Forecast Grade of House in Seattle

This project will employ several supervised learning algorithms to forecast the house grade in Seattle. The goal is to construct a model that accurately predicts house grade based on the 19 feature columns. The best candidate algorithm will be chosen from preliminary results and will be further optimized based on the historical house transaction data.

The target measure is the overall grade given to the housing unit, based on King County grading system. It is whole number varying from 1 to 13.

Problem Statement

I live in Seattle and am pretty interested in the following questions

- What are the key features determining house grade in Seattle?
- What are the busiest times of the year to buy/sell house in Seattle? By how much more do transactions spike?
- Can I predict house grade based on the given 19 features?

Datasets and Inputs

I will use Kc_house_data.csv from Kaggle. It contains following 19 house features along with 21613 observations from 2014 May to 2015 May.

- id: a notation for a house
- date: Date house was sold
- price
- bedrooms
- bathrooms
- sqft_living
- sqft_lot
- floors
- waterfront
- view: Has been viewed
- condition
- sqft_above
- sqft_basement
- yr_built
- yr_renovated
- zipcode
- lat
- long
- sqft_living15: Living room area in 2015(implies-- some renovations) This might or might not have affected the lotsize area
- sqft_lot15: lotSize area in 2015(implies-- some renovations)

- grade

Solution Statement

I will

- calculate the correlation between features and price to get the primary factor determine price.
- analyze the price distribution along time to answer question 2.
- run basic linear regression model, compare it with AdaBoost, KNN algorithms to see what are each model's strengths/weakness. And will pick the best candidate algorithm based on the F score and accuracy score.

Benchmark Models

Typical supervised learning models can solve this problem. Some benchmark models are:

- Gaussian Naive Bayes (GaussianNB)
- Decision Trees
- Ensemble Methods (Bagging, AdaBoost, Random Forest, Gradient Boosting)
- K-Nearest Neighbors (KNeighbors)
- Stochastic Gradient Descent Classifier (SGDC)
- Support Vector Machines (SVM)
- Logistic Regression

Evaluation Metrics

R2 Score will be used to measure the accuracy of the classifier and evaluate the performance.

Project Design

Here is the project outline

1. Clean data and summary stats (distribution, correlation matrix and heatmap)
2. Data Preprocessing
    a. Feature scaling and normalizing
    b. Creating training and testing groups
3. Three models
    a. Basic linear regression model (least squares regression)
    b. KNN
    c. AdaBoost
4. Compare the 3 models and tune the best candidate algorithm.
5. Conclusion