A Decision Tree Model and Misclassification Metric to Predict Relative Risks of Property Tax Sales in Chicago's South Side

Jonathan Zhu & Dr. Peter Jantsch

Department of Mathematics and Computer Science, Wheaton College

Abstract

Hundreds of families have lost their homes in Chicago due to tax purchasers buying their unpaid taxes at a tax sale and then charging extreme interest; this disproportionately impacts minority homeowners. Benevolence programs to help families keep their homes do exist but are not highly effective as they lack a method for reaching out to homes most at risk. As such, we developed a decision-tree model to determine relative risk levels of a property being sold and minimized model undeprediction with a modified misclassification metric. The model was the best performing among some high-performance algorithms, with 65% accuracy and the lowest misclassification; additionally, we used the model to predict risk levels of properties in an upcoming tax sale.

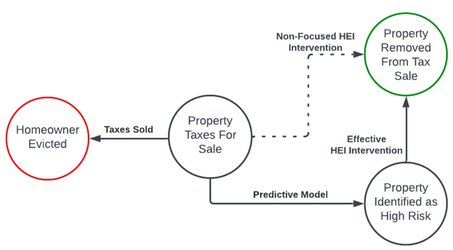
Introduction

Every year, Cook County releases a list of property owners with unpaid property taxes and puts them on a tax sale.

- Tax purchasers buy a property owner's delinquent taxes, owner pays back with **interest rates up to 36%**
- Owner cannot pay off this interest within 30 months of the sale ⇒ tax purchaser gains ownership of the property
- **Thousands of homes placed on tax sale every year; especially harmful for minority homeowners**

Organizations like Sunshine Gospel Ministries and their recently launched Housing Equity Initiative (HEI) aim to help people keep their homes. **Their Tax Sale Benevolence Program (TSBP) aims to pay off property owners' unpaid taxes and remove them from the tax sale list** (see figure).

- Several homes already kept by owner due to TSBP
- Specifically focused on benefiting the Woodlawn neighborhood in Chicago's 20th Ward on the South Side
- **No good way of identifying homes most at risk of having taxes sold** (see below)



Sunshine and HEI suggested a **machine learning model to predict a property's risk of having its taxes sold** to make tax sale benevolence more effective, i.e. reaching those most at risk.
Additonally, we sought to include:

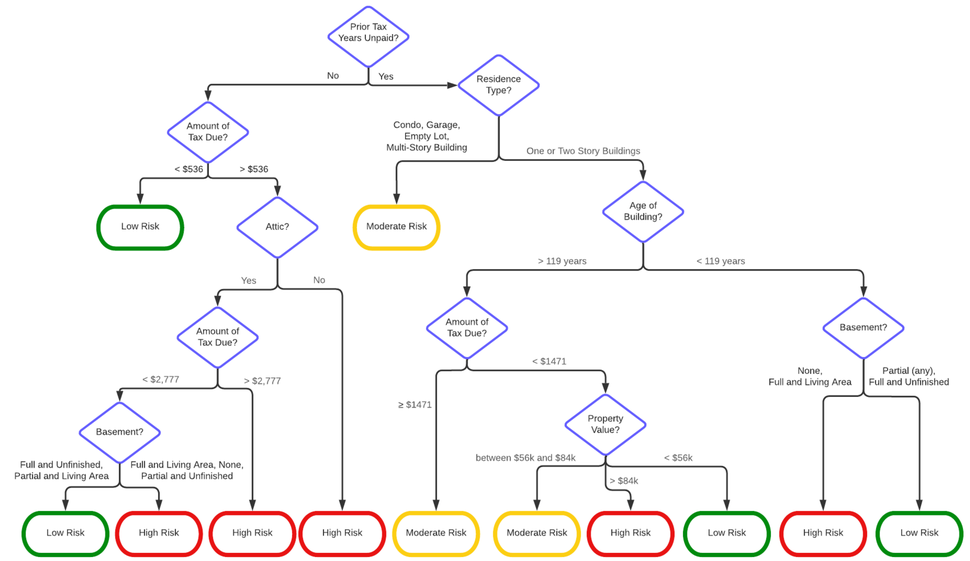
- *simplicity*, so that nonspecialists could understand the model;
- *high accuracy*, for obvious reasons, and;
- *minimal underprediction*, so that homes most at risk are not classified as low risk.

After preliminary data exploration and understanding, **we chose and finalized a decision tree model** as a deliverable to Sunshine and HEI.

Decision Tree Model

The Decision Tree algorithm was chosen for its simplicity and its ability to take categorical data as inputs.

- Decision Tree algorithm splits data at each split in the tree until all splits have the same class
- Data: Properties from 2019 Tax Sale with predictor variables as property and tax characteristics scraped from Cook County Assessor's and Treasurer's websites, outcome variable as relative risk level, classified as low, moderate, or high (done by hand for development data)
- Data first preprocessed to remove identifying variables and variables with heavy class imbalance
- Data split into training and testing sets (80% and 20% of the full set, respectively)
- Decision tree fit to entire training dataset with `rpart` R package [3], tuned with the `rpart.control` function, and the first few layers of the tree were plotted with the `rpart.plot` package and function [2]



Metric Development

Accuracy alone is not necessarily a good assessment metric; here, accuracy fails as it does not account for over- and under-predicting. For risk assessment, underpredicting is far worse than overpredicting.

- Wanted a custom metric to account for accuracy and under- and over-predicted items
- Modified a procedure by George et al. [1] to compute a cost matrix

Based on our isolated training set, we compute the cost matrix here by the following element-wise matrix multiplication:

$$C = \begin{bmatrix} \frac{\sum_{i \neq j}^{M} n_i}{n_i} \end{bmatrix} * \begin{bmatrix} 0 & 0.2 & 0.4 \\ 1 & 0 & 0.2 \\ 2 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0.3628 & 1.0698 \\ 1.4881 & 0 & 0.3194 \\ 6.1142 & 2.2286 & 0 \end{bmatrix},$$

where M is the number of classes and n_i is the number of training samples in class i, for all $i \neq j$. We then determine the total misclassification cost metric (TC) as

$$TC = \sum_{i=1}^{M} \sum_{j=1}^{M} c_{i,j} * f_{i,j},$$

where $c_{i,j}$ is the component in the i^{th} row and j^{th} column of C; similarly, $f_{i,j}$ is the same for F, the model's confusion matrix. Good model performance corresponds to a lower TC value, and the TC values are best understood when compared to the maximum cost; based on our isolated testing set, we determine the maximum cost as 108.48.

Results

Accuracy and total misclassification cost (TC) for the decision tree and other algorithms used in preliminary model development are calculated to be:

Algorithm	Accuracy	TC
Decision Tree	**0.658**	**13.46**
Neural Net	0.579	57.229
Support Vector Machine	0.632	15.712
Random Forest	0.579	31.46
Boosted Forest	0.579	52.336

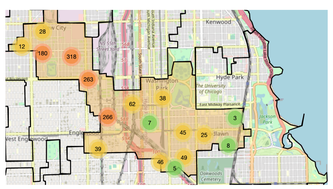
Additionally, the decision tree model was also applied to the list of properties from the 2020 Tax Sale.

- Predicted 295 high risk, 333 moderate risk, 186 low risk
- Sunshine and HEI responded favorably and hope to use the predictions in the upcoming tax sale

Conclusion

Overall, the decision tree model was shown to be effective by multiple metrics and relatively understandable. The modified misclassification metric also performed as expected, penalizing underprediction well. The model is applicable to future tax sales but is not incredibly robust due to a low number of training samples. Therefore, next steps include:

- Making model more applicable long-term, possibly using past tax sale data if available
- Improving model performance (ideally 80% accuracy)
- Extending model efficacy to Chicago's 49 other wards; requires additional cooperation with nonprofits for real-world impact



Above: Chicago's 20th Ward, the focus of Sunshine Gospel Ministries' TSBP, and general locations of the properties at risk of tax sale (*Figure credits: Claire Wagner*)

References

Nysia I George, Tzu-Pin Lu, and Ching-Wei Chang.
Cost-sensitive performance metric for comparing multiple ordinal classifiers.
Artificial intelligence research, 5(1):135, 2016.

Stephen Milborrow.
rpart.plot: Plot 'rpart' Models: An Enhanced Version of 'plot.rpart', 2022.
R package version 3.1.1.

Terry Therneau and Beth Atkinson.
rpart: Recursive Partitioning and Regression Trees, 2022.
R package version 4.1.16.

Acknowledgements

We would like to thank Sunshine Gospel Ministries, the Housing Equity Initiative, and the BLOCK movement with special thanks to Janice Miller, Kimberley Salley, and Tracy Stanciel for their direct involvement. We also thank the research team of Dr. Paul Isihara and his students Claire Wagner and Samuel Carlson for their collaboration.