Using Machine Learning to Predict Property Values in Buffalo, New York

Mateo Jácome and Kathryn Weissman, Computer Science Department, Universitat Politècnica de Catalunya - BarcelonaTech

Context: The City of Buffalo collects data for all properties within the city in order to assess their values for annual taxation. Developing a Machine Learning model to predict the property values has several challenges:

Heterogeneity: There are nearly 150 different types of properties which may include homes, businesses, schools, vacant land, parks, and structures like billboards and bridges. Multiple models would be necessary to cover all properties.

Complex Target: A subpart of the total property value is the land value, however most of the data is related to building features. Simplifying the model to predict a single target may not be nuanced enough to achieve high accuracy.

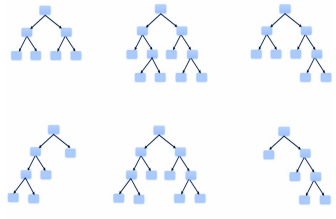
Useless Data: The score for overall condition is an intuitively useful indicator of property value, however 90% of properties have the same rating of 3 on a 5-point scale. There is not enough variety in the ratings to distinguish properties.

Outliers: Most of the homes have common characteristics and are a similar size, however models that generalize well are limited in their capacity to predict outliers.

Icon provided by Vecteezy.com

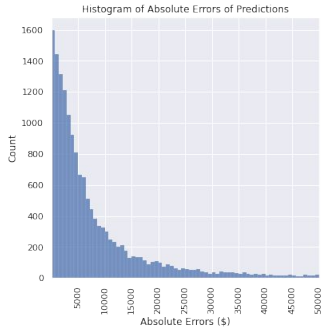
Proposal: Implement a Random Forest regression model to predict total values for the two largest property classes, including one and two family dwellings, which make up 68% of the total number of properties in the city.

Complexity: Random Forest models can become quite complex. Our final model was composed of 100 estimator trees. Our model's predictions are the average of all the trees' results, each based on thousands of decisions.



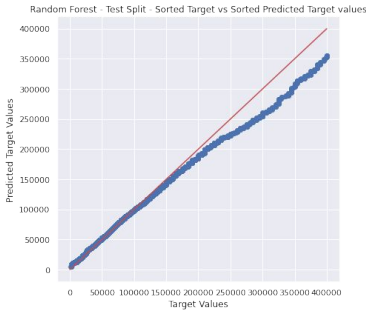
Our 100 trees had a mean depth of 43 nodes, making it a model with thousands of variables.

Accuracy: The median absolute error of predictions is around $5,000, which is 10% of the median property value of $50,000.



25% of the absolute prediction errors are less than $2,000, 50% are less than $5,000, and 75% are less than $11,000.

Limitations: The best model tends to undervalue homes when trying to predict total values that are high above the third quartile, which is around $75,000.



The best model tends to undervalue properties that are worth over $150,000.

Conclusions: Machine Learning could be used to assess property values for taxes, depending on the city's threshold for accuracy. Properties that are significantly different from the typical home are not well predicted. With further research and parameter tuning, it is likely that the model's accuracy could improve.