**Data Sources**

For this Analyis, we utilized several data sources to gather information on various aspects of Chicago's neighborhoods. The data sources used were:

* Census Data: We obtained census data from the American Community Survey (ACS) 5-year estimates for the years 2013-2018. The data was downloaded using the census data library in Python. The variables included in the dataset were population, median age, median household income, total housing units, median number of rooms, median year built, median gross rent, and mean household size.
* Crime Data: Crime data was obtained from the City of Chicago's Data Portal. The data included crime incidents from 2013 to 2018, with variables such as case number, year, primary type, and location.
* CTA Data: CTA (Chicago Transit Authority) data was obtained from the City of Chicago's Data Portal. The data included information on CTA stations, including their location and distance from each neighborhood.
* Sales Data: Sales data was obtained from the Cook County Assessor's Office. The data included information on property sales, including sale price, age, and location.

**Data Cleaning and Preprocessing**

Before analyzing the data, we performed several preprocessing steps to ensure the data was clean and consistent. The preprocessing steps included:

* Merging datasets: We merged the census, crime, CTA, and sales datasets based on the GEO\_ID variable, which represents the unique identifier for each neighborhood.
* Handling missing values: We handled missing values in the datasets by dropping rows with missing values or imputing them with median values.
* Data transformation: We transformed some variables, such as the sale price and age, to ensure they were in a suitable format for analysis.
* Data reshaping: We reshaped the data from long format to wide format to facilitate analysis.
* Data cleaning: We cleaned the data by removing duplicates and ensuring that the data was consistent across all datasets.

After preprocessing the data, we got a combined datasets with 400k rows and 32 columns, To make m analysis easier I divided this into train and predicting dataset and now Both these datasets are ready for analysis and visualization.

**Exploratory Data Analysis**

A map of a city

Description automatically generatedThe visualization depicted on the left highlights distinct variations in crime severity across different areas of Chicago. Specifically, it reveals that the South Side, Downtown, and West Side exhibit noticeably higher levels of severe crimes compared to the North Side. This spatial disparity underscores significant differences in crime dynamics throughout the city, reflecting varying socioeconomic factors, population density, and possibly differing levels of law enforcement presence and community resources. Understanding these geographical patterns is crucial for local policymakers, law enforcement agencies, and community organizations aiming to implement targeted interventions and resource allocations to address crime prevention and community safety effectively.

A graph showing the number of cities in the united states

Description automatically generated

The figure above illustrates that neighborhoods on the North side of Chicago generally exhibit higher sale prices. This trend can be attributed to several factors, including lower crime rates and better access to transportation infrastructure, particularly robust transit networks. These areas benefit from a combination of factors that enhance desirability and livability, such as safer environments and convenient commuting options. This spatial disparity underscores how urban amenities and safety considerations significantly influence real estate values, highlighting the importance of location-specific factors in shaping property markets.

A graph showing the average price of a neighborhood

Description automatically generated

From this plot, it's evident that crime rates are notably higher on the south side of Chicago, contributing to lower average housing prices in those areas. Moving northward, we observe a gradual decrease in crime rates, corresponding with an increase in average housing prices. However, the downtown Loop area stands out as an exception: despite higher crime rates compared to some northern neighborhoods, it maintains high average housing prices. This anomaly suggests that other factors, such as urban amenities, job opportunities, and historical significance, play significant roles in determining property values in the downtown core.

**Model Development**

The primary objective of our analysis is to understand how the introduction of a new CTA (Chicago Transit Authority) line in a neighborhood affects property pricing and crime rates in that area. This requires a comprehensive examination of various factors influencing both property values and crime incidents, necessitating the use of advanced machine learning techniques to predict these outcomes accurately.

Our problem can be framed as a regression task where the goal is to predict the average house price per Census block group. This involves examining the relationships between a diverse set of features derived from demographic data, crime statistics, transportation accessibility, public school availability, housing sales attributes, and neighborhood characteristics. Specifically, we aim to answer the following questions:

1. How does proximity to a new CTA line impact property prices in Chicago's neighborhoods?

2. What is the effect of a new CTA line on crime rates in these neighborhoods?

1. Model Selection

To address these questions, we employed a variety of regression models, conducting a grid search to identify the best-performing algorithms. The models used include:

Linear Regression, Ridge Regression, Lasso Regression, Elastic Net, Decision Tree Regressor, Random Forest Regressor, Gradient Boosting Regressor

These models were chosen due to their suitability for predicting quantitative outcomes with high levels of accuracy, which we measured using Root Mean Squared Error (RMSE).

**Feature Selection**

Our feature set includes a comprehensive array of variables to capture the multifaceted influences on property prices and crime rates:

1. Demographic Data: Total Population, Median Age, Median Household Income, Total Housing 6 Units, Median Number of Rooms, Median Year Built, Median Gross Rent, Mean Household Size, Percent White Population, Percent Black Population, Percent Household with Children, and Percent Housing Vacant.

2. Crime Data: Total Number of Crimes (all types of crime) and Total Number of Crimes per Capita (Total Number of Crimes / Total Population from ACS).

3. CTA Data: Distance to the nearest L train station (distance is calculated in miles). All distance data is calculated for the Census block group level from the original data set.

4. Sales & Housing Attributes Data: Number of Sales, Average Age of House Sold, Price per Square Feet, Price per House, and Average Square Feet of House Sold. Sales data were calculated for the Census block group level and were 3 years lagged. After filtering the original dataset, we focused on approximately 220,000 housing sales within the City of Chicago.

5. Neighborhood and Side Indicators: Dummy variables indicating which neighborhood (i.e., Hyde Park, Woodland, etc.) and which side (i.e., North, West, etc.) the Census block is in.

**Feature Engineering and Correlation Analysis**

During the analysis, we examined the correlations and distributions of all our features. No additional feature transformations were found necessary beyond those initially included. We ensured that our features were relevant and adequately captured the variability needed for accurate prediction of property prices and crime rates. By integrating these diverse data sources and employing a rigorous model selection process, we aimed to create robust predictive models to inform urban planning and policy decisions regarding the impact of new CTA lines on Chicago neighborhoods.

Temporal Data Splitting

Given our objective to use past features data (i.e., 3 years lagged) to predict future average house prices per Census block group, we employed a temporal splitting strategy for our train, validation, and test sets. This approach ensures that the models are trained and evaluated in a manner that mimics real-world forecasting scenarios. The splitting was done as follows:

A chart with text overlay

Description automatically generated

The years mentioned above indicate the target variable years and not the feature years. For example, the "2016 Train" set includes average house prices per Census block group for 2016 and features data from 2013. This lagging of feature data is crucial because the oldest available data for all features is from 2013, making 2016 the earliest year for which we can train our models to predict.

Due to this limitation, we have two validation sets and one test set. To further validate our models and ensure they do not overfit, we performed non-temporal 5-fold cross-validation. This involved splitting each training set (before 2017 and before 2018) into five random folds and reporting the average test error (measured as RMSE) of the validation fold. This cross-validation process helps corroborate the robustness of our final model.

Models and Hyperparameters Grid Search

We initiated our modeling approach with a simple Linear Regression to establish a foundational baseline, which inherently does not require tuning hyperparameters. Subsequently, our strategy involved delving into more sophisticated models where we systematically explored various sets of hyperparameters to optimize their predictive performance. These models included Ridge Regression, Lasso Regression, Elastic Net, Decision Tree Regressor, Random Forest Regressor, and Gradient Boosting Regressor. The objective of our grid search was to methodically evaluate different combinations of hyperparameters across these models to identify the optimal settings that would maximize predictive accuracy and enhance the generalizability of our results.

Our grid search encompassed essential hyperparameters such as regularization strength for Ridge and Lasso, the balance between L1 and L2 regularization for Elastic Net, maximum tree depth and features for Decision Tree and Random Forest, and boosting parameters like number of estimators, learning rate, and subsample size for Gradient Boosting. Each model underwent rigorous evaluation to determine which hyperparameter settings yielded the most promising outcomes in predicting the impact of new CTA lines on property prices and crime rates within Chicago neighborhoods.

For further insights into our grid search results and the final model selection process, detailed documentation is available in our Git repository under the "train-models-final" notebook. Our comprehensive approach, coupled with a robust feature set, aims to uncover actionable insights that illuminate the complex interplay between transportation infrastructure developments and urban dynamics, particularly focusing on housing markets and crime trends in Chicago.

**Evaluation Metric and Model Selection**

For evaluating the performance of our regression models, we opted for Root Mean Squared Error (RMSE) as our primary metric. RMSE is widely used in regression tasks because it provides a clear measure of prediction accuracy that aligns with the scale of the target variable. This choice allows for straightforward interpretation of how well our models predict average house prices per Census block group.

Given the limited number of temporal validate sets (2017 and 2018), we acknowledged the potential challenge of assessing model generalizability. Therefore, in addition to evaluating RMSE on these temporal sets, we also conducted non-temporal 5-fold cross-validation. This approach helped us validate the robustness of our models across different splits of the data.

**Model Performance:**

A screenshot of a graph

Description automatically generated

As illustrated in the charts, our initial exploration revealed that the simple Linear Regression model performed inadequately, suggesting the need for more sophisticated techniques. Among the models evaluated, the Random Forest Regressor consistently demonstrated superior performance based on RMSE. Consequently, we selected it as our final model class for its robust predictive capabilities.

Through a rigorous grid search process, we identified optimal hyperparameter configurations for the Random Forest Regressor. The top-performing sets of hyperparameters were determined based on their average rank of temporal validate RMSEs across 2017 and 2018 datasets. This approach ensured that our chosen configurations were consistently effective across different temporal splits.

The table below summarizes the top three sets of hyperparameters selected based on their average temporal validate RMSE ranks:

A screenshot of a computer

Description automatically generated

These configurations not only performed well in temporal validation but also showed promising results in non-temporal 5-fold cross-validation, with RMSEs within a negligible margin of the best obtained results. This consistency across different validation methods provided confidence in the robustness of our chosen hyperparameters.

**Final Model Selection**

Considering concerns about potential overfitting, we opted for the set of hyperparameters that offered a balance of complexity and performance: {'max\_depth': 40, 'max\_features': 'sqrt', 'min\_samples\_split': 2, 'n\_estimators': 100}. This decision was reinforced by observing that these parameters achieved competitive RMSEs on the temporal test set (2019 data) and maintained favorable rankings across various evaluation metrics.

**Result and Findings**

A table with numbers and letters

Description automatically generated

A graph of a number of people

Description automatically generated with medium confidence

Based on the feature importance scores obtained from our model, it is evident that certain variables significantly influence the average house prices per Census block group in Chicago neighborhoods. Notably, variables such as 'prev\_year\_price\_p\_house', 'prev\_year\_price\_p\_sf', 'Median HH Income', 'Percent White Population', and 'Median Gross Rent' emerged as key determinants. These features underscore the socioeconomic dynamics and housing market characteristics that play pivotal roles in shaping property values across different neighborhoods.

Moreover, the inclusion of 'Norm distance\_miles' among the influential factors suggests that proximity to CTA lines indeed impacts property pricing and crime rates. As anticipated, our analysis revealed that areas closer to new CTA lines experience observable changes in both housing prices and localized crime rates. This finding aligns with our initial hypothesis, indicating that transportation infrastructure developments can have significant socio-economic implications on urban neighborhoods.