Syeda Jaisha, Rahim Rasool, Carly Schippits

Machine Learning for Public Policy

April 13, 2021

Project Proposal

**Title**

Residential property price estimation using property features and spatial-temporal data

**Team**

Syeda Jaisha (syedajaisha), Rahim Rasool (rahimrasool), Carly Schippits (cschippits)

**Summary**

Each year, state and local governments collect property taxes from homeowners. Specific tax amounts depend on a home’s assessed value. While home values can easily be gleaned if a home was recently sold, state and local governments must come up with a method for estimating the value of homes that have not recently sold. This estimated value is then used to calculate the amount of property taxes owed by the homeowner. Ensuring a good prediction is important, because underpredictions result in less revenue for state and local governments, while overpredictions may result in homeowners being overly burdened by property taxes.

Conventional models used by property tax officials use outdated models to estimate a home's assessed value. These models rely on the average value of individual attributes for the homes that have sold. Assessors then apply the resulting averages to the attributes found at each residence. (Source: [https://www.bloomberg.com/news/features/2021-03-09/racial-inequality-broken-property-tax-system-blocks-black-wealth-building).](https://www.bloomberg.com/news/features/2021-03-09/racial-inequality-broken-property-tax-system-blocks-black-wealth-building)

Our aim is to calculate a fairer and accurate estimate of the home's value using factors that impact the market value of the house such as square footage, number of bedrooms, location, etc.

Moreover, given the context of the current [affordable housing crisis](https://berniesanders.com/issues/housing-all/) in the country, our model also aims to facilitate prospective house buyers make an informed decision based on their preferences, push for a fair valuation of the houses they’re interested in and prevent unfair [costs such as land development costs](https://www.brookings.edu/research/whos-to-blame-for-high-housing-costs-its-more-complicated-than-you-think/) from being passed along to them. This study will also be useful for real estate investors to understand what features influence house prices the most and how to capitalize on those features to get the highest returns on their investment.

References: [This publication](https://jamescitycountyva.gov/DocumentCenter/View/15181/Understanding-Real-Estate-Assessments-Guide-PDF) from the Virginia Department of Taxation explains the importance of accurate property tax assessments for the fiscal health of state government. [This Washington Post article](https://www.washingtonpost.com/business/2021/03/12/property-tax-regressive/) explains that inaccurate property assessments can lead to overburdening homeowners with property taxes.

In recent years, machine learning models that can predict the sale price of a house with a significant degree of accuracy, have been of major interest to stakeholders. This is evident by the traction, competitions designed around the same problem, have gained on Kaggle.

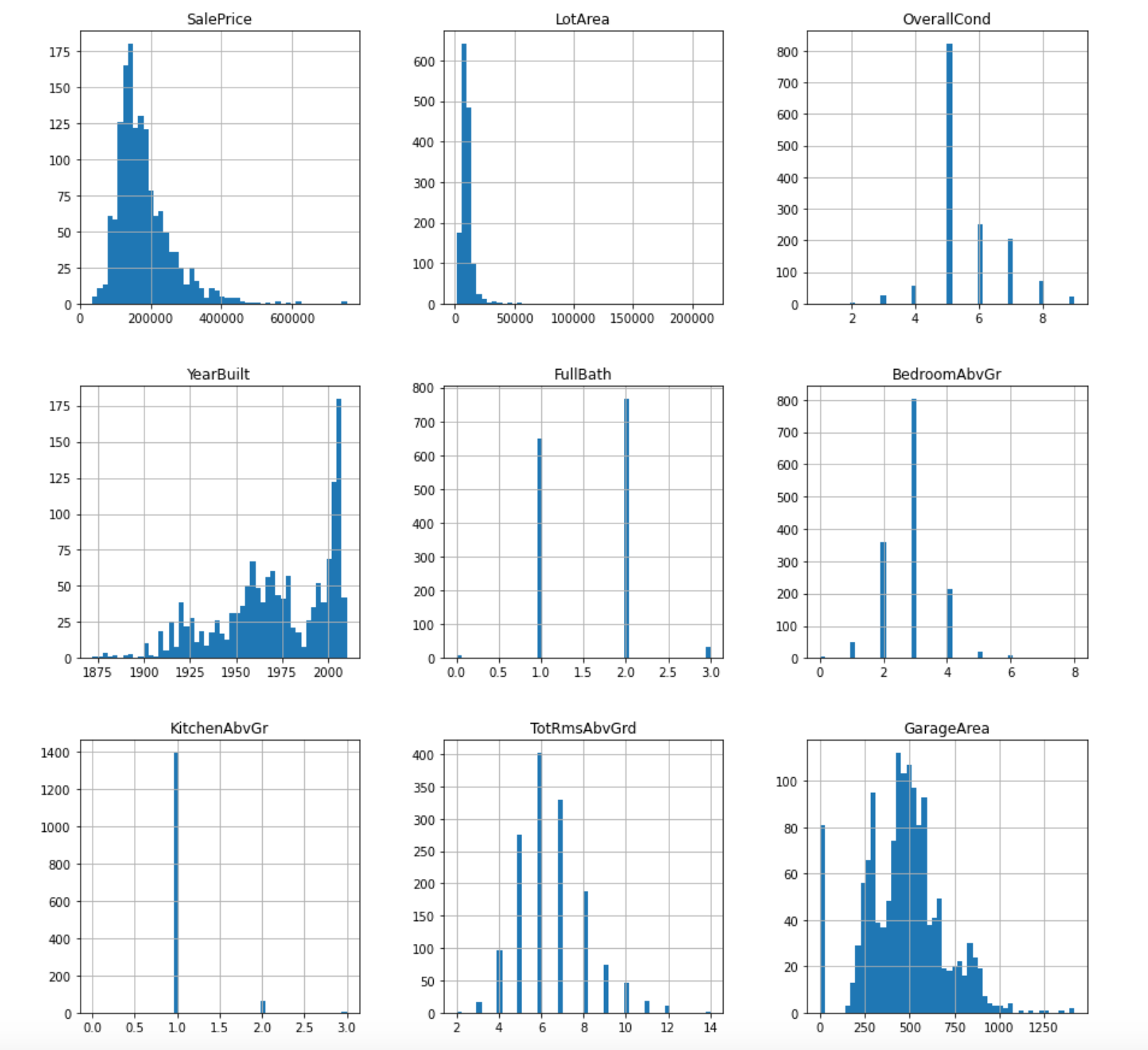
**Data**

We will use a [dataset](https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data?select=train.csv) on homes sold in Ames, Iowa from 2006-2010. The dataset contains 1,460 homes, the sales price for each home, and a set of 77 attributes for each home. The attributes include square footage, number of bedrooms, existence of a basement, etc.

Apart from using the individual attributes of similar homes, we're aiming to combine the dataset with spatial features (such as walkability to the nearest subway, park, market, etc), crime data, socioeconomic and demographic indicators of the region, and temporal sale price/rent data to achieve better valuation.

The data sports a wide range of houses valued at anywhere between $34900 and $755000 and having an area between 1300 and 215245 sq. ft. Each house has both numerical and categorical features associated with it. Some of the notable numerical features include size, year built, overall house condition rating, number of bedrooms and baths while some of the notable categorical features include neighborhood, building type, type of heating and whether the house has central AC or not.

The following frequency distributions of some of the most significant numerical features give a better sense of the overall distribution of the data.



We plan on checking correlation of all the existing features with sale price and then selecting the most relevant ones and deriving new variables from the existing ones that may better explain the variation in our data such as bedroom/square foot of the house area or bath/bed ratio etc. We will also run other exploratory analysis to ensure that our data is representative of the house distribution in Ames. We will also check for and decide on the most suitable way to handle missing values.

**Machine Learning**

This is a regression-based prediction problem. Our results will allow us to predict the values of unsold homes based on the sale price of recently sold homes.

We plan to create a set of relevant indicator attributes and use regression techniques to estimate the property valuation. The regression techniques we’ll use will involve a set of simple regression to bagged and boosted models in order to discover which model gives the optimal performance. We will also experiment the use of clustering techniques (such as DBSCAN) to spatially group homes with similar walkability features. Additionally, we will also check the impact of outliers on our model performance and resultantly detect and remove such data points using outlier detection algorithms.

**Evaluation**

Before beginning the model selection process, we will set aside part of our dataset for testing. After training various models, we will use this test set to assess the models. Because our prediction problem does not involve classification, our primary metric for assessing the fit of the learned model as well as evaluating the accuracy of the predictions will likely be mean squared error (MSE). However, we may decide to use other metrics after we learn about additional metrics for evaluating quantitative regression-based models in class. We would also be using K-fold cross validation to ensure the validity of our training model before finally using it to make predictions for our testing set.

**Ethics**

Our project topic has direct ethical implications. Based on [existing evidence](https://www.washingtonpost.com/business/2021/03/12/property-tax-regressive/), property assessments used by state and local governments tend to overestimate the value of homes owned by low-income individuals, but tend to underestimate the value of homes owned by high-income individuals. The result is that low-income homeowners face a disproportionate property tax burden, while high-income homeowners are more likely to be undertaxed. When evaluating our models, we will be conscious of this bias and use various metrics to assess the fairness of each model. We hope that, due to the number of home features included in our dataset, our predictions may be more accurate (and therefore more fair) than the simpler predictions used by many state and local governments.