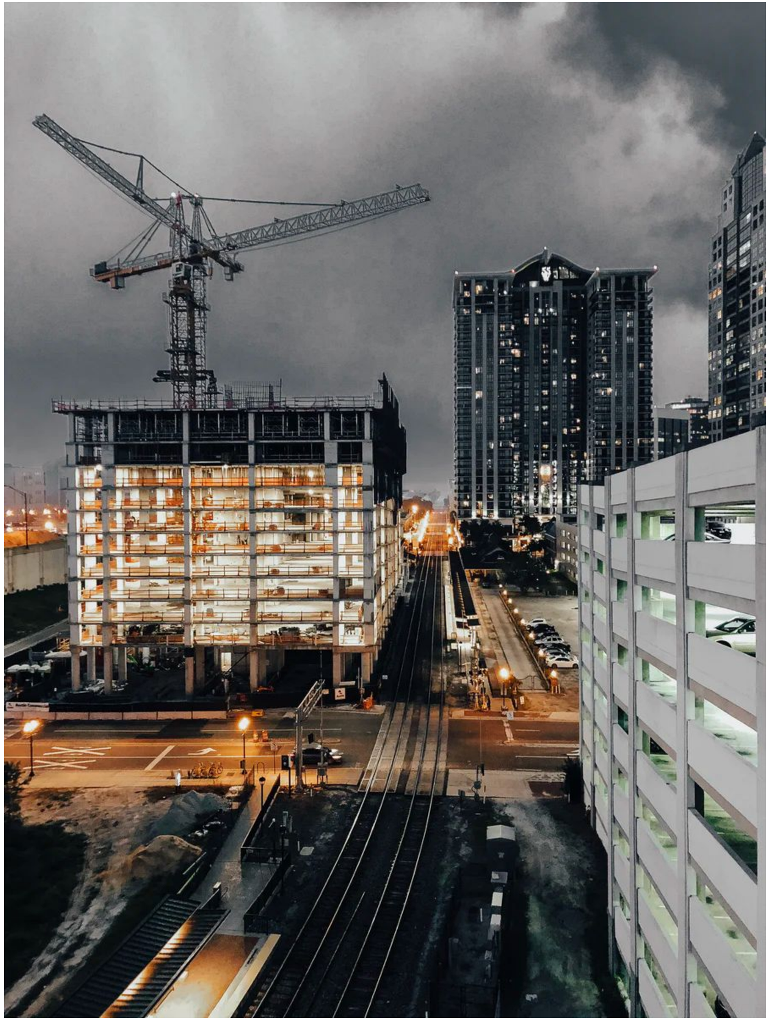
Finding the next trendy neighborhood in chicago

Problem Statement

- Old and neglected neighborhoods can quickly transform into a popular and trendy neighborhood

- This cause high real estate demand and sharp increase in the real estate prices.

- If detected early, these areas can be a good real estate investment opportunity.



Goal

- Build a model that identifies next trendy zip codes

- The model predicts the change in housing price over the next three years using historical housing prices and other factors that can affect the housing price in a neighborhood.

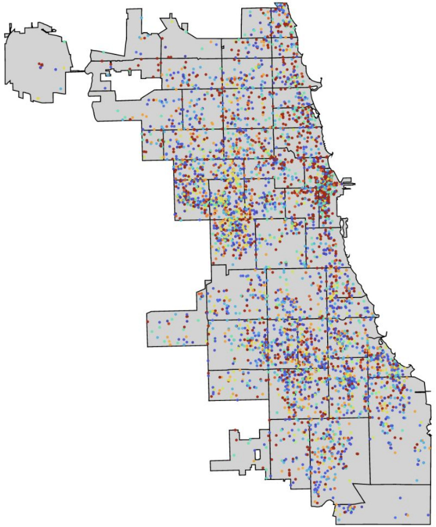
Some limitations:

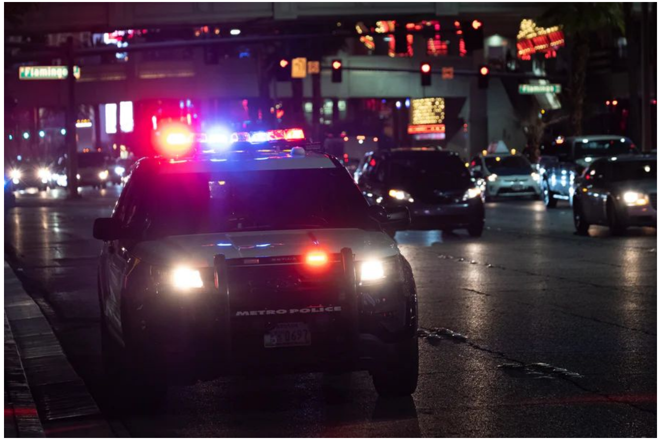
- Data has to be publicly available

- Data has to include location information (zip code, latitude, longitude)

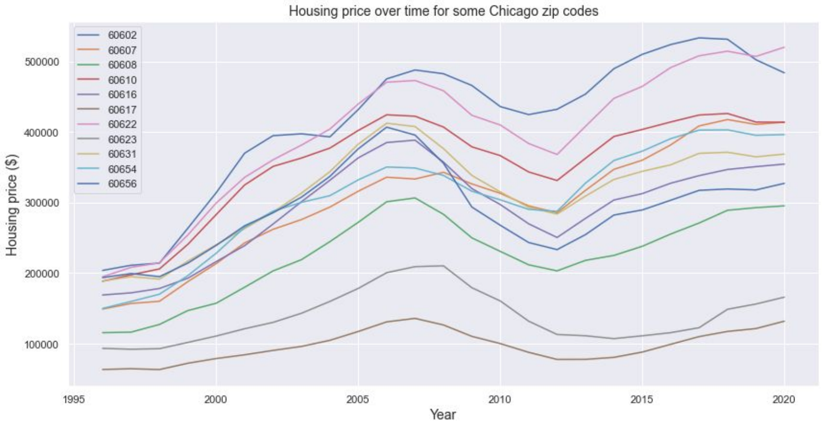
Data Sources

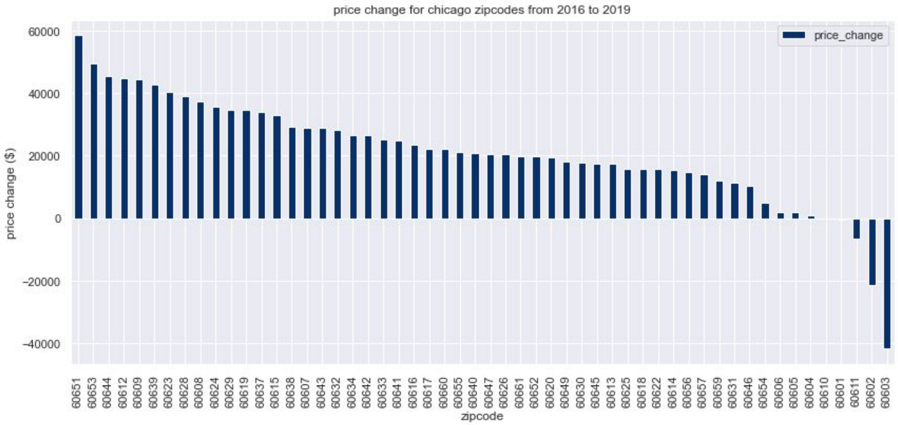
- building permits
- valid retail food licenses
- crime rate
- historical housing prices
- Zip code information





Housing Price Over Time

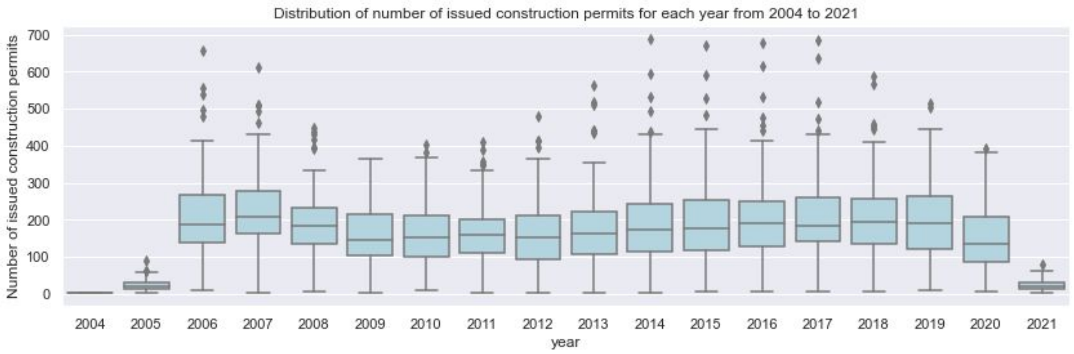


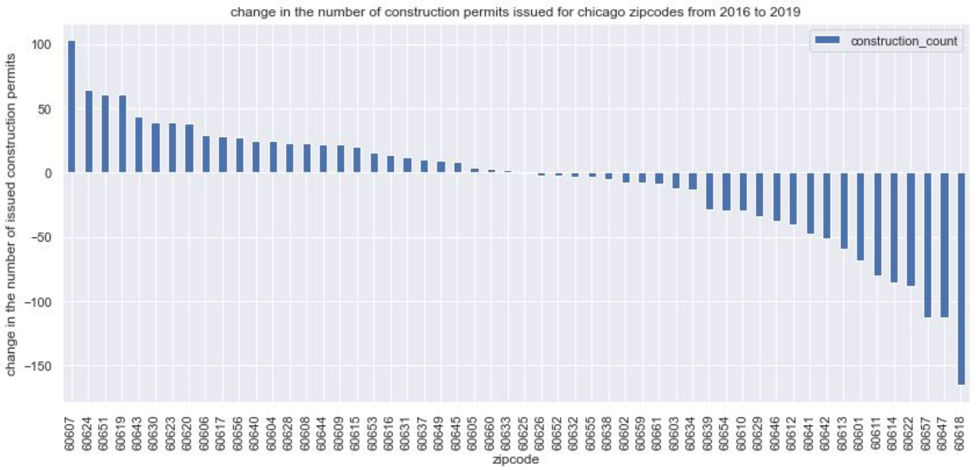


Construction and Renovation Permits Data

- In each year there are few zip codes that are outliers and have much more construction counts compared to other zip codes.

- The median construction count dropped during the housing market crises but slowly increased afterwards.

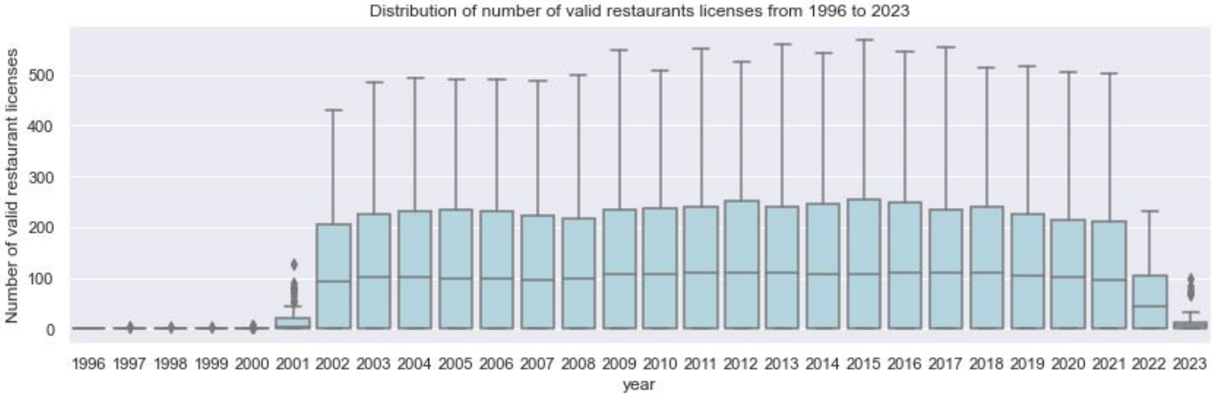


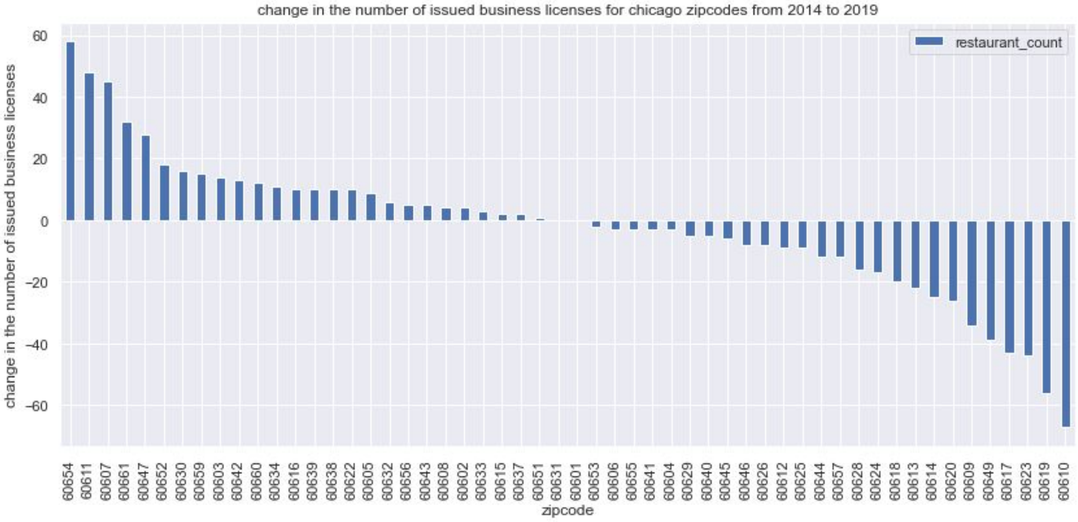


Issued Restaurants license Data

- Yearly distribution of valid business licenses is almost uniform over time.

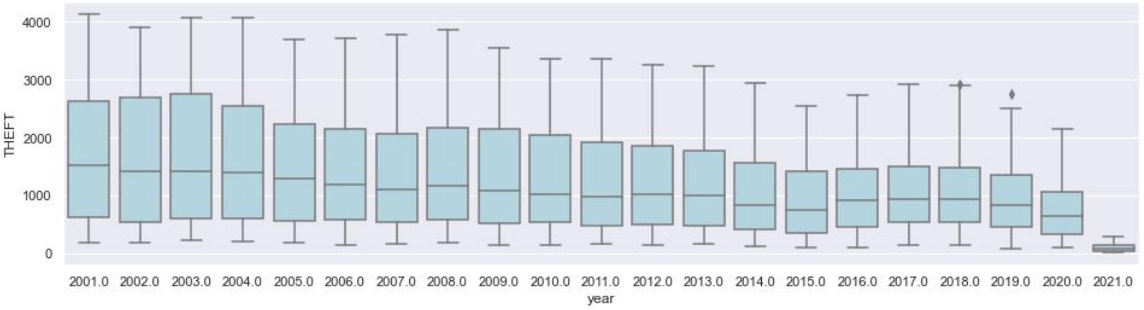
- Number of businesses in some zip codes such as 60654 60611 60607 increased from 2014 to 2019 however, some zip codes such as 60610, 60619, and 60623 lost some businesses.

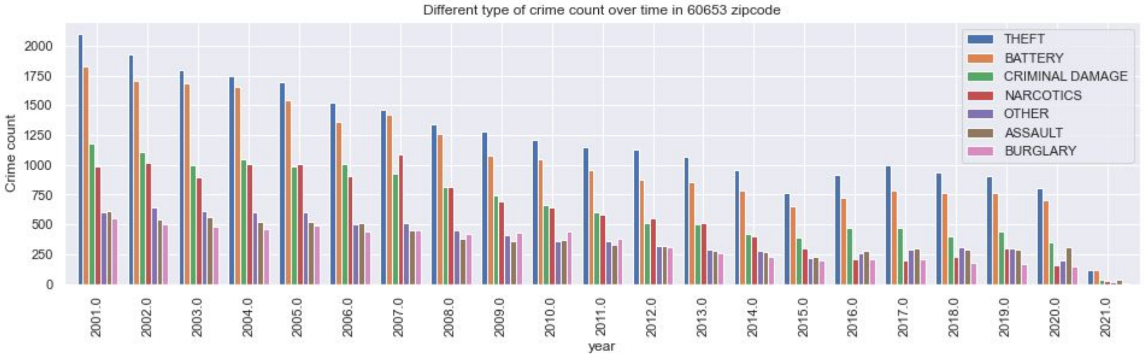


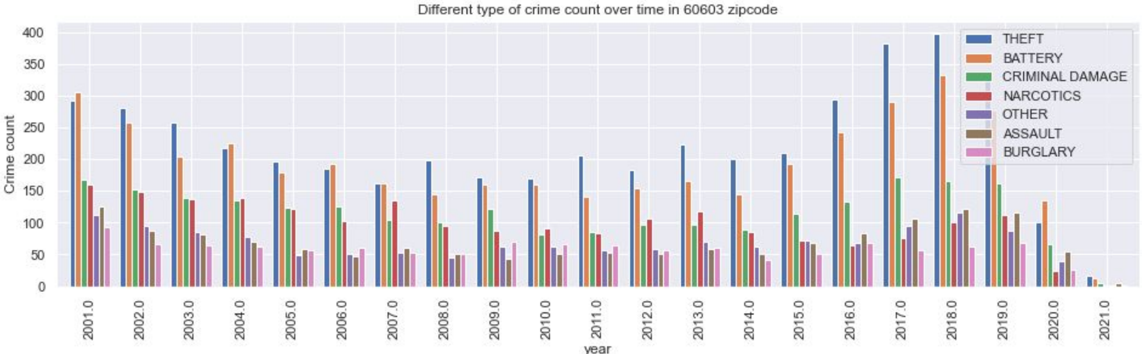


Crime

- Crime rate has decreased in Chicago over time.
- Zip code 60603 has seen a decrease in housing value over the last 5 years. Crime rate has increased in this zip code over this time period.
- Zip codes 60653 saw the highest increase in housing value over this time period and we can see that the crime rate has decreased for this zip code.





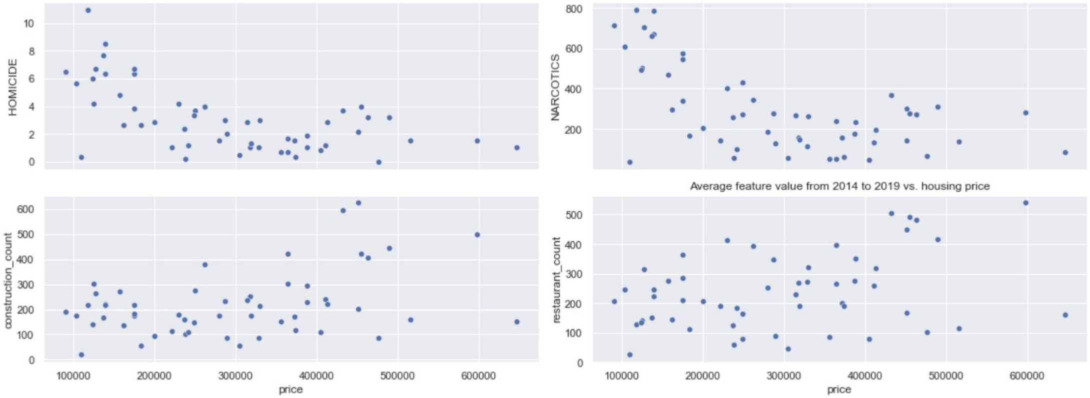


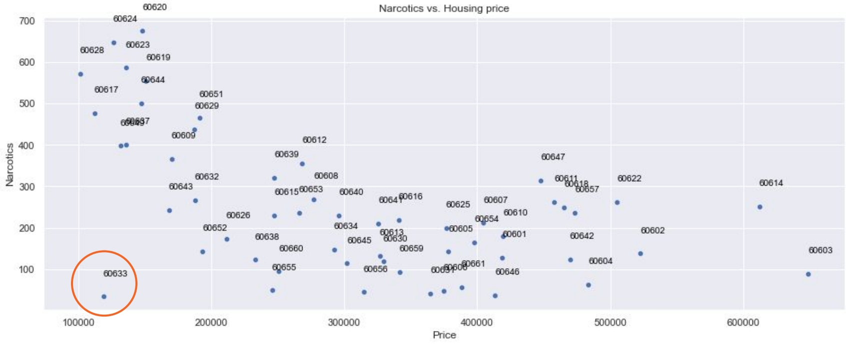
Machine Learning

- **preprocessing**
  - One hot encoding categorical values
  - Scaling numerical values

- **Feature engineering**
  Other than the lagged features and the change in the lagged values, based on the EDA results I added the squared value of some of the features to the feature dataframe as well.

- **Removing outliers**
  Based on the EDA results zip code 60633 was an outlier





Machine Learning

Trying vanilla models

- Trying few different vanilla models with cross validation to see which ones have a better initial performance

- compared model performance to a baseline model.

- The Ridge model had the best performance.

	r2	mae	rmse
DummyRegressor	-0.016509	45179.678083	52563.660426
LinearRegression	0.809325	17815.507418	22532.591576
Ridge	0.848368	15792.808242	20253.714657
ElasticNet	0.604103	25140.266433	32711.531128
RandomForestRegressor	0.663020	21418.122908	30205.066289
SVR	-0.056677	45034.171238	53645.347122
XGBRegressor	0.638123	22133.815950	31167.823736

Model selection (Hyperparameter tuning)

Ridge model performance:

used mean absolute error as the scoring metric because it's less sensitive to outliers compared to root mean squared error.

Results:

train MAE: 18613

test MAE: 19864

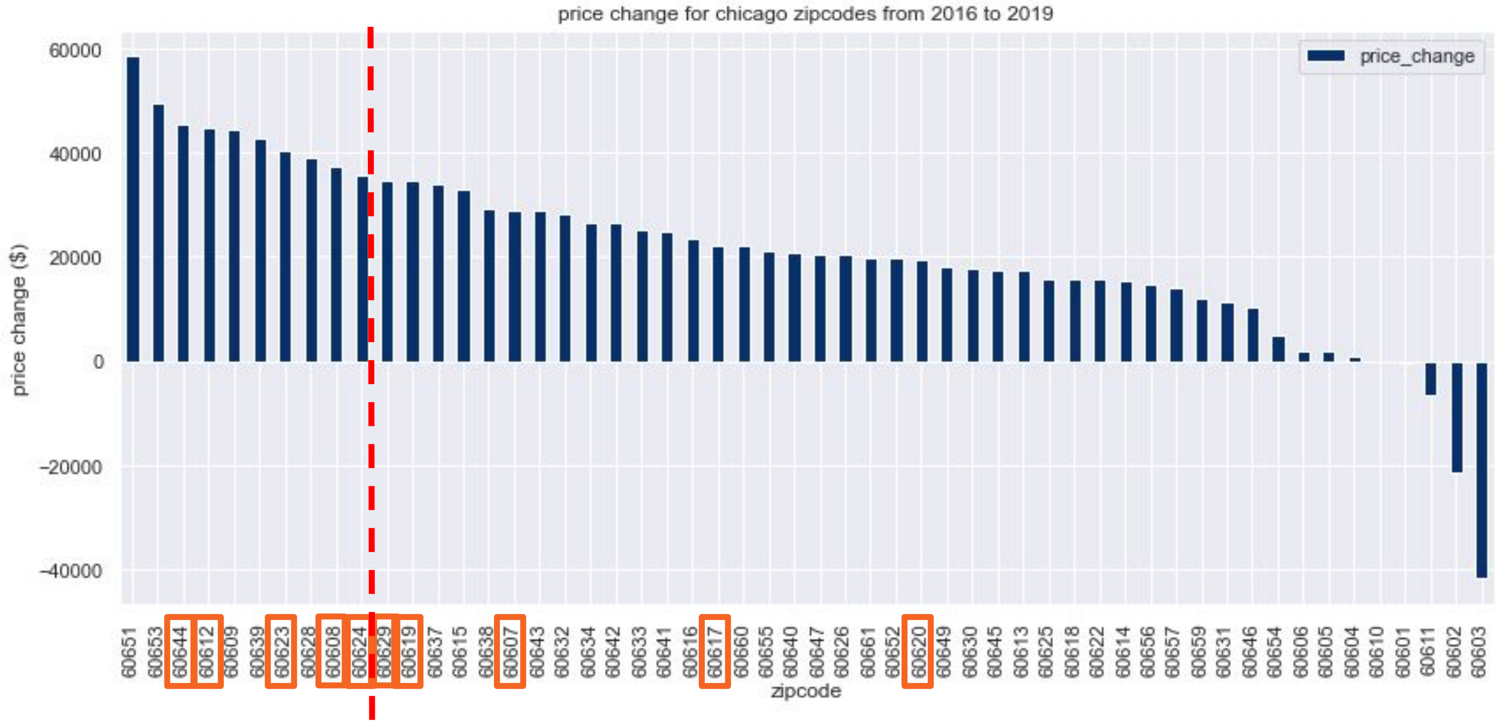
test RMSE: 25827

	feature_importance
price_lag3_value	0.111911
price_change_lag3_lag4	0.072490
construction_count	0.068175
construction_count_lag5_value	0.056521
NARCOTICS	0.049741
NARCOTICS_squared	0.043937
HOMICIDE_squared	0.043780
construction_count_change_lag3_lag5	0.043260
construction_count_lag4_value	0.039280
construction_count_lag3_value	0.032380

Results

Top 10 zip codes recommended by the model:
60619, 60608, 60629, 60644, 60620, 60612, 60617, 60623, 60624, 60607

Zip codes identified correctly by the model:
60644, 60612, 60623, 60608, 60624



Improvements

- More data:
    - Zip code specific data such as demographics and population over time
    - Adding other type of data (economy, GDP,..)
- More feature engineering