**The Assessment is Too Damn High: Overvaluation of Bungalows in the Bell School Area of Chicago**

**Ross Hyman**

**A picture containing building, outdoor, house, grass

Description automatically generated**

**Introduction and Problem Statement**

According to the Cook County Assessor, the estimated market value of my home, a one-story ninety-six-year-old Chicago bungalow in the Bell School area of Chicago jumped from $343,360 in 2020 to $720,000 in 2021. Many of my neighbors received similar shocking assessments. A quick search on Redfin, which I conducted upon receiving my assessment, shows that in 2018, three bungalows in our neighborhood sold for $300,000, $485,000, and $510,000. One sold in 2021 for $445,000. How is it possible for the Assessor’s estimated market value of our homes to be so much higher than the values that they are being sold for? This is the question that was on my mind and wallet when I was asked to choose a capstone project for the data science course I am taking at Springboard. This is the report for that project.

**Hypotheses for Overassessment**

I approached this project with two hypotheses for the causes of the overvaluations. I developed a third after an analysis of the data. I start with the fact that the Bell School area is rapidly gentrifying. Amongst the bungalows in our neighborhood, are many mansions that have sold for over a million dollars each. These mansions were each built on land that used to contain a bungalow built for working class families almost a hundred years ago. Bungalows are sold to developers, those with severe settling are torn down, and mansions built in its place. This leads a bimodal distribution of homes. One reason for the overvaluation that I consider is that the feature set used by the Cook County Assessor’s office includes some economic and demographic features that are used for each home in the same census tract, such as median income. It could be that homes in our heterogeneous community, in which some homes are sold for $300,000 and some for $1,100,000, are being compared to homes in a more homogeneous community, where most homes are being sold for $700,000, because of sharing the same economic and demographic features. To prevent this, I model the value of homes in the Bell School area by training my model only on home sales in the Bell School area. A second possible source of overvaluation is that although machine learning models should determine the price of bungalows using only the sales price of other bungalows, no method is perfect and mansion sales could be impacting the predicated price of bungalows. To prevent this, I consider models that determine the value of newer homes only on the sale prices of newer homes and that determine the value older homes only on the sales prices of older homes. Lastly, I discovered by examining the Cook County Assessor’s data file for the homes in the Bell School area, that the data for many of the mansions includes feature values common for bungalows. It appears that for these mansions, the mansion’s sales value was correctly recorded but the data describing the features of the home were not updated. This incorrect data, fed into the Assessor’s machine learning model, causes overvaluation of bungalows and undervaluation of mansions.

**Implications**

Assessed values are used to determine property taxes. Assuming sales price is related to income, bungalows are owned by relatively lower income people and mansions owned by relatively higher income people. Overvaluing bungalows and undervaluing mansions, therefore, has the effect of over-taxing lower income people and under-taxing higher income people. There are many neighborhoods in Chicago undergoing the same kind of rapid gentrification experienced in the Bell School area, in which bungalows and other working-class cottages are being torn down and replaced by mansions. The same overvaluation of low-priced homes and overvaluation of high-priced homes that is present in the Bell School area homes values is therefore potentially present for all these neighborhoods. In addition, for many of these neighborhoods, the bungalows and working-class cottages are predominantly owned by people of color. Therefore, the overvaluations could have disparate impact on people of color making it potentially actionable under the civil rights act.

**Data Source**

The Assessor’s data set was downloaded from the Cook County Assessor’s data catalog at <https://datacatalog.cookcountyil.gov/Property-Taxation/Archive-Cook-County-Assessor-s-Residential-Modelin/8f9d-wy2d>.The Assessor also maintains a Gitlab repository at <https://gitlab.com/ccao-data-science---modeling/models/ccao_res_avm>

**Background**

Property taxes are a main revenue source for county, schools, and city governments in Illinois. The property tax on a house is a percentage of the value of the house. Since the sales price of a house is only determined when it is sold and only a fraction of homes are sold each year, assessors use machine learning models to form a prediction for the value on all homes by training the model on those houses that have been sold.

**Property Tax is Regressive**

A tax is considered regressive if, when interpreted as a tax on income, people of lower incomes are taxed at a higher rate than people of higher income. Likewise, a tax is considered progressive if people of higher incomes are taxed at a higher rate than people of lower income.

The data provided by the Cook County Assessor includes the property tax paid for each home and the median family income for the census tract of each home. From this data, we can calculate an approximate effective income tax rate for the median household in each census tract by calculating the median property tax paid for a given median income divided by household median income. The result is an approximation because the median household income and the median home sale need not be for the same household.

A graph of the approximate effective tax rate vs income is shown below. The data appears to be made up of several separate curves, each curving downward. The low-income earners on each curve have a higher effective income tax rate than high income earners, indicating that the property tax is regressive.

Chart, scatter chart

Description automatically generated

The reason for the separate curves is that that Cook County is composed of many cities and taxing districts, each taxing the assessed value of properties at different rates. The data is consistent with the proposition that the modeled prices of the homes of low-income homeowners is a higher fraction of their income than the modeled price of the homes of higher income homeowners. Property taxes are therefore an example of a regressive tax.

A common measure of the regressiveness is found by determining the best fit linear regression line to the data. The slope of the best fit line is -0.01% per $1000 meaning that for every additional one hundred thousand dollars of income, the effective income tax percentage the homeowner pays goes down by 1 percent. However, the regressive nature of property taxes is greater than the linear coefficient implies because the curves are visibly non-linear with a steep negative slope much steeper than -1% per $100,000 at low incomes and much smaller at high incomes.

**Regression is Regressive**

There are two possible contributing factors to the modeled price of low-income homes being a larger fraction of income than high-income homes. One is that the actual sales prices of low-income homes are a larger fraction of income than high-income home. A second reason is that the Assessor’s model could overvalue low-priced homes and undervalue high-priced homes. As machine learning models are based on averages of similar homes, all such models will tend to overvalue low priced homes and undervalue high priced homes. This is so basic to the linear model that it is named linear regression. However, non-linear models also exhibit regression as they are also based on averaging. Modern machine learning models should effectively weigh the sales prices of homes in the averages according to the closeness of features to reduce the effect of regression to the mean. However, no model is perfects and if the data is inaccurate, so that the value of low-priced homes is determined from the sales price of high prince homes, machine learning models will fail no-matter how sophisticated. Garbage in, garbage out.

**Homes in the Bell School Area**

In the Bell School area of the city of Chicago in Cook County, relatively-low-priced bungalows are sold to developers, torn down, and relatively-high-priced mansions are built in their place. This leads to a bimodal distribution of homes in which most homes are either more than eighty years old or less than twenty years old.

An anecdotal sampling of recent sales in the Bell School neighborhood, which I conducted to appeal my home’s property assessment, shows a significant over-valuation of the bungalows and under-valuation of the mansions. In particular, three bungalows were assessed at 124%, 126% and 148% of their sale prices which were $510,000, $525,000 and $445,000 respectively. Two mansions were assessed at 70% and 86% of their sale prices which were $1,025,000 and $1,225,000 respectively. My motivation for this project was to see if this bias is systemic and if it can be minimized. I applied different models to predict Bell School area home prices with the goal of minimizing this bias.

**Criteria for success**

The model used by the Cook County Assessor has the following characteristics:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model Type | R2 | RMSE | MAE | MAPE |
| Linear | 0.77 | $153,764 | $89,910 | 29% |
| LightGBM | 0.84 | $125,977 | $67,657 | 24% |

The metrics are the mean absolute percentage Error (MAPE), the coefficient of determination (R2), the root mean square error (RMSE) and the mean absolute error (MAE). Recognizing that the MAPE was comparable to the errors I was seeing, my initial goal was to use MAPE as a metric to minimize to determine an optimal model see if that decreased the overvaluation of bungalows.

**Stratifications**

As many people chose homes according to the local elementary school, and the perceived quality of schools strongly varies even between neighboring schools I predict home values in the Bell School neighborhood by training only on home sales in the Bell School neighborhood.

Since old bungalows are being systemically demolished to make way for new mansions, we hypothesize that the sales price should be only weakly correlated with home features for bungalows compared to features for mansions, since to a first approximation, the bungalows are only being sold for their land value. To this end we also consider models with a further stratification separating homes that are older and younger than 80 years.

A significant downside to stratifying by elementary school is that the number of homes sales within a school boundary can be small. For the Bell school district this is only 293.

**Features**

The data provided by the Assessor’s office includes many features for each home. Features can be divided into those that are features of the individual home, such as the number of rooms and the number of square feet of the lot and of the building, and features of the home’s neighborhood such as the median income of the census track, the school district, and demographic racial percentages in the neighborhood. Furthermore, the features for each home can be divided into those for which there is independent verification through permits and title documents such as square footage and number of rooms, and for which the features are self-reported and there are no confirming documents such as if the basement is finished. For this project I focus on a model that incorporates verifiable features.

**Missing Feature Example: Bricktown Clay Pit Garbage Dumps**

Every neighborhood in Cook County is unique. Some have features that impact the price of homes that are missing from the Cook County Assessor’s feature set. An example of one such feature that impacts the price of homes in the Bell School area are the Bricktown clay pit garbage dumps. Western Avenue, a North-South street that cuts through the Bell School area, is so named because it used to be the Western edge of the city. The area centered around Grace Street West of Western Avenue and East of the Chicago River was the source of clay for Chicago’s brick mills which were in high production to rebuild the city after the calamitous 1871 Chicago fire. The area was so dominated by the brick industry’s clay pits and brick mills that it was referred to as Bricktown. Deep pits for extracting clay throughout the area were dug. When all of the clay was extracted, the brick industry moved on and the pits became the city garbage dumps. They were then covered over, and in the early 1900’s ago, a community of working-class bungalows were built over the covered garbage-filled clay pits. Since that time, some of the bungalows have been sinking into the ground.

The bungalows of Bricktown compose the western end of the Bell School area. Bell Elementary School is one of the highest rated elementary schools in Chicago. So high-priced housing in the neighborhood is in high demand. For bungalows on firm ground, a developer can chop off the attic of the bungalow and build a second floor using the existing bungalow’s foundation. But for bungalows on sinking ground, the entire bungalow must be demolished, and pilons pounded into the earth until they reach solid ground with a new foundation built on top of the pilons to support a two- or three-story house. In these cases, the internal features of the house have no effect on the selling price as the property’s only value is in the land. Machine learning models trained mostly on homes that are sold to live in, will greatly overvalue these bungalows. I do not have data on the places where air pockets or less dense garbage has led to more settling. However, by stratifying by the Bell School area and homes older than 80 years old, our model concentrates these homes together so that the value of a home in this group will be determined from an average dominated by similar homes.

**Data Cleaning**

The data is converted from parquet form to a Pandas Data Frame. The sales data and test data is checked to insure they have the same column names. The data is filtered for “BELL” in the 'geo\_school\_elem\_district' feature. Also filtered for are numerical features for which there is verifiable data. These include features for: year of sale, sale price, land area, building area, age of house, number of rooms, number of bedrooms, number of full bathrooms, number of half bathrooms, number of fireplaces (determined by the number of external flues), longitude, and latitude. In addition, data from a property classification category feature is used to create new category features to indicate if a home is one story or more than one story, and a new category feature to indicate if a home is attached or not.

Median income for a census tract is a feature that is used in the Assessor’s model that we exclude from our model on the grounds that its effects, in the case of the Bell school area are regressive. Representing the income distribution in a tract by its median value exclusively, has the effect of making the bimodal distribution of Bricktown with its relatively-lower income bungalows and relatively higher-income mansions comparable to census tracts with a monomodal income distribution with the same median as Bricktown’s. This amplifies the tendency towards regression towards the mean that we are endeavoring to minimize.

Only single-family residences are retained and the resulting data set is checked for missing and values.

**Exploratory Data Analysis**

A screenshot of a computer

Description automatically generated with low confidence

Histograms of features were examined. The bimodal distribution in the age was noted as consistent with the history of the community.

A picture containing background pattern

Description automatically generated

A heat map shows that age has the strongest ant-correlation with sales price while building area, number of full bathrooms, and number of bedrooms have the strongest positive correlations.

Diagram, engineering drawing

Description automatically generated

Graphing the sales price against numerical features we can see that the lot sizes are nearly constant, the older age homes contain a subset of low-priced homes that are lower priced than any of the newer homes, that the strongest linear correlation is with building area, and that location within the Bell School area is not very correlated with sale price.

The graph of sales price vs age of home is surprising because it shows many old homes selling for over a million dollars. This is an indication that there could be something wrong with the data. How likely is it that people would spend over a million dollars for homes originally constructed for working class people over eighty years ago? We flag this potentially very serious issue for now and first present results assuming that the data set is accurate. Later we question this assumption.

**Modeling**

The following are metrics for several sklearn machine learning models applied to the full Bell School data set including the potentially incorrect rows. Models considered are linear regression, elastic net, random forest, and gradient boosting. Grid search is used over hyperparamers of the later three methods to find the model that minimizes the mean absolute percentage error (MAPE). We consider the MAPE the relevant metric to minimize because it is the relative not the absolute error that is important for determining errors in the effective income tax rate. We also present R2, the root mean square error (RMSE) and the mean absolute error (MAE).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model Type | Train or Test | MAPE | R2 | RMSE | MAE |
| Linear | Train | 0.17 | 0.80 | $195,059 | $153,166 |
| Linear | Test | 0.21 | 0.60 | $290,606 | $213,838 |
| Elastic Net | Train | 0.17 | 0.80 | $195,602 | $155,055 |
| Elastic Net | Test | 0.21 | 0.61 | $288,069 | $211,546 |
| Random Forest | Train | 0.07 | 0.97 | $77,737 | $54,946 |
| Random Forest | Test | 0.16 | 0.67 | $265,565 | $165,065 |
| Gradient Boost | Train | 0.11 | 0.87 | $159,320 | $107,173 |
| Gradient Boost | Test | 0.16 | 0.69 | $257,753 | $179,624 |

There are only 293 rows in the data set. The training set includes 234 sales, 80% of the total, and the test set contains 59 sales, 20% of the total. Because of the small numbers, the data distributions of the training and test sets are not close. One can see this in the dramatic differences in the metrics for the training and test sets for the same method. Nonetheless, the MAPE, is less than for the Cook County models. The other metrics are higher. This is related to the fact that my data set is small and that, aside from MAPE and R2 they determine error from absolute instead of relative differences. This makes my metrics higher because the average sale price in the Bell School area is higher than the average sale price in the County. Random Forest and Gradient Boost are a small improvement over linear regression and elastic net. I presume this is because Random Forest and Gradient Boost are decision tree methods so they can treat features differently for bungalows and mansions.

**Bias Metrics**

To see if the models are overvaluing low priced homes and undervaluing high priced homes we determine bias metrics recommended by the International Association of Assessing Officers (IAAO). These metrics are all relative metrics based on the ratio of the predicted value divided by the actual sales value. The metrics are the Coefficient of Dispersion (COD), the Price Related Differential (PRD), and the Coefficient of Price-Related Bias.

The Coefficient of Dispersion (COD) is a measure of the variance of the ratio that is very similar to MAPE. Its value is the percentage variation of the ratio away from the median (instead of the mean as it is for MAPE). It does not tell us if low priced homes are being treated differently than high priced homes.

The Price Related Differential (PRD) does tell us if low priced homes are being treated differently than high priced homes. It is the mean ratio divided by the sales price weighted mean ratio. If the ratio for large sales price homes is lower than the ratio for low sales price homes, then the mean ratio over the weighted mean ratio will be larger than 1.

The Coefficient of Price Related Bias (PRB) also tells us if low priced homes are being treated differently than high priced homes. It is the coefficient of liner regression between the log2 of the price and the ratio. A negative number means that if one increases the price, the ratio goes down. As with the case of effective income tax rate, linear regression coefficients can downplay bias if the curves they are modeling are non-linear.

The IAAO recommends calculating these metrics after removing outliers in a prescribed way. I do not remove outliers because my data set is small. Therefore, my metrics show more bias than if the metrics were calculated using official IAA methods.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model Type | Train or Test | COD | PRD | PRB |
| Linear | Train | 16.842 | 1.047 | -0.136 |
| Linear | Test | 19.131 | 1.066 | -0.137 |
| Elastic Net | Train | 16.779 | 1.049 | -0.149 |
| Elastic Net | Test | 18.949 | 1.067 | -0.149 |
| Random Forest | Train | 6.413 | 1.021 | -0.096 |
| Random Forest | Test | 16.540 | 1.067 | -0.209 |
| Gradient Boost | Train | 6.448 | 1.022 | -0.100 |
| Gradient Boost | Test | 16.054 | 1.068 | -0.206 |

There is no simple intuitive relationship between the error measures MAPE and COD and the bias measures PRD and PRB. Indeed, it appear that methods that have a decreased error have an increased bias. We now turn to age stratification models.

**Age Stratification Models**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model Type | Train or Test | MAPE | R2 | RMSE | MAE |
| Linear | Train | 0.16 | 0.82 | $189,778 | $143,160 |
| Linear | Test | 0.17 | 0.70 | $238,215 | $174,463 |
| Elastic Net | Train | 0.16 | 0.81 | $193,862 | $144,434 |
| Elastic Net | Test | 0.16 | 0.72 | $231,015 | $165,815 |
| Random Forest | Train | 0.06 | 0.97 | $76,746 | $53,386 |
| Random Forest | Test | 0.15 | 0.77 | $208,382 | $143,301 |
| Gradient Boost | Train | 0.11 | 0.87 | $159,320 | $107,173 |
| Gradient Boost | Test | 0.16 | 0.69 | $257,753 | $179,624 |

Determining the value of older (newer) homes only on the sales price of older (newer) decreases MAPE but only about one percent.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model Type | Train or Test | COD | PRD | PRB |
| Linear | Train | 15.717 | 1.040 | -0.095 |
| Linear | Test | 17.822 | 1.021 | 0.021 |
| Elastic Net | Train | 15.385 | 1.049 | -0.146 |
| Elastic Net | Test | 16.497 | 1.031 | -0.033 |
| Random Forest | Train | 5.755 | 1.019 | -0.076 |
| Random Forest | Test | 15.353 | 1.039 | -0.105 |
| Gradient Boost | Train | 7.518 | 1.028 | -0.095 |
| Gradient Boost | Test | 15.233 | 1.034 | -0.071 |

Stratifying by age has a much larger impact on PRB cutting it by half.

**Predictions**

We now compare our predicated sales values of bungalows to those of the Cook County Assessor. We use our optimized Random Forest model as it has small COD and PRB for the training and test sets. The following table shows the Cook County assessments for several bungalows in my neighborhood assessed by the Cook County Assessor at much higher values than the actual sales values for bungalows in the neighborhood.

|  |  |  |  |
| --- | --- | --- | --- |
| Address | pin | Cook County Assessment | Age Stratified Random Forrest |
| 2419 W Berenice | 13-24-205-013-0000 | $720,000 | $537,821 |
| 2425 W Berenice | 13-24-205-011-0000 | $720,000 | $452,134 |
| 2429 W Berenice | 13-24-205-010-0000 | $730,000 | $528,915 |
| 2435 W Berenice | 13-24-205-008-0000 | $657,730 | $702,487 |
| 2443 W Berenice | 13-24-205-006-0000 | $660,000 | $472,955 |

Training on homes in the same school area and home age range has significantly reduced the predicated value for most but not all of the homes. We now determine to what extent inaccurate data could be overvaluing homes.

**Questionable Data**

Recall that the graph of sales price vs age of home shows many old homes selling for over a million dollars. We now consider the possibility that this data is incorrect. The Cook County Assessor’s data set includes 56 one-story home sales in the Bell school area. Of these, seventeen of the sales, more than thirty percent of the total, are for a sales price of more than $700,000, with a mean sale price of $890,000. Are these really one-story homes?

Below left is a picture of a home listed in the data set as a one-story home selling for over $1,000,000. On the right is the picture of the home in the Cook County Property Tax portal website, <https://www.cookcountypropertyinfo.com/pinresults.aspx>, for the same address and pin number. The picture on the right appears to be of the bungalow that has been replaced by the mansion on the left.

**A picture containing outdoor, building, house, curb

Description automatically generatedA picture containing building, house, outdoor

Description automatically generated**

2429 W Byron 13-24-204-010-0000 $1,050,000

I have run the models with and without the questionable data. The following table shows the Cook County assessments for several bungalows in my neighborhood that are still bungalows, along with the predications from the Random Forrest model without age stratification with and without the questionable data. Note that we are comparing models without age stratification whereas the previous model had age stratification.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Address | pin | Cook County Assessment | Random Forrest predication using questionable data | Random Forrest Prediction not using questionable data |
| 2419 W Berenice | 13-24-205-013-0000 | $720,000 | 771,642 | 721,817 |
| 2425 W Berenice | 13-24-205-011-0000 | $720,000 | 645,215 | 597,333 |
| 2429 W Berenice | 13-24-205-010-0000 | $730,000 | 736,963 | 695,532 |
| 2435 W Berenice | 13-24-205-008-0000 | $657,730 | 523,533 | 455,545 |
| 2443 W Berenice | 13-24-205-006-0000 | $660,000 | 466,899 | 407,165 |

Removing the questionable data reduced the predicated value in all cases. Overall, age stratifying the data has a greater impact on correcting overvaluation of homes than removing questionable data.

**Conclusion**

I have modeled the market value of bungalows in the rapidly gentrifying Bell School area of Chicago. The Bell School area has a bimodal distribution of home ages containing low-priced bungalows that are over eighty years old and high-priced mansions that are less than twenty years old. The Cook County Assessor’s model for market value of homes predicts bungalow sale values that are much higher than actual sales. By modeling the price of older (newer) homes only on the older (newer) homes, I predict bungalow prices that are generally lower than the Assessor’s values. I have also noticed data rows that appear to have correct mansion sales data but features for the bungalows that they replaced. These incorrect data rows also contribute to the overvaluation of bungalows. As there are many areas of the city that are gentrifying, the data set for all such areas should be checked for errors and age or some other type of stratification should be used to separately model the values of low- and high-priced homes in gentrifying areas.