COGS 209 Mini-Project Proposal

Lead Author: Allen Zou
Co-authors: Ting-An Lu, Zachary Chao, and Casey Hild

Title: Deciphering House Price Dynamics Using Predictive Modeling on Housing Data

Research Question

The real estate market is influenced by many factors, ranging from the physical attributes of a property to its location and market trends. Understanding the interplay of these factors and their impact on house prices is important for various stakeholders, including homebuyers, sellers, and real estate investors. Therefore, our research question seeks to uncover the underlying factors that drive house prices and investigate the feasibility of constructing a predictive model to estimate house prices based on these factors. With this in mind, our research question seeks to delve into the complexities of house price dynamics by investigating the following:

- What factors influence house prices?
  We aim to identify and quantify the key determinants of house prices. These factors may include structural characteristics such as square footage (lot area), number of bedrooms, and number of kitchens, as well as locational attributes such as neighborhood, proximity to amenities, and regional economic conditions. Additionally, we will explore the influence of housing features such as heating type, garage capacity, and overall condition on property values.
- Can we build a predictive model to estimate house prices based on these factors?
  Leveraging the rich dataset, we seek to develop a robust predictive model that accurately estimates house prices based on a comprehensive set of housing-related features. By employing regression techniques, we aim to capture the relationships between predictor variables and house prices.

Understanding the factors influencing house prices allows for informed decision-making in real estate. From guiding homebuyers to helping sellers strategically price their properties and aiding investors in optimizing their strategies, insights into housing market dynamics are invaluable. Leveraging machine learning techniques and housing-related datasets offers an opportunity to bridge theoretical knowledge with practical applications, empowering stakeholders to navigate the real estate market more effectively.

Data/materials

We will analyze housing transaction data from an ongoing Kaggle competition featuring the Ames Housing dataset. This dataset consists of 79 explanatory variables that describe various aspects of residential homes in Ames, Iowa. The dataset was compiled by Dean De Cock for data science education and serves as a modernized and expanded version of the often-cited Boston Housing dataset. You can download the dataset from Kaggle. It contains the following files.

File name	Format	Description
train.csv	Comma Separated Values (CSV)	The dataset consists of a tabular format with 1,460 rows and 81 columns. Each row represents a distinct property sale transaction, serving as a unique sample in the dataset. The first column, labeled "Id", contains a unique identifier for each property sale. The target variable "SalePrice" is included as the last column, representing the sale price of each property. The features of each property sale transaction are represented by the columns between the first and last

		columns of the dataset. These columns capture various aspects of the property, including its characteristics, amenities, local details, and attributes relevant to the sale. There are more details about the features in the data_description.txt file.
test.csv	Comma Separated Values (CSV)	The dataset consists of a tabular format with 2,919 rows and 81 columns. Each row represents a distinct property sale transaction, serving as a unique sample in the dataset. The first column, labeled "Id", contains a unique identifier for each property sale. The target variable "SalePrice" is included as the last column, representing the sale price of each property. The features of each property sale transaction are represented by the columns between the first and last columns of the dataset. These columns capture various aspects of the property, including its characteristics, amenities, local details, and attributes relevant to the sale. There are more details about the features in the data_description.txt file.
data_description.txt	Text file	This file provides detailed information about the features (columns) present in the train.csv and test.csv datasets. It outlines the meaning and possible values for each feature.

It's important to note that the dataset specifically pertains to residential homes in Ames, Iowa. While the insights gained from this analysis can be valuable, it’s worth noting that the findings may not generalize to other geographical areas. Additionally, the train-test split is a bit strange in this dataset, so some reshuffling and redistributing may be required.

Course impact/relevance

This project intersects with key topics in COGS 209, including linear regression, handling categorical predictors, model selection, and potentially dimensionality reduction.

Outcome(s)

This project aims to develop a predictive model for estimating house prices based on various housing-related features. The primary outcome will be the creation of a robust predictive model using linear regression or any other suitable model (e.g., KNN, support vector machines, perceptron, etc.). Performance evaluation of the model will be conducted using metrics such as mean absolute error (MAE) or root mean squared error (RMSE) to assess its accuracy in predicting house prices. Additionally, we will interpret the coefficients of the linear regression model to gain insights into the relative impact of different features on house prices.

Another potential outcome could involve exploring dimensionality reduction techniques to improve model efficiency and interpretability. Principal Component Analysis (PCA) or feature selection methods could be used to reduce the number of features while preserving as much information as possible. Evaluation of model performance and visualizing features after dimensionality reduction could also provide more insight into which features or information from the dataset are most relevant to the prediction of house prices.