PROJECT PROGRESS REPORT

Team Members

1. Avinash Ponnada
2. Muttaki I. Bismoy

Progress So Far

Completed Tasks:

Data Acquisition and Initial Exploration:

1. Successfully loaded the dataset from here.
2. Conducted initial exploratory data analysis (EDA) to understand the structure and content of the dataset.

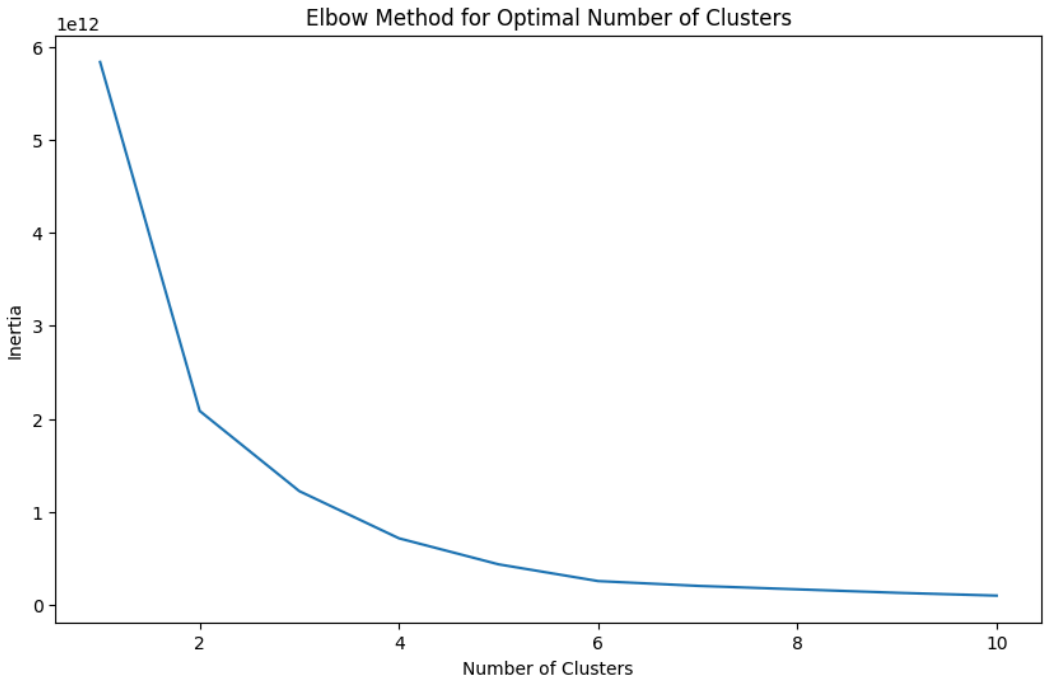

Fig. Elbow Method for Optimal Number of Clusters

Data Cleaning and Preprocessing:

1. Handled missing values by filling in median values for numerical columns.

2. Removed duplicate records to ensure data quality.
3. Addressed inconsistencies and outliers, particularly in the sqft_living feature, by removing properties with abnormal square footage.
4. Engineered new features such as age and has_basement for better predictive modeling.

Clustering Analysis:

1. Applied KMeans clustering to the dataset to identify meaningful clusters.
2. Performed PCA to reduce the dimensions to two for visualization purposes.
3. Visualized the clusters using PCA, providing a clear representation of the cluster assignments in a 2-dimensional space.

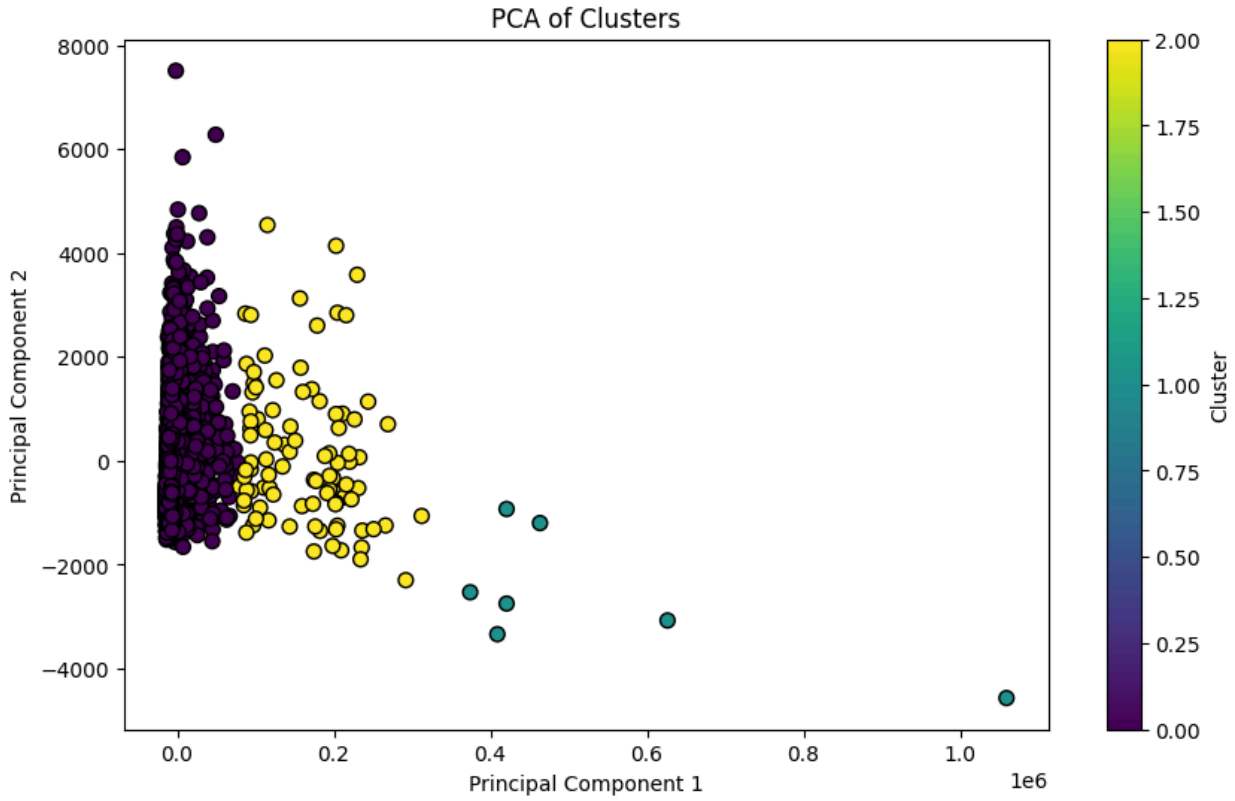

Fig. Principal Component Analysis of Clusters

Regression Models:

1. Trained and evaluated multiple regression models including Linear Regression, Decision Tree Regression, Gradient Boosting Regression, and Random Forest Regression.
2. Implemented GridSearchCV for hyperparameter tuning to improve model performance.
3. Assessed model performance using metrics such as MSE, RMSE, and R-squared.

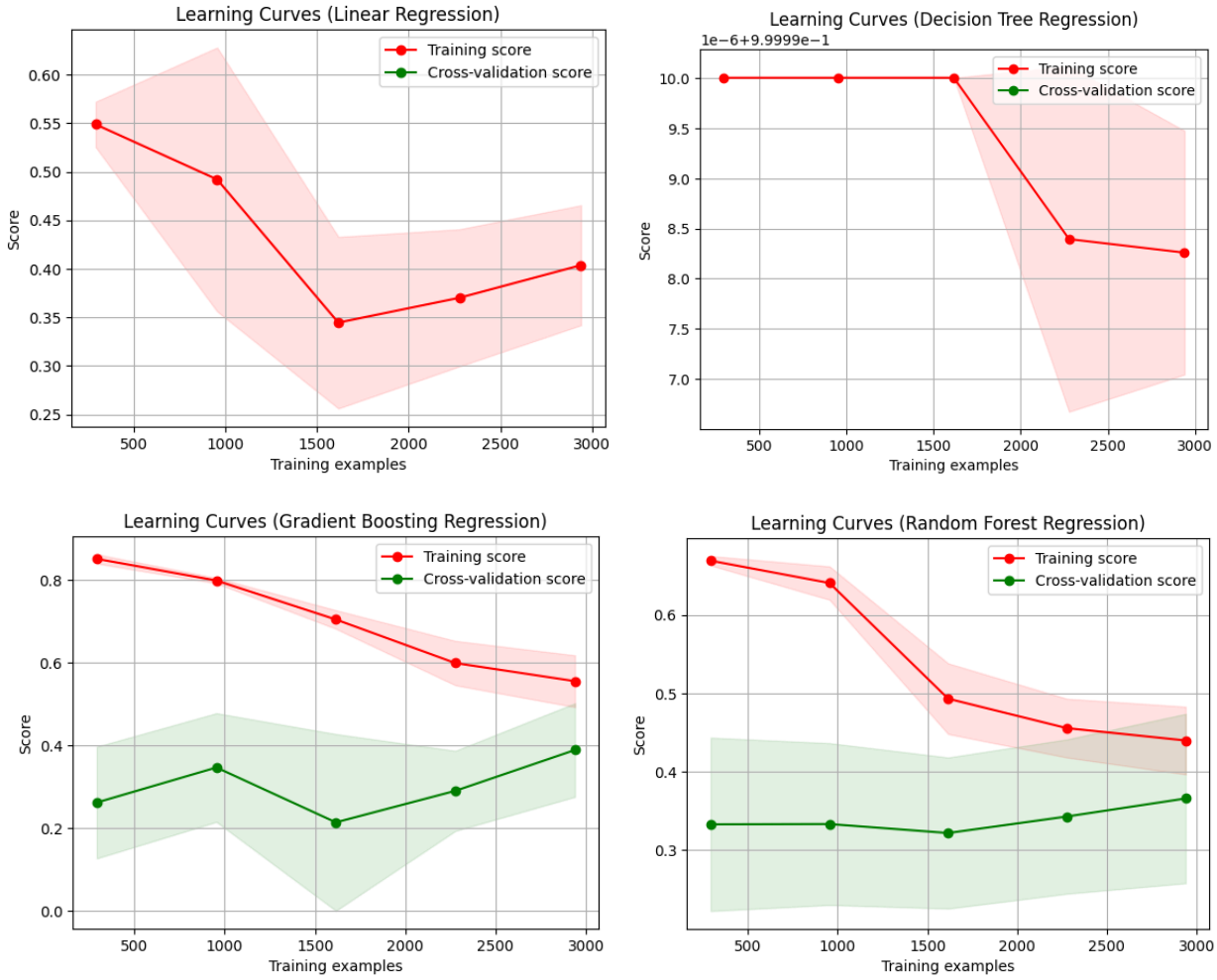

Fig. Regression Models

Data:

Dataset: House Pricing Dataset

Challenges

Encountered Difficulties:

Data Quality Issues:

1. Encountered significant missing values which required careful handling to avoid bias in the analysis.
2. Presence of duplicate records that needed to be removed to maintain data integrity.
3. Inconsistent and outlier values in the sqft_living feature, which affected the clustering results.

Technical Difficulties:
1. Implementing PCA and ensuring accurate visualization of clusters.

Addressed Challenges:
1. Missing Values: Filled missing numerical values with the median to maintain the dataset's integrity.
2. Duplicates: Removed duplicate entries to ensure unique and reliable data.
3. Outliers: Applied filters to remove unrealistic values in the sqft_living feature.
4. PCA Visualization: Successfully reduced dimensions using PCA and visualized clusters to interpret the clustering results effectively.

Collaboration

Meeting Frequency:
- Our team meets twice a week for one-hour sessions each time.

Team Contribution:
- All team members are equally and sufficiently contributing to the project.
- There have been no issues with participation or workload distribution.

Next Steps

Remaining Tasks:

1. Boxplot Visualization for Cluster Comparison:
   Create boxplots for various features (bathrooms, bedrooms, sqft_living, sqft_lot, floors, waterfront, view, condition, sqft_above, sqft_basement, yr_built, yr_renovated, age, has_basement) to compare clusters.
2. Model Fine-Tuning and Validation:
   Further fine-tune the regression models and validate their performance on unseen data.
3. Documentation and Reporting:
   Compile the final report, including detailed documentation of the methodology, results, and conclusions.

Plan to Complete Remaining Tasks:

1. Boxplot Visualization: Utilize the existing cluster assignments to generate and analyze boxplots for the specified features. This will be done in the next team meeting.
2. Model Fine-Tuning: Continue hyperparameter tuning and cross-validation to ensure robust model performance.
3. Documentation: Assign team members specific sections of the report to draft, followed by a collaborative review session.

Potential Challenges:

1. Time Management: Ensuring all tasks are completed within the project timeline while maintaining quality.
2. Data Visualization Clarity: Ensuring the visualizations are clear and interpretable.
3. Handling Outliers: Properly addressing or highlighting outliers in the visualizations.
4. Consistency in Data Interpretation: Ensuring all team members consistently understand and interpret the visualizations correctly.
5. Integration of Visualizations: Seamlessly integrating the visualizations into the final report or presentation.