**Predicting Home Prices in NYC**

**95891 - Intro to Artificial Intelligence**

Final Project Proposal

**Project Description**

There are many products and services that exist to help individuals looking to purchase or rent a home. Many of these services such as Zillow and Redfin offer estimates for the value of homes, but this is becoming increasingly difficult to do in today's real estate market, particularly in red-hot areas like NYC. We aim to use a conglomeration of data sources related to NYC (crime, government public works spending, community/job growth) to bolster real estate data from sources such as Zillow and Redfin, to build models to better predict the cost of real estate in New York City.

**Datasets**

* [Redfin](https://redfin.com/) (Home values, Forecasts, List/Sale Prices)
* [Zillow](https://www.zillow.com/research/data/) (Home values, Forecasts, List/Sale Prices)
* [NYC Open Data](https://data.cityofnewyork.us/Housing-Development/DOB-Job-Application-Filings/ic3t-wcy2) (DOB Job Application Filings)
* [NYPD Crime Data](https://www1.nyc.gov/site/nypd/stats/crime-statistics/citywide-crime-stats.page) (Arrest Data, Complaint Data, Shootings Data)
* [Great Schools](https://www.greatschools.org/new-york/new-york/schools) (School Ratings for Prek-12)
* [Housing Maintenance Code Violations](https://data.cityofnewyork.us/Housing-Development/Housing-Maintenance-Code-Violations/wvxf-dwi5) (units that that have been verified to violate the New York City Housing Maintenance Code (HMC) or the New York State Multiple Dwelling Law (MDL)

**Main Challenges**

* Merging datasets and predicting based on added variables on top of Redfin data, such as crime rate, schools, etc.
* Determining the appropraite level of prediction. Some datasets are aggregated at the zip code level, neighborhood level, or individual address

**Evaluation Metrics to Measure Success**

* For supervised regression analysis, we can use Mean Squared Error (MSE), Root Mean Squared Error (RMSE), or Mean Absolute Error (MAE) as model evaluation metrics
* Furthermore, to benchmark our models, we would like to employ a few different methods. Namely:
  + Comparing the “Zestimate” (Zillow Estimate) or Redfin estimate against the actual sale price for properties and calculate error metrics. Then compare our team’s champion model against the actual sale price and calculate error metrics. Finally, we would compare these two errors to understand where our model stands.
  + Another method to evaluate our model is to compare our estimates against the actual sale price and calculate errors (same as above) then compare our model’s errors to a naive model which strictly produces a median or mean sale price at a certain granularity (zip code, neighborhood, census tract, etc.). If our model cannot outperform this simple benchmark, we will need to begin the end to end modeling process again
* Some of the techniques our team hopes to employ is:
  + random forest regression
  + KNN regression
  + Data visualization techniques to explore our data. For example, we can use a correlation matrix to visualize existing relationships between our features