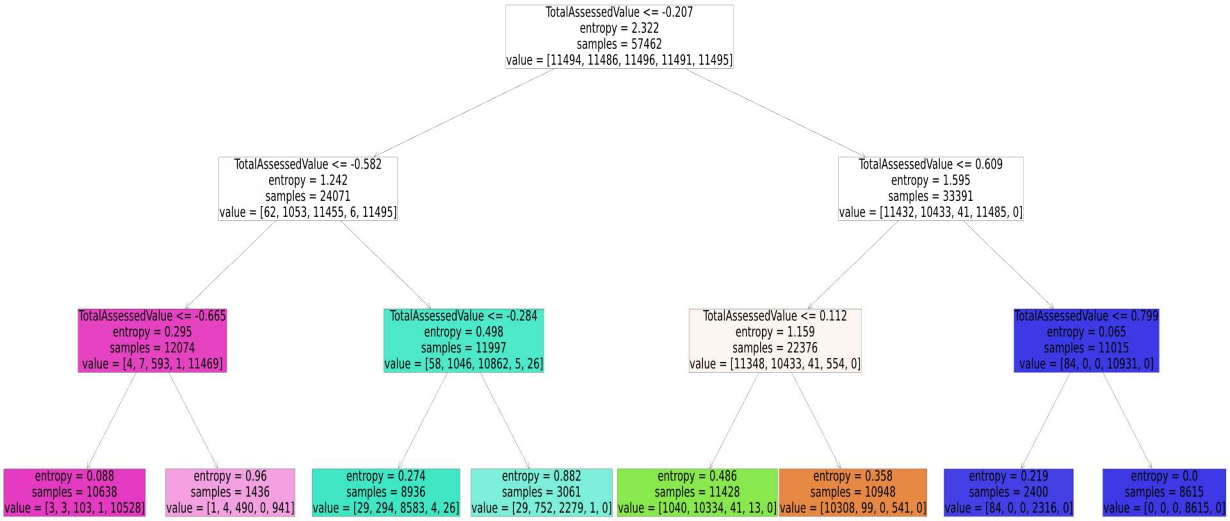
TotalAssessedValue <= -0.207
entropy = 2.322
samples = 57462
value = [11494, 11486, 11496, 11491, 11495]
TotalAssessedValue <= -0.582
entropy = 1.242
samples = 24071
value = [62, 1053, 11455, 6, 11495]
TotalAssessedValue <= 0.609
entropy = 1.595
samples = 33391
value = [11432, 10433, 41, 11485, 0]
TotalAssessedValue <= -0.665
entropy = 0.295
samples = 12074
value = [4, 7, 593, 1, 11469]
TotalAssessedValue <= -0.284
entropy = 0.498
samples = 11997
value = [58, 1046, 10862, 5, 26]
TotalAssessedValue <= 0.112
entropy = 1.159
samples = 22376
value = [11348, 10433, 41, 554, 0]
TotalAssessedValue <= 0.799
entropy = 0.065
samples = 11015
value = [84, 0, 0, 10931, 0]
entropy = 0.088
samples = 10638
value = [3, 3, 103, 1, 10528]
entropy = 0.96
samples = 1436
value = [1, 4, 490, 0, 941]
entropy = 0.274
samples = 8936
value = [29, 294, 8583, 4, 26]
entropy = 0.882
samples = 3061
value = [29, 752, 2279, 1, 0]
entropy = 0.486
samples = 11428
value = [1040, 10334, 41, 13, 0]
entropy = 0.358
samples = 10948
value = [10308, 99, 0, 541, 0]
entropy = 0.219
samples = 2400
value = [84, 0, 0, 2316, 0]
entropy = 0.0
samples = 8615
value = [0, 0, 0, 8615, 0]

Figure 2

Analysis 2: Predicting Market Values

We now concentrate on building a regression model to predict the market value of houses in Madison. As part of the setup, we:

1. Randomly sampled 15000 observations.
2. Transformed features such as school districts and neighborhood vulnerability into binary labels to reduce the number of columns.
3. Performed standard scaling on non-binary numerical variables (for feature selection purposes) and one-hot encoding on categorical features.

Next, we conducted feature selection by LASSO regression. We implemented grid search to find the best alpha value, which happens to be 10. Subsequently, this reduces the number of columns from 583 to 403.

Then, we did a similar train/validation/test split, rescaled numerical variables on the training data, performed another grid search for hyperparameter tuning, and compared the R^2 score of multiple models on validation data.

Decision Tree	k-NN	Linear Regression	SGD	Random Forest	Gradient Boosting
0.75	0.77	0.84	0.83	0.8	0.87

Table 2

Table 2 reveals that gradient boosting with Huber loss, a max depth of 9, and a 0.1 learning rate outperforms other algorithms; thus, we chose this as our final model. This regressor's R^2 score on test data was 0.88; therefore, overfitting is not an issue.

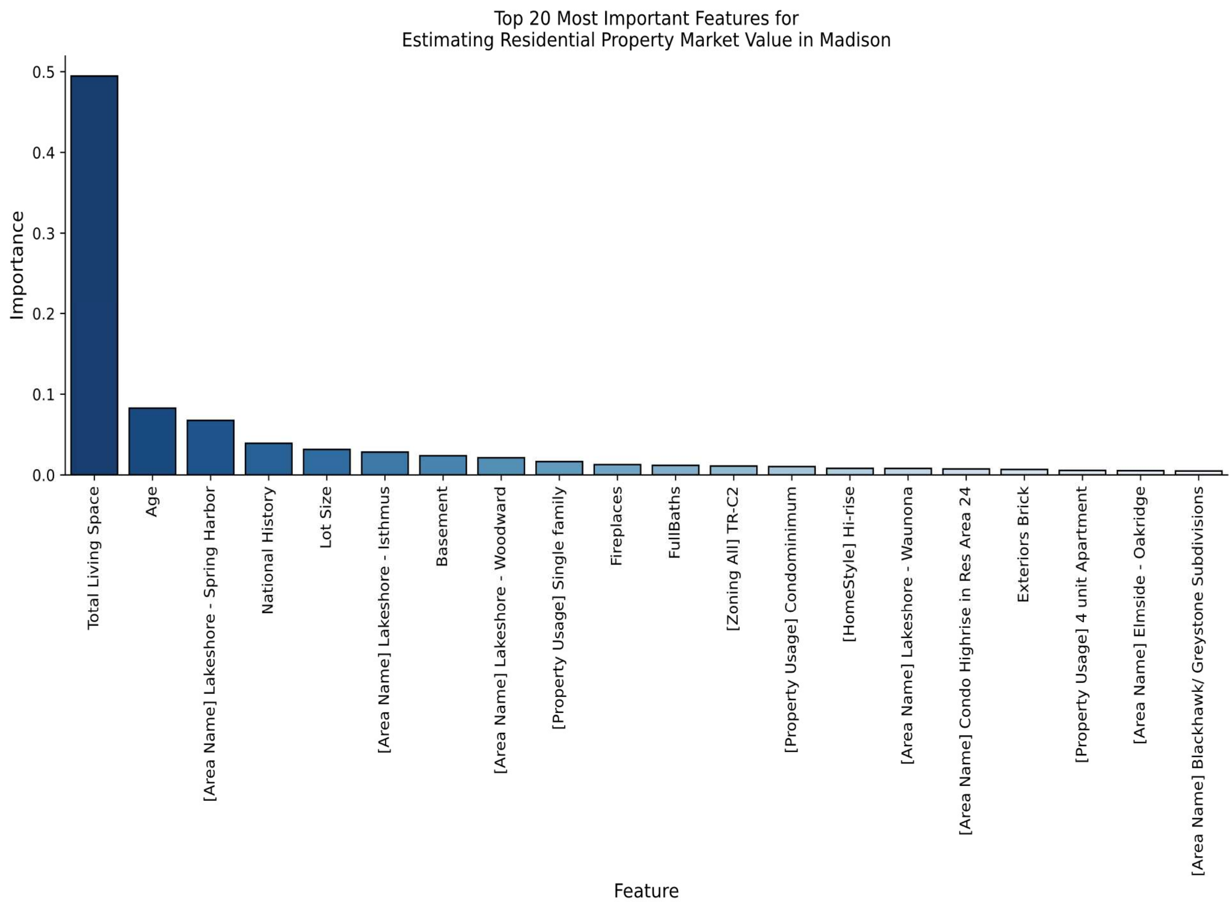

Figure 3

From the above feature importance graph, we can confirm that location is a significant contributor to the market value of houses. However, total living space is far more important than others. Interestingly, environmental factors are not as influential.

Conclusion

To conclude, we can both accurately classify tax amounts based on a residential property's assessed value and confidently predict its market value with adequate data. Looking ahead, it would be captivating to expand such analyses to other property types as well.

Table 3 summarizes our contributions to the project:

Member	Proposal	Coding	Presentation	Report
Ahsan Fawwaz	1	1	1	1
Faris Hazim	1	1	1	1
Imran Iskander	1	0.8	1	1
Nick Elias	0.1	0.3	0.4	0.6
Tyler Kelly	0.1	0.3	1	1

Table 3