CAPP 30254 Final Project:
Machine Learning for Public Policy (Spring 2022)

Final Report: Predicting Gentrification in Chicago with Tract-level Neighborhood Characteristics

The Perceptrons
Claire Hemmerly
Dustin Marshall
Phoebe Shihui Collins
Sabrina Yusoff
Yuki Yu Qi Chong

(I) Introduction

Gentrification is a complex, multifaceted and highly contested issue in cities today. Defining gentrification is challenging in itself; the Urban Displacement Project at the University of California, Berkeley defines it as “a process of neighborhood change that includes economic change in a historically disinvested neighborhood — by means of real estate investment and new higher-income residents moving in — as well as demographic change — not only in terms of income level, but also in terms of changes in the education level or racial make-up of residents” (Urban Displacement Project, 2021).

Why is gentrification important?
While there are social and economic advantages to gentrification, the rapid nature of the changes often results in disproportionately detrimental effects on low-income residents in these gentrifying neighborhoods, such as removal of affordable housing, cultural conflicts and the most pernicious problem of all being forced displacement. At the same time, gentrification potentially brings with it the promise of integration and investment that can increase residents’ quality of life — but only if disadvantaged residents are set up to be included in these benefits of increased neighborhood investment.

Therefore, identifying neighborhoods that are at risk or are in the early stages of gentrification using open source census data that reflect neighborhood characteristics is crucial in helping policymakers to implement appropriate intervention policies (e.g. anti-displacement policies) to mitigate the harms that accrue to disadvantaged communities and ensure that these residents stand to benefit from gentrification.

Research question
Our group is particularly interested in urban policy issues and how we can leverage data to improve urban equity — specifically in Chicago, one of the most segregated cities in the United States. Our project sought to explore the research question: **How can we use neighborhood characteristics to predict the neighborhood’s gentrification status in three-years’ time?** To do so, we used 2016 census tract data of neighborhood characteristics to predict a binary classification of a neighborhood’s gentrification status in 2019.

Other issues of interest such as housing affordability, neighborhood investment and revitalization, and racial integration are tightly linked to gentrification, thus we hope that our project will be able to contribute to existing research to inform policy that works towards ethically integrating urban neighborhoods and fostering sustainable and inclusive community development. It took overcoming several conceptual hurdles in order to answer the question in a truthful and policy-relevant way, the details of which are explained in the method and experiment setup section.

(II) Dataset

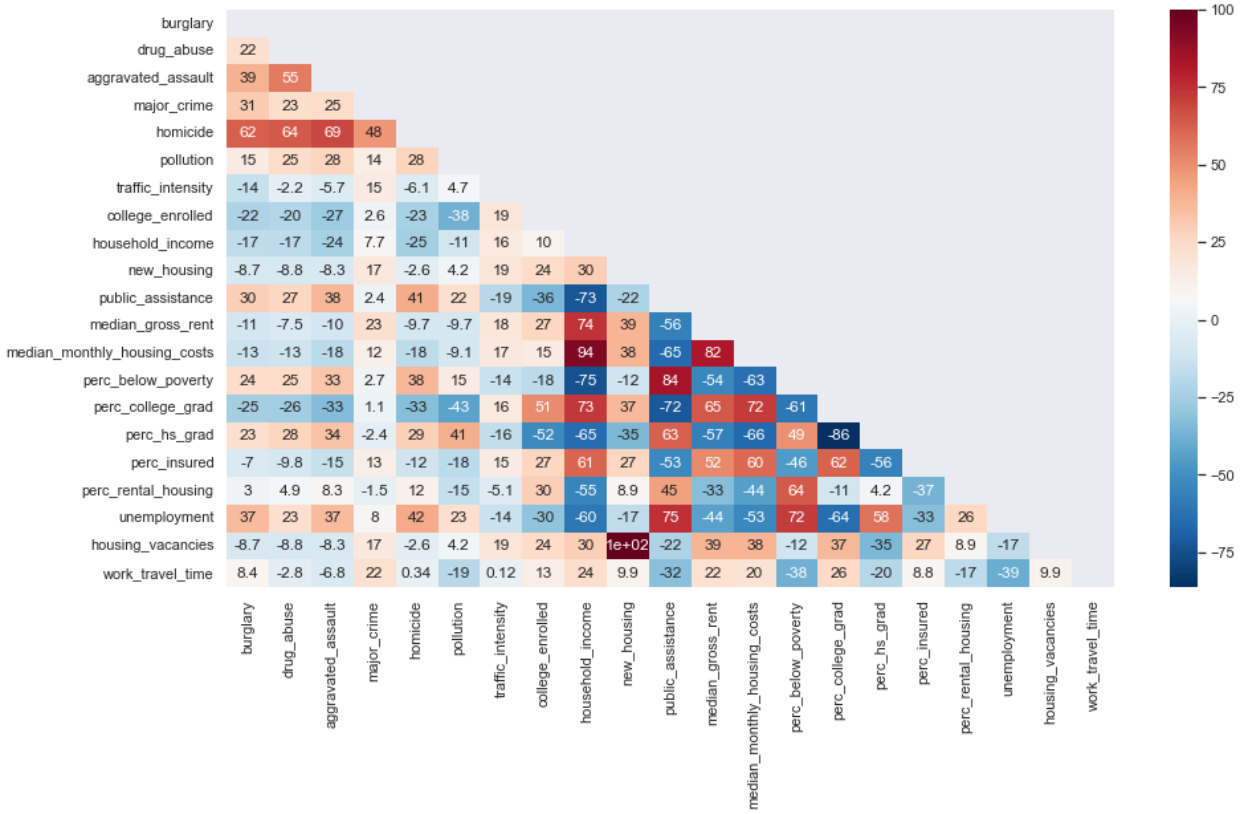
Feature Variables

Our final dataset has 778 observations with 22 feature variables indicating socioeconomic status, educational attainment, criminal activity, pollution level, transportation equity and housing equity in the year 2016, with the unit of observation being census tracts within the City of Chicago. The data was collected by American Community Survey and Chicago Health Atlas (see Appendix for more details on feature variables).

Descriptive Statistics

Before building our model, we visualized feature values using histograms to determine feature engineering. As feature values varied significantly in scale, we normalized features to improve model performance and interpretability. We did not observe outliers that needed to be removed from the data. While most features generally have an even distribution, some features such as assault, drug abuse and traffic intensity were significantly skewed with high density close to zero. This indicated that some features might need to be pruned if the variance is too low.

To further evaluate feature relevance and redundancy, we created a correlation heatmap where an increasing intensity of red represents increasing positive correlation and an increasing intensity of blue represents increasing negative correlation.



The correlation heatmap indicated possible multicollinearity between features, for instance between crime-related features (homicide, burglary, drug abuse, and aggravated assault). We explored collapsing features with low variance or were highly correlated with other features into a composite feature, however this caused model accuracy to deteriorate. Ultimately, we decided to choose a model that would be able to handle redundant features, such as random forest.

Y labels

To create a binary classification of a neighborhood's gentrification status in 2019, we used median home values between 2016 to 2019 at the census-tract level from the American Community Survey. We describe generating the y labels in the next section on Experiment Setup. The dataset comprises 175 positive labels and 603 negative labels.

There was a heavy skew towards 0 in our y labels, with only 175 out of 778 (22%) census tracts assigned a positive label. An imbalanced dataset makes it more challenging to model as the model might overfit to the over-represented class. However, the classification of gentrifying neighborhoods is inherently imbalanced as only a small proportion of neighborhoods undergo gentrification at any one point in time while most neighborhoods remain relatively stable. This skewed distribution would subsequently inform our choice of evaluation metrics in the model selection and results sections of our project.

(III) Experiment Setup

Balancing classification model's accuracy and policy relevance

The first hurdle for our project was to build a model that could accurately predict gentrification using neighborhood characteristics while also being relevant for allowing timely policy interventions. Our first model which nowcasted neighborhoods' median home values using neighborhood characteristics from the same year had a decent accuracy score. However, our second model forecasted median home values using neighborhood characteristics from previous years performed very poorly. Recognizing that forecasting was key to allow timely policy interventions, our third model retained the forecasting task but simplified the y variable into a binary classification label (gentrifying or not gentrifying) instead of a continuous variable. This allowed us to balance model accuracy with policy relevance.

Using median home value as proxy for gentrification

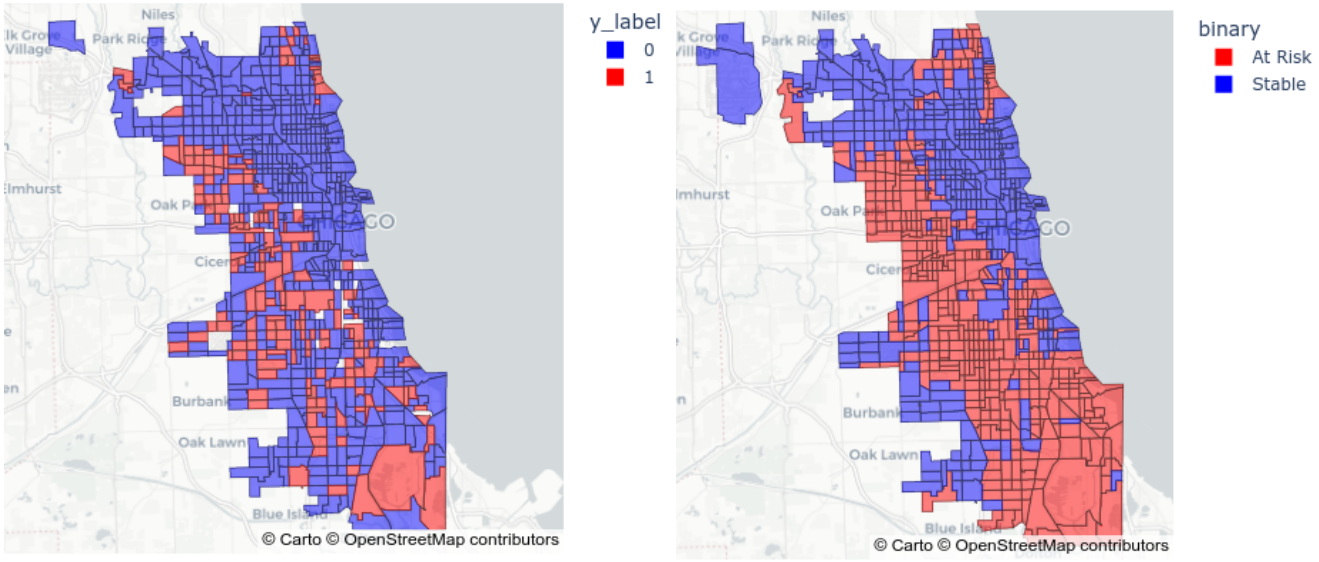
The next hurdle was to determine a variable that could serve as a proxy for gentrification, which is characterized by physical, demographic and socio-economic changes in neighborhoods. Since gentrification is a complex process that is hard to define, any ground truth estimation of it is a composite of many neighborhood characteristics.

Numerous papers on gentrification use home value as a measure of gentrification. Turner and Snow (2001) from the Urban Institute defined gentrification as "sales prices that are above the DC average". Bates (2013) developed an approach to classify census tracts into "early gentrified", "dynamic gentrified", and "late gentrified" using housing appreciation rates and housing values. Since our literature review pointed to median home value as a key indicator of gentrification, we chose to focus on it as the key variable for generating class labels.

Generating class labels using median home value

Next, we determined the conditions for generating class labels. One key aspect of gentrification is an increase in median home value. However, we realized that this alone could not sufficiently encapsulate the complexities of gentrification. In nowcasting gentrification with socioeconomic and Airbnb data, Jain (2021) focused solely on disadvantaged neighborhoods for two reasons. First, when median home prices rise, non-disadvantaged neighborhoods experience significantly less change. Second, and more importantly, the concept of gentrification is usually discussed only in the context of disadvantaged neighborhoods; while affluent neighborhoods may experience some socioeconomic changes due to rising home prices, this is typically not considered to be gentrification (Zuk et al., 2018). As such, our group decided on a more nuanced classification approach involving two conditions: If a neighborhood's median home value was 1) below the median of all median home values in Chicago in 2016 and 2) increased from 2016 to 2019 at a rate greater than the median rate of change in Chicago from 2016 to 2019, we classified that tract as "gentrifying". We assigned neighborhoods the label "not gentrifying" if they did not meet either of these two conditions.

To validate our approach, we compared our generated class labels with an established gentrification typology index from the Urban Displacement Project. Since our project involves a binary classification, we collapsed the ten classes used by the Urban Displacement Project into two broader categories: At Risk (comprising ‘Low-Incomes/Susceptible to Displacement’, ‘Ongoing Displacement’, ‘At Risk of Gentrification’, and ‘Early/Ongoing Gentrification’) and Stable (comprising the remaining classes). In the plot below, our label distribution is on the left and Urban Displacement Project is on the right. Overall, our labels generally align with the typology index, lending validity to our approach for generating class labels.



(IV) Method

For all models, we used an 80/20 stratified train-test split that preserves the same proportions of examples in each class as observed in the original dataset to account for the skewed distribution of class labels. To develop our model, we referenced an article by Wagle (2020) who used machine learning to predict house prices using house characteristics at the individual-house level.

For each model, we:

1) Fitted the model on the training data
2) Obtained predicted values from the fitted model
3) Calculated accuracy scores as the proportion of correctly classified observations
4) Applied 10-fold cross-validation and obtained the mean accuracy and standard deviation - we used cross-validation since our training set is too small to have a separate validation set

Model Selection

Given that our research question involved class prediction, supervised learning methods were considered to answer our question. We used logistic regression as our baseline model, which we compared to three other models:

i) decision tree: we included a decision tree as it is a simple and interpretable model
ii) random forest: we included this ensemble model for improved accuracy
iii) gradient-boosted decision tree: we included this ensemble model for improved accuracy

To evaluate the performance of the aforementioned binary classification models, we used the evaluation metrics: accuracy, precision, recall and F1 score. Precision represents the fraction of positive predictions that belong to the positive class, recall represents the fraction of positive predictions out of all positive instances in the dataset and F1 score is the harmonic mean between precision and recall. We use these metrics because our data is imbalanced, so it may be skewed by a large number of negative examples.

Based on the performance of the four classification models using the metrics accuracy, precision, recall and F1 scores (Table 1), we selected the gradient-boosted decision trees model as our final model. The model had the best overall performance, with the highest baseline accuracy, precision and F1 score - especially since the F1 score strikes a balance between precision and recall, this highlights the model's relatively better performance overall.

Baseline: 0.22 (% positive instances)

Table 1: Model Metrics

	Metrics			
Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Logistic Regression	0.76	0.41	0.10	0.15
Decision Trees	0.69	0.30	**0.28**	0.28
Random Forest	0.76	0.38	0.11	0.16
Gradient-Boosted Decision Trees	**0.77**	**0.50**	0.21	**0.29**

Hyperparameter Tuning

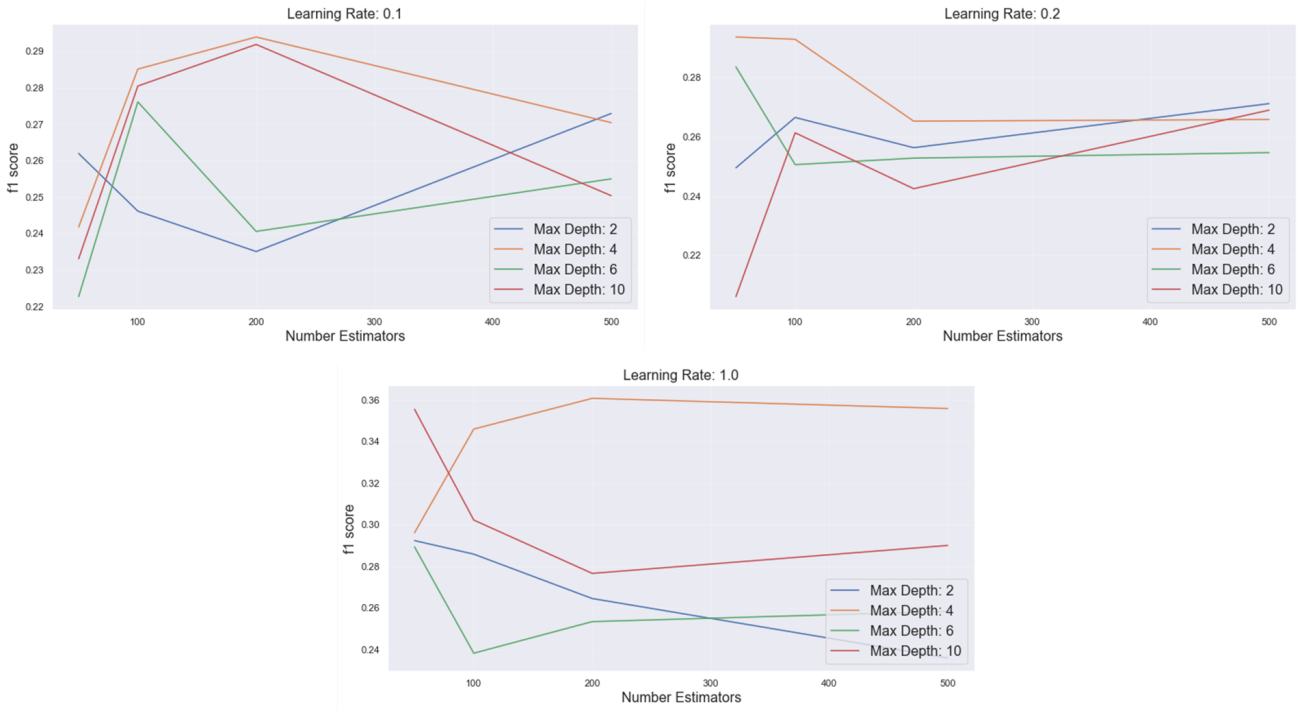
We then used grid search to tune the hyperparameters of the final gradient boosted model. We created a grid of hyperparameter values to try:

- n_estimators (no. of sequential trees to build): 50, 100, 200, 500
- max_depth (maximum depth of each tree): 2, 4, 6, 10
- learning_rate (amount of contribution each model has on the ensemble prediction): 0.1, 0.2, 1.0
- subsample (no. of examples used to fit each tree): 0.5, 0.7, 1.0

A total of 4 x 4 x 3 x 3 = 144 models were generated using a combination of values on these four hyperparameters, for which the accuracies were computed and compared to produce the best model with highest accuracy on 5 fold cross-validation.

Grid Search Results

As plotting combinations of values on all four hyperparameters is complex, we focus on examining the number of estimators, max depth and learning rate. The above plots summarize the F1 scores on all combinations considered for these 3 parameters, with a subsample value of 0.5. Each individual plot shows one of the 3 learning rate values tested. Within each plot, each line represents the max depth of the tree tested with the number of estimators parameter as the x-axis and F1 score as the y-axis.

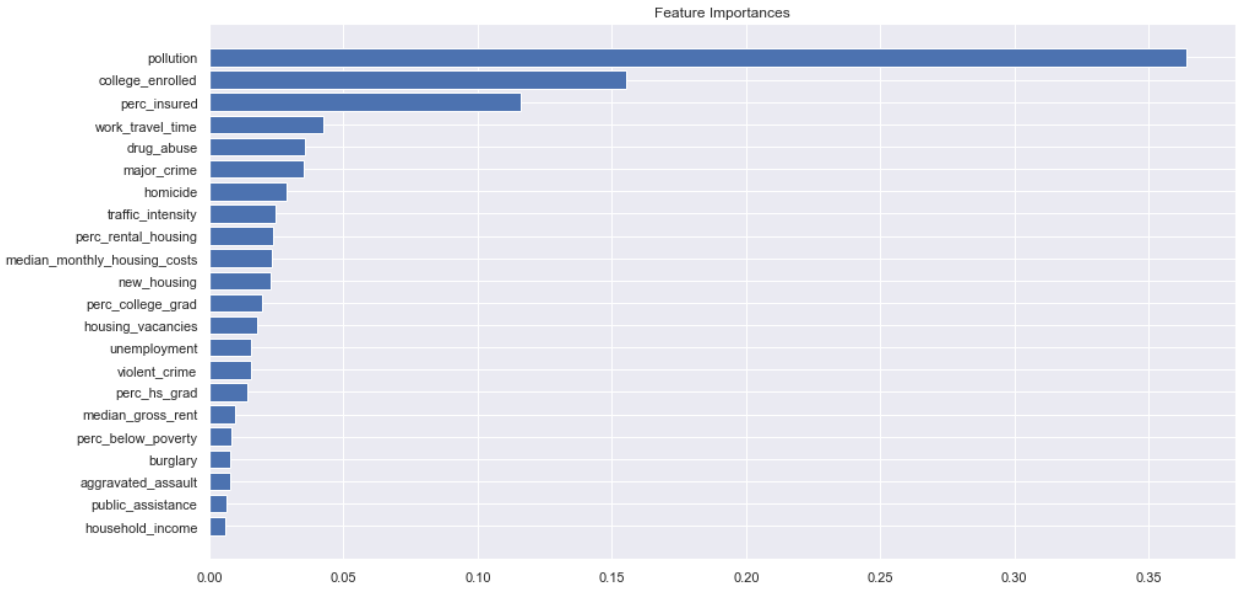


The hyperparameter values of the final model that produced the highest f1 score are:

- n_estimators: 200
- max_depth: 4
- Learning_rate: 1.0
- Subsample: 0.5

Feature Importance

We plotted the relative importance of all features in the final model. The top three features were pollution, percent college enrolled and percent insured. While education level has been previously cited as a relevant factor in gentrification, the high relevance of pollution and percent insured for predicting gentrifying neighborhoods was surprising to us, suggesting areas to be explored in future research.



(V) Results Analysis

Using the test data, the performance of our final model with tuned hyperparameters on the four evaluation metrics are as follows:

Model	Accuracy	Precision	Recall	F1 Score
Gradient Boosted Decision Trees	0.692	0.324	0.343	0.333

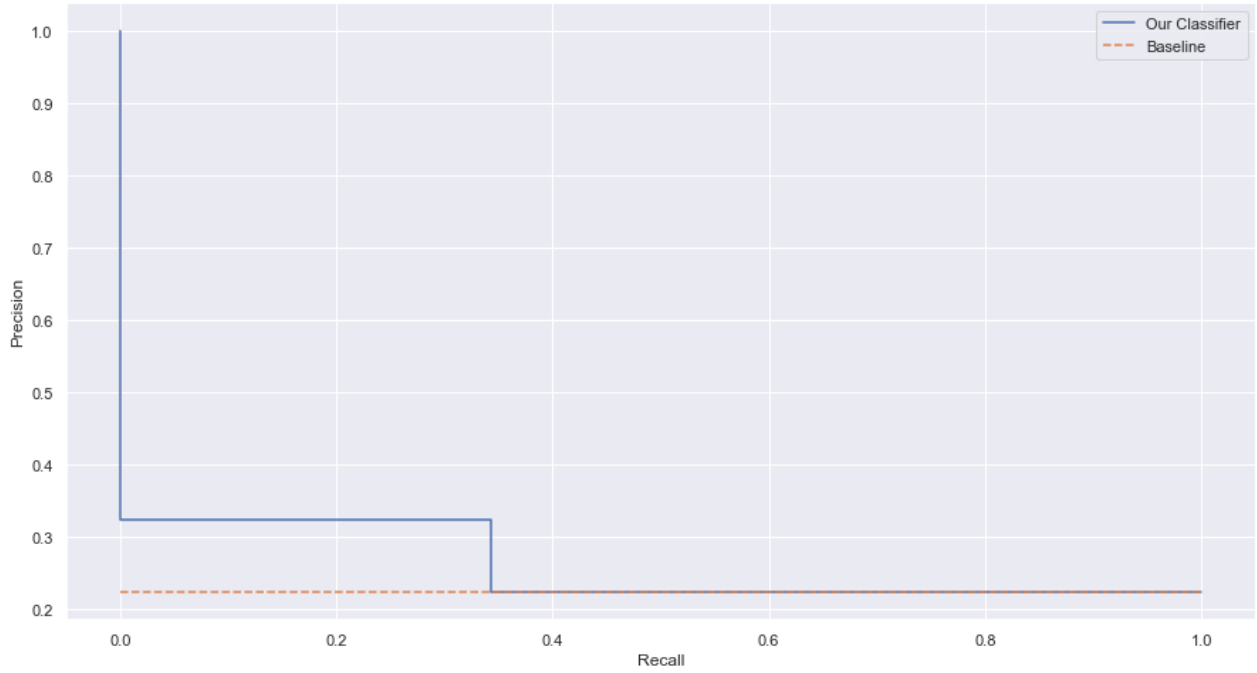
Precision-Recall Curve

Given the imbalance in binary class labels, we plotted a prediction-recall curve, which is often used when class distribution is heavily skewed as in our dataset. This is more informative than the ROC which would paint an overly-optimistic picture of model performance and overestimate the predictive power of our model. In addition, precision-recall analysis is usually conducted to evaluate binary classification models which are primarily interested in the positive instances. In our model which is mainly concerned with the accurate classification of gentrifying tracts, this further justifies the use of this evaluation technique.

Based on our precision-recall curve, we observed fairly low values for both precision and recall, where we performed worse than a random classifier with an AUC score of 0.5. Possible reasons for the lackluster performance of our model include:

1) Feature variables used lacked predictive power - we did not include demographic features like age and race
2) Skewed distribution of class labels which likely required more optimization during the data preparation process
3) Gentrification itself is a complex phenomenon, thus using home prices alone as a proxy for gentrification might not have sufficed. The feature variables used in our model were

meant to be relevant to gentrification, and this did not necessarily translate to high relevance to home appreciation rates



(VI) Conclusion & Key Lessons

Summary

Our project has achieved its main objective of understanding how neighborhood characteristics, including socioeconomic, crime and housing data on a tract level, can be used to predict the neighborhood's gentrification status in three-years' time, with housing appreciation as a proxy for gentrification. We explored various supervised machine learning algorithms that were suitable for binary classification, including logistic regression, decision trees, random forest and gradient-boosted decision trees before deciding on our final gradient boosted decision trees model with grid search due to its relatively better performance in the evaluation metrics accuracy, precision and F1 score as compared to the other three models. However, the overall performance of our final model was less-than-ideal, as reflected by the low AUC in the precision-recall curve plotted.

Next steps

Given more time and resources, we would consider the following to improve our model:

- **Expand our dataset** to include census tracts across Cook County
- **Consider a composite class label** such as one that includes other important indicators of gentrification like household income and educational attainment in addition to home prices.
- **Diversify our data sources** to explore categorical data like building permits and unstructured data such as Twitter data and Airbnb reviews, since they may provide more

granular insights on human behavior in changing neighborhoods thereby improving the overall robustness and performance of our model.
- **Optimizing skewed data with resampling methods** such as over-sampling examples from the minority class by duplicating random examples from this class, or creating synthetic data.
- **Perform a multiclass classification** to provide a more nuanced and comprehensive view of tracts at different stages of gentrification and neighborhood change.

Key takeaways

- Overall, working on this project has provided us with greater insight into what it means to **develop a feasible and meaningful research question** - exploring a complex phenomenon like gentrification has given us the opportunity to integrate existing literature and our own conceptual interpretation of gentrification into our model design.
- Secondly, our experience has certainly verified the saying: “A machine learning model is only as good as the data it is being trained on”. Our initial struggle with data collection and wrangling, which resulted in our baseline approach performing poorly, has highlighted how our dataset forms the bedrock for all the downstream decisions - thus, it is essential to **align our research question with available data** we have access to and set aside time for data preparation.
- Working on this project underscored the importance of **using appropriate evaluative techniques based on use cases** to ensure that we do not overestimate the predictive power of our model - we started off using the ROC curve to evaluate our model, only to realize we were painting an overly optimistic picture of our model’s predictive power. With the skewed distribution of our class labels, we changed our approach to use the precision-recall curve to more accurately evaluate our model performance.
- Above all, this project has been a humbling reminder for us to constantly **keep in mind the policy impetus when making different algorithmic decisions** in the model creation process. While high model accuracy should be prioritized, we should only be working towards that after ensuring that our model is answering a meaningful question and is truly adding value to the problem we are trying to address.

Work Division

- Claire Hemmerly - Data visualization, analysis, and write-up
- Dustin Marshall - Data collection, wrangling, and write-up
- Phoebe Shihui Collins - Feature engineering, model selection, data analysis, and write-up
- Sabrina Yusoff - Feature engineering, model selection, data analysis, and write-up
- Yuki Yu Qi Chong - Data collection, wrangling, and write-up

References

Bates, L. (2013). Gentrification and Displacement Study: Implementing an Equitable Inclusive Development Strategy in the Context of Gentrification. *Urban Studies And Planning Faculty Publications And Presentations*. https://doi.org/10.15760/report-01

Chapple, K., & Thomas, T., and Zuk, M. (2021). Urban Displacement Project website. Berkeley, CA: Urban Displacement Project. https://www.urbandisplacement.org/about/what-are-gentrification-and-displacement/

Jain, S., Proserpio, D., Quattrone, G., & Quercia, D. (2021). Nowcasting Gentrification Using Airbnb Data. *Proceedings Of The ACM On Human-Computer Interaction*, *5*(CSCW1), 1-21. https://doi.org/10.1145/3449112

Knorr, D. (2019). *Using Machine Learning to Identify and Predict Gentrification in Nashville, Tennessee* (Masters). Vanderbilt University.

Turner, M., and Snow, C. (2001). "Leading Indicators of Gentrification in D.C. Neighborhoods." Presentation at the Urban Institute D.C. Policy Forum, Washington, D.C., June 14, 2001. Washington, DC: Urban Institute.

Wagle, M. (2020). *Predicting House Prices using Machine Learning*. Medium. Retrieved 1 June 2022, from https://medium.com/@manilwagle/predicting-house-prices-using-machine-learning-cab0b82cd3f.

Zuk, M., Bierbaum, A. H., Chapple, K., Gorska, K., & Loukaitou-Sideris, A. (2018). Gentrification, Displacement, and the Role of Public Investment. Journal of Planning Literature, 33(1), 31–44. https://doi.org/10.1177/0885412217716439

Appendix

American Community Survey

No.	Predicted variable (Y)	Column name
1	Median home value ($)	home_value

No.	Feature variable (X)	Column name
1	Aggregate travel time to work (mins)	work_travel_time
2	Median gross rent ($)	median_gross_rent
3	Median household income ($)	household_income
4	Median monthly housing costs ($)	median_monthly_housing_costs
5	Percent below poverty level (%)	perc_below_poverty
6	Percent college graduate (%)	perc_college_grad
7	Percent high school graduate (%)	perc_hs_grad
8	Percent housing built after 2000 (%)	new_housing
9	Percent housing units vacant (%)	housing_vacancies
10	Percent insured (%)	perc_insured
11	Percent living in household w/ public assistance income and/or food stamps (%)	public_assistance
12	Percent population enrolled in college (%)	college_enrolled
13	Percent rental housing (%)	perc_rental_housing
14	Unemployment rate (%)	unemployment

Chicago Health Atlas

No.	Feature name (X)	Column name
15	Aggravated assault/battery	aggravated_assault
16	Burglary	burglary
17	Drug abuse	drug_abuse
18	Homicide	homicide
19	Major crime	major_crime
20	Particulate matter (PM 2.5) concentration (µg/m3)	pollution
21	Traffic intensity	traffic_intensity
22	Violent crime	violent_crime