ILLINOIS
Gies College of Business

Big Data Analytics Final Project

WHAT IS IT WORTH? PROPERTY ASSESSMENT CHALLENGE FOR COOK COUNTY ILLINOIS

Authors: Riley League
Updated: Fall 2024
Prepared for: FIN 550 Big Data Analytics

1 CASE BACKGROUND

Property taxes account for a large portion of revenue for local governments, allowing them to fund police, schools, pension liabilities and many other services provided to citizens. In the United States, property taxes are collected at the county level, requiring owners to pay a sum that is based on tax rates, along with levies for their locale, multiplied by the estimated or assessed value of the property. It is the role of a government entity, commonly knowns as the Assessor, to determine the fair market value of each property.

For this project we use the example of the Cook County Assessor's Office (CCAO), which is responsible for the second most populous county in the United States, and home to the City of Chicago and over 130 other municipalities. To get a sense of the magnitude of the task for this organization, in FY 2019, CCAO was in charge of valuing over 1.8 million properties, totaling$15.58 billion in taxes.

Determining what is a fair market value with this diversity of parcels that are spread across over 130 municipalities is one of the most complex valuation tasks in the United States. Despite its complexity, prior to 2018, CCAO's valuation process was obscure and inaccurate. Compounded with unusually high tax rates and the flawed modeling process, CCAO, which is overseen by an elected official, faced intense public and political pressures. One outcome was the surprising, yet decisive win in the 2018 election of Fritz Kaegi, who promised to reform CCOA and make it transparent.

To address this challenge, the new Assessor invested in technology, has increased the transparency of data and has formed and later expanded the data science team in the agency. The modeling techniques underlying the valuation pipeline have also shifted from old underperforming technology using SPSS, to a more capable implementation in R. With respect to transparency, CCAO is making all their choices, including modeling and data decisions, publicly available.

For additional background and context on the data science efforts at CCAO, review the documentation the data science team at CCAO generated, especially for their machine learning models. Additionally, view the video recording of the presentation CCOA's data scientist Dan Snow gave last year about their machine learning modeling and data.

CCAO data documentation is located on GitLab.

CCAO video presentation on valuation methods, models, and results is located on YouTube.

2 CASE DESCRIPTION

You have been recruited to join the CCAO data science team. Your first work assignment is to assess the residential property value (i.e., you are to predict the market price of the home, if it were to sell) given a set of features. For this endeavor, you have been provided with three files (found in the data.zip file):

- "**predict_property_data.csv**" contains data on 10,000 properties whose values you are to assess. Each row in the dataset is a property. The variable "pid" contains a unique identifier for each property.
- "**historic_property_data.csv**" contains data on 50,000 other properties that have recently sold, including their sales price (variable "sale_price"). Each row in the dataset is a property.
- "**codebook.csv**" describes the variables included in each property file.

Objective. Your objective is to predict the value of each home in "predict_property_data.csv" as closely as possible to its actual value. You will produce and submit a data file that contains the property ID (variable "pid") and your value assessment for each of these properties (see the deliverables section below for how your data file must be formatted).

We will assess the quality of your work by comparing your property value predictions to the actual sales price when these homes sell and computing the Mean Squared Error (MSE) of your predictions. Your objective is to minimize the MSE of your predictions.

Approach. You can use the data in "historic_property_data.csv" to build models of the relationship between property characteristics and their value, as reflected in sales prices. All analysis, including training and selecting models and making predictions, should be performed in R and the code should be placed into a single script. You will submit this script as supporting documentation to your written report.

Tips. Here are a few suggestions to help guide your analysis.

- You are encouraged to use any modeling methods covered in the class. You may use other modeling methods not covered in class, but doing so is not expected and we cannot help you if you encounter technical issues. You will be graded primarily on how well you explain and justify your approach to generating predictions. Using methods covered in class is sufficient for receiving full credit on this assignment.
- You are encouraged to start by building simple models that use only a few predictors. Including many predictors can (but will not always!) improve the quality of predictions. Adding many predictors will increase the technical challenges you may face, especially for categorical variables that have many levels. We advise that you avoid starting with "kitchen sink" models of the form `Y ~ .`, i.e., models that throw in all available predictors. It is also advisable to know the variables you include in your models. Referring to the codebook will help you gain familiarity with the variables in the project datafiles.

- Make sure to take care when handling variables with missing values!

3 CASE DELIVERABLES

You should **deliver the following three files** as part of your final project:

1. **Executive Summary**. Using the Word template posted to Canvas, produce an executive summary report in PDF format of your approach and findings. Final project results will be reported publicly to the class, so give your project an appropriate Team Name to preserve your identity. The document should then have three sections: a concise overview of the case and objectives, a description of your methodology, and an actionable conclusion which should include a summary of your recommendations to the Cook County Assessor's Office, along with an estimate of the accuracy of your models. Your executive summary needs to be thoughtful, but it should not be more than 3 double-spaced pages, font 12, in length. Tables and figures should be placed at the end in an appendix. The appendix will not count towards the page limit.

2. **R script**. Turn in the R script used to perform your analysis. It should be designed such that if the data files are in the same folder as the script we can run your script and generate exactly the same results. (Depending on the analysis you use, you may need to set the random seed at the beginning for this to occur.)

   Annotate your R script with brief comments and follow recommended R programming style practices. **Place a comment at the top of the script with how much time a full run takes.**

3. **Assessment File**. Export your property assessment values (predictions) into a csv file called "assessed_value.csv". The file should contain one record for each property included in "predict_property_data.csv" and should contain only two variables:

   - "pid" (the property identification number). This column should be identical to the pid column in "predict_property_data.csv" which consists of consecutive integers from 1 to 10,000.
   - "assessed_value" (your predicted property value). Follow these guidelines:
     - The values in this column should be of a numeric format, without any special characters (e.g., save values as 100000, NOT $100,000).
     - Ensure that all pid and assessed_value values are non-missing and non-negative before submitting your file.
     - The final file should be formatted as follows (assessed values will differ):

pid	assessed_value
1	156787.00
2	2879298.00
3	365198.00
…	…

We will compile each team's product and will report to the class the predictions versus actual property value realizations for each team, and we will see which team realizes the lowest MSE!

4 CASE GRADING

1. **Executive Summary (40%)**. Your grade on this component will be based primarily on your ability to clearly communicate your objectives, methodologies, and results. Avoiding jargon, using correct grammar, and using proper sentence and paragraph structure will all be taken into consideration.
2. **R script (40%)**. Your grade on this component will be based on three factors. First, how easy is it to follow your script? Clearly commenting your code and using informative naming conventions will help. Second, are your results replicable? If the raw data files are in the same folder as your script, we should be able to run your code without any modification and replicate the results in your final project. Third, is your code reasonably efficient? Code that takes an inordinate amount of time to run may negatively impact this component grade.
3. **Assessment File (20%)**. Your grade on the assessed value file will be based on valid formatting and the MSE your prediction generates. A valid assessed_value.csv file should contain 10,000 records (pid should take on consecutive values from 1 to 10,000) and the assessed_value variable should be a non-missing numeric value for each record. If the file is properly formatted and complete, the lowest grade you can receive on this component is a B+.

NOTE: As a special hat-tip, the project achieving the most accurate valuation in the class will get an A+ on the Assessment File component.