Final Project Report

CPT_S 437

Project Title: House Price Prediction

Team Name: **Project 8**

Team Members:

Ryder Swanson, Yanlei Song, Bowen Wang, Jiaming Chu,

Table of Contents:

1. Introduction

1.1. Background and Motivation:

Predicting housing prices is a complex and important problem in real estate, benefiting stakeholders such as buyers, sellers, and agents. By using machine learning techniques, this project aims to provide an accurate and interpretable model for predicting housing prices based on available data.

Our goal is to evaluate and compare different machine learning models to predict housing prices. We specifically focus on the following:

- Experimenting with various algorithms: GBM, XGBoost, and Random Forest.
- Optimizing the models by tuning hyperparameters and filtering noisy features.
- Assessing the performance of each model using RMSE, R^2, and MAE.

2. Dataset and Preprocessing

2.1. Dataset Description:

Our dataset was sourced from publicly available resources provided by the State of Washington. We utilized a comprehensive dataset containing information on all parcels of land within the state, which we subsequently filtered to focus specifically on Whitman County. We then leveraged this information to access corresponding county assessor website details, thereby obtaining relevant data points such as number of rooms and square footage.

2.2. Data Scraping:

The most labor-intensive aspect of this project was developing a custom data scraping tool to retrieve information from the county assessor website. This required designing and implementing involved pattern matching algorithms and parallel processing capabilities to optimize the program's efficiency. Additionally, the website's anti-scraping

measures required circumvention, introducing an extra layer of complexity to the development process.

2.3. Data Cleaning :

The data cleaning process for this dataset involved a straightforward approach. Initially, we eliminated parcels lacking a valid address or without any structures. Subsequently, we removed parcels with incomplete data, including those missing square footage or age information. This refining process yielded a robust dataset consisting of approximately 10,000 properties with sufficient data.

3. Methodology

3.1. Overview of Models:

We compared three machine learning algorithms:

1. **Gradient Boosting Machine (GBM)**: A tree-based boosting method optimized for high performance.
2. **XGBoost**: A highly efficient gradient boosting library.
3. **Random Forest**: A bagging-based ensemble method with interpretable results.

3.2. Hyperparameter Tuning:

We extensively tuned each model's hyperparameters to balance bias and variance.

GBM Tuning

- **Number of Trees**: 3,000
- **Interaction Depth**: 8
- **Shrinkage (Learning Rate)**: 0.005
- **Minimum Observations per Node**: 8

- **Bagging Fraction**: 80%
- **Cross-Validation Folds**: 5

XGBoost Tuning

- **Number of Rounds**: 2,000
- **Max Depth**: 6
- **Learning Rate**: 0.01
- **Subsampling**: 80%
- **Column Sampling per Tree**: 80%

Random Forest Tuning

- **Number of Trees**: 500
- **Features per Split**: √(number of features)
- **Node Impurity Measure**: Gini Index.

3.3. Adjustments During Experiments:

Feature Selection:

- Removed low-impact features like **age_squared** and **bedrooms** based on preliminary results to reduce noise and overfitting.

Outlier Removal:

- Focused on 3rd–97th percentile ranges for **Total_Value** and **Total_Area**.

Cross-Validation:

- Used 5-fold cross-validation to ensure robust performance.

Metrics Analysis:

- Focused on reducing RMSE and improving R^2 by fine-tuning the learning rate, tree depth, and number of iterations.

4. Results and Analysis

4.1. Model Comparison:

The comparative performance of the three models—Gradient Boosting Machine (GBM), XGBoost, and Random Forest—revealed that all three yielded very similar predictive capabilities. This observation holds not only for single evaluations but also for results averaged over 100 iterations. The summarized metrics from the 100 iterations are as follows:

Metric	Model	Range (Min, Max)	Average	Median
RMSE	GBM	(85,639.7, 93,816.46)	89,050.01	88,808.52
	XGBoost	(85,116.68, 93,318.99)	88,677.78	88,408.12
	Random Forest	(85,526.78, 94,166.42)	89,390.64	89,231.21
R^2	GBM	(0.499, 0.585)	0.55	0.55
	XGBoost	(0.505, 0.585)	0.55	0.56
	Random Forest	(0.492, 0.582)	0.55	0.55
MAE	GBM	(65,915.59, 71,348.66)	68,133.65	67,966.34
	XGBoost	(65,196.45, 70,292.42)	67,380.51	67,337.89
	Random Forest	(65,055.38, 70,893.94)	67,747.22	67,506.29

4.2. Observations:

Similar Performance Across Models:

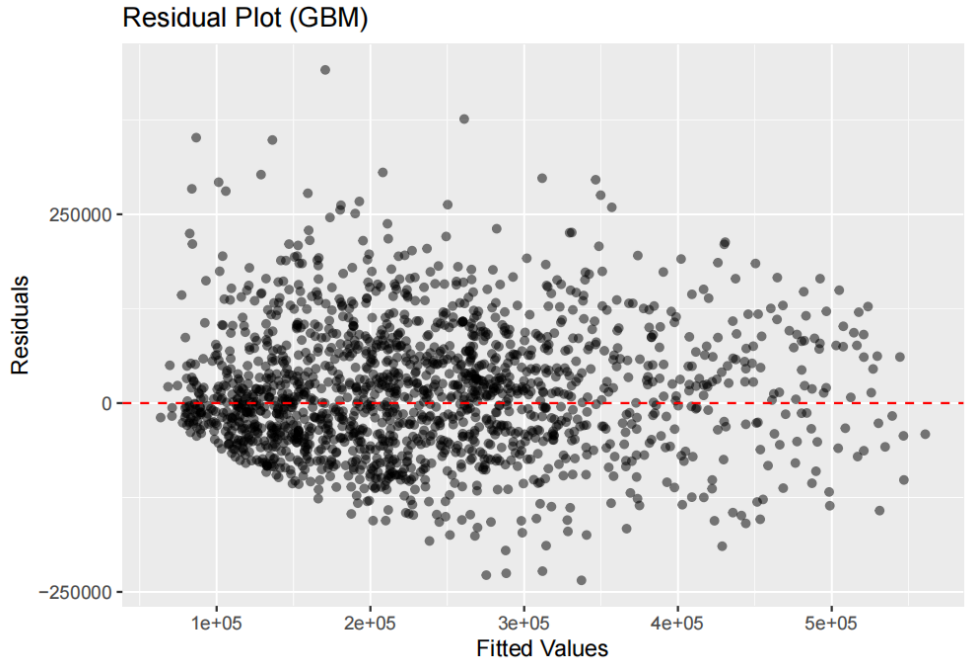
- The RMSE, R^2, and MAE metrics across the three models show consistent ranges, averages, and medians. For example, the RMSE averages across the models differ by less than 1%, with GBM slightly outperforming Random Forest.
- XGBoost achieved the smallest average RMSE (88,677.78) and the highest median R^2 (0.56), but the differences are negligible and fall within the expected statistical variation.

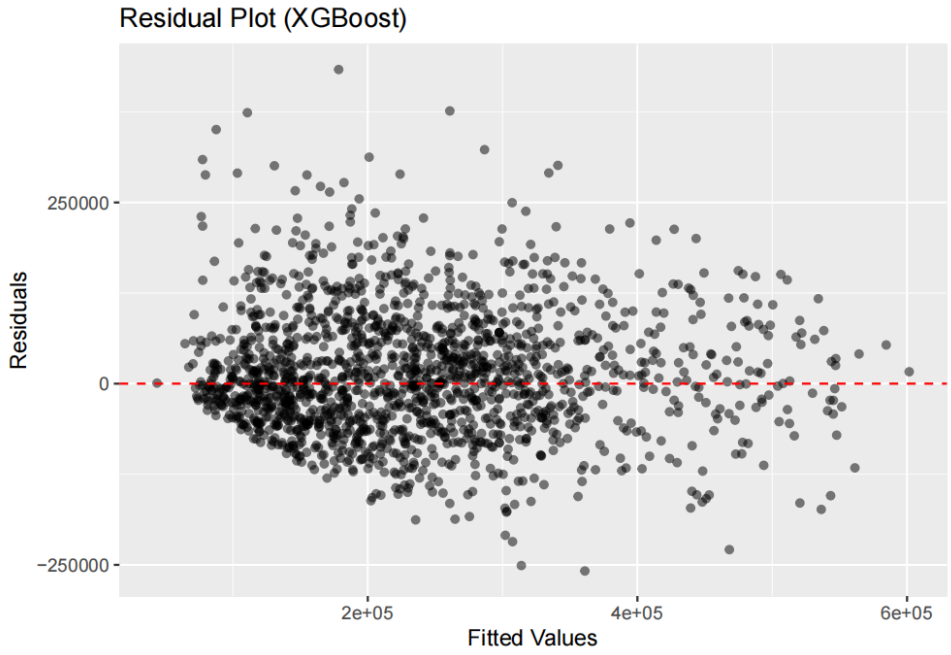
Insights from Repeated Evaluations:

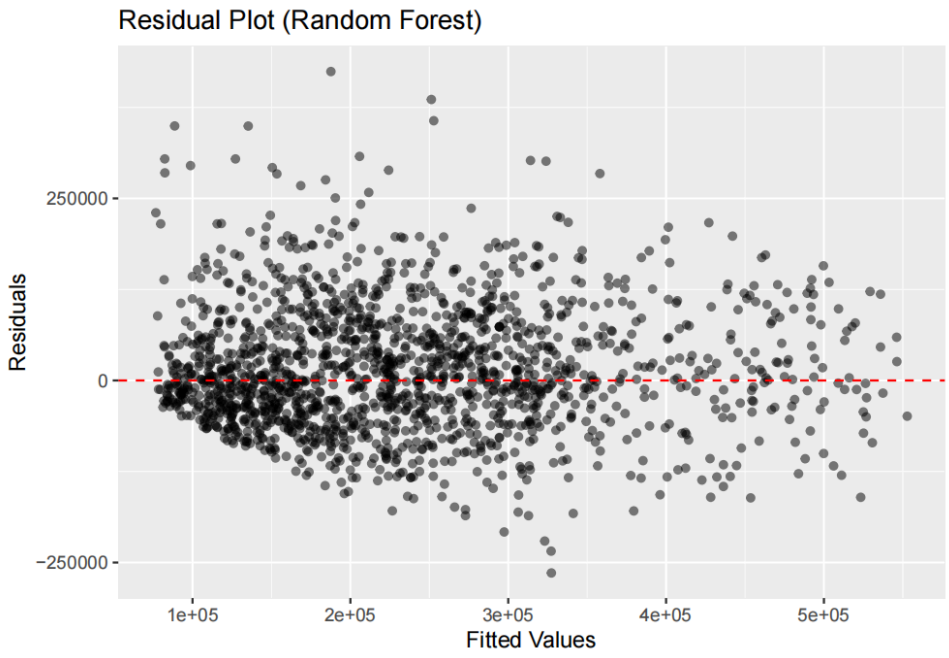
- The variation in RMSE and R^2 across iterations highlights the inherent variability in machine learning predictions, likely due to data splits and random initialization of model parameters.
- The range of R^2 values (approximately 0.49 to 0.59) indicates moderate predictive power, suggesting that the models explain about 50-59% of the variance in housing prices.

Residual Patterns and Consistency:

- The residual plots for all models show similar patterns, with residuals clustered around the horizontal zero line but significant dispersion for larger predicted values.
- Outliers in all models suggest the presence of unaccounted-for factors affecting housing prices, such as neighborhood or location-specific attributes.


Residual Plot (GBM)
Residuals
250000
0
−250000
1e+05
2e+05
3e+05
4e+05
5e+05
Fitted Values


Residual Plot (XGBoost)
Residuals
250000
0
−250000
2e+05
4e+05
6e+05
Fitted Values

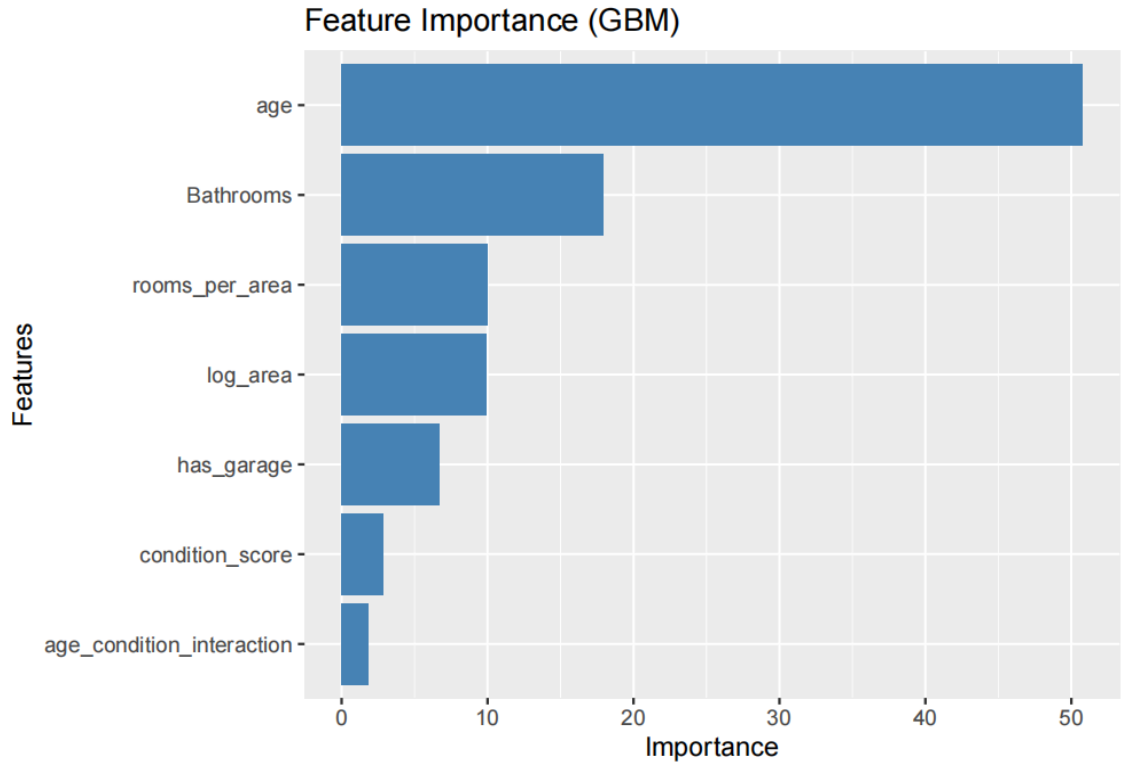


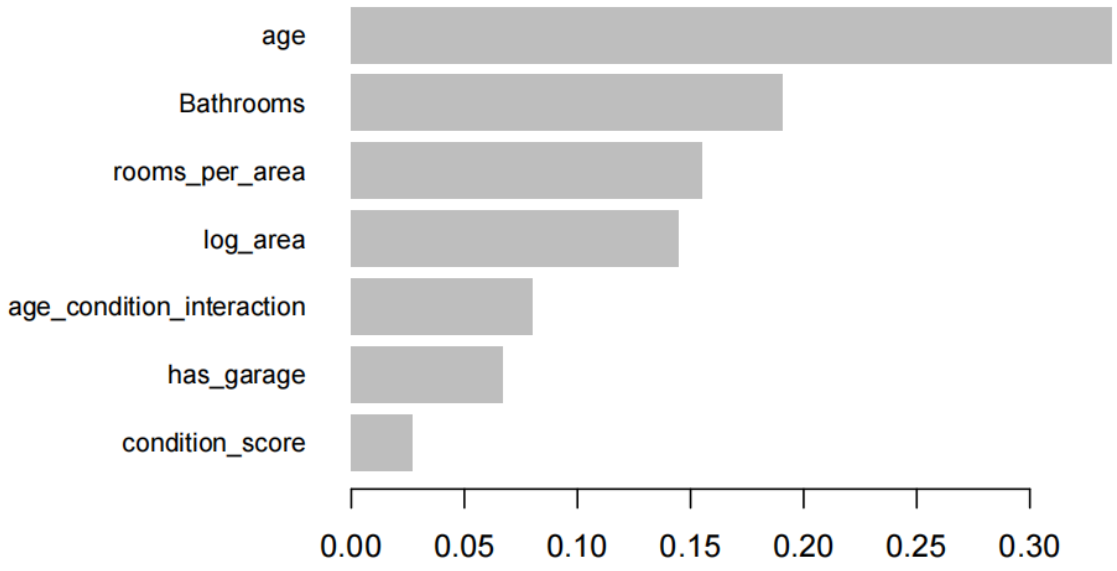
4.3. Feature Importance:

From the feature importance visualizations:

- **GBM and XGBoost:**
  - The **age** of the house consistently emerged as the most important feature, followed by **Bathrooms** and **rooms_per_area**.
  - **log_area** and **has_garage** were moderately important, while **condition_score** and **age_condition_interaction** played smaller roles.

The consistency of feature rankings across GBM and XGBoost suggests that these models are capturing similar underlying patterns in the data.


Feature Importance (GBM)
Features
age
Bathrooms
rooms_per_area
log_area
has_garage
condition_score
age_condition_interaction
0
10
20
30
40
50
Importance


age
Bathrooms
rooms_per_area
log_area
age_condition_interaction
has_garage
condition_score
0.00
0.05
0.10
0.15
0.20
0.25
0.30

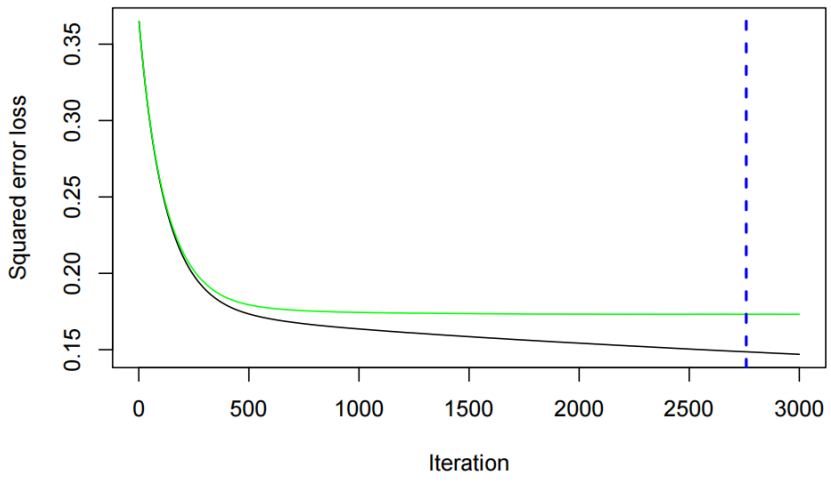
4.4. Challenges and Adjustments:

Dataset and Feature Limitations:

- The absence of location-based features (e.g., proximity to amenities or neighborhood attributes) is likely the primary reason for the moderate R^2 values. Housing prices are inherently location-dependent, and the current dataset only contains address information, which makes it difficult to search for surrounding information one by one.
- The dataset size and variability in house prices further contributed to the models' limited ability to generalize.

Efforts to Address Limitations:

- **Outlier Removal:** By filtering out extreme values (e.g., top and bottom 3% of house prices and areas), we reduced noise and improved model stability.
- **Feature Engineering:** Derived features such as **rooms_per_area** and **age_condition_interaction** were introduced to capture potential interactions, though their impact was limited.
- **Hyperparameter Tuning:** Models were optimized with cross-validation, as evidenced by the GBM loss plot showing convergence around 2,500-3,000 iterations.



5. Future Work

5.1. Future Directions:

Enrich Feature Set:

- Incorporate geospatial and neighborhood data to capture location effects, which are critical for predicting housing prices accurately.
- Use external APIs or datasets (e.g., Zillow, Google Maps) to extract additional features.

Increase Dataset Size:

- A larger dataset with greater variability in house prices and attributes could improve generalizability and model performance.

Explore Alternative Models:

- Experiment with LightGBM or CatBoost for potential improvements in performance and training efficiency.

Advanced Feature Engineering:

- Use clustering techniques to group houses by neighborhood and include these clusters as features.
- Employ domain-specific insights, such as proximity to schools or shopping centers, to refine predictions.

6. Code Demo Link(Slides & Video Included):

https://drive.google.com/file/d/15kXF5jI6aTOE1K2AajyIOxcp35KLgsbw/view?usp=sharing