Data Science For Business - Project Proposal

Team G1 - Anna Dominic, Siri Desiraju, Manisha Goyal, Karthikeyan Shanmugam

---

House Value Prediction for Strategic Real Estate Decision-Making

OBJECTIVE

This project aims to develop a predictive model for estimating house values based on a wide range of features, primarily focusing on residential homes in Ames, Iowa. The model will facilitate informed decision-making in various real estate contexts, including purchasing, renovating, and selling properties.

DATASET

House Prices Dataset Ames, Iowa (from Kaggle): This dataset provides a rich set of features for over 1,460 residential homes located in Ames, Iowa, including both the physical attributes of the houses (such as square footage, number of bedrooms, and overall quality) and their location (like proximity to various amenities). The dataset is particularly well-suited for regression analysis, with the target variable being the sale price of the house.

METHODOLOGY

The main goal of our project would be to build a regression model to predict the sale price of houses in Ames, Iowa. To this end, we envision ourselves carrying out the following techniques and methodologies:

- **Data Preprocessing**: Handle missing values, encode categorical variables, and normalize numerical features to prepare the data for modeling. The preprocessing steps are crucial for improving model performance and ensuring that the input data is in a suitable format for analysis.
- **Feature Engineering**: Identify and create new features that could have a significant impact on house values, such as the total living area (sum of basement, first floor, and second floor areas), presence of luxury amenities (such as a pool/ fireplace), as well as location of the house. From the 80+ features present in the dataset, feature selection techniques will be applied to reduce dimensionality and focus on the most predictive attributes.
- **Model Development**: Experiment with various regression techniques, including linear regression, decision trees, random forests, and gradient boosting machines, to predict house values. Advanced models like ensemble methods or deep learning architectures could be explored to enhance predictive accuracy.
- **Model Evaluation and Optimization**: Use cross-validation and appropriate metrics (e.g., RMSE, R^2 score) to evaluate model performance. Hyperparameter tuning and model optimization techniques will be employed to achieve the best possible predictions.

IMPACT

- **Investment Decisions**: The model can serve as a tool for investors and homebuyers to assess the fair market value of properties, aiding in buy or pass decisions and identifying under- or overvalued homes.

- **Renovation Insights**: By understanding the features that most significantly impact house values, homeowners and investors can make targeted renovations that are most likely to increase property value.
- **Market Analysis for Builders**: Builders and developers can use insights from the model to determine which house features and designs are most in demand, guiding construction and development projects to align with market preferences.

CHALLENGES

- The heterogeneity of the housing market means that a model trained on this dataset might not generalize well to all locations or types of properties, necessitating localization or adaptation of the model for different markets.
- The dynamic nature of the real estate market, influenced by economic factors, interest rates, and trends, may affect the model's accuracy over time, requiring regular updates and retraining with new data.
- The psychological and subjective factors influencing house buying decisions, such as curb appeal or neighborhood feel, may not be fully captured by the dataset.

FURTHER ANALYSIS

If resource and time constraints permit, the project will also try to explore comparisons with housing markets in other cities and states, leveraging average income data to assess similarities and differences in housing trends. Some of the datasets that we have sourced for this purpose are given below:

- Housing Prices Seattle (from Kaggle): This dataset comprises extensive details of 505 residential properties in Boston, Massachusetts, USA, covering the period from August to December 2022. It includes various physical attributes and location specifics for each property.
- Housing Prices California (from Kaggle): The California Census Data, published by the US Census Bureau, encompasses a comprehensive set of metrics for each block group in California, including population statistics, median income, median housing price, and more. With 20,640 districts represented in the dataset, the project's objective is to construct a model capable of predicting median house values in California.
- Housing prices New York (from Kaggle): This dataset contains prices of New York houses, providing valuable insights into the real estate market in the region. It includes information such as broker titles, house types, prices, number of bedrooms and bathrooms, property square footage, addresses, state, administrative and local areas, street names, and geographical coordinates.
- Housing prices Miami (from Kaggle): The dataset provides information on 13,932 single-family homes sold in Miami. It includes various features such as the sale price, land area, floor area, distance to amenities like rail lines, ocean, water bodies, central business district, subcenter, and highways. Additionally, it covers attributes like the age of the structure, presence of airplane noise exceeding an acceptable level, structure quality, sale month in 2016, latitude, and longitude coordinates.

By combining these datasets and using features common across them, and leveraging the insights from our primary regression model to predict house prices in Ames, Iowa, we envision building a secondary, generalizable model that scales across different cities.