Bengaluru House Price Prediction Report

1. Introduction

- Bengaluru's rapid growth has made house price prediction essential for stakeholders like buyers, sellers, and investors.
- The project aimed to predict house prices based on features like location, square footage, number of bathrooms, and BHK.

2. Problem Statement

- The goal was to predict house prices in Bengaluru accurately, considering influential factors and using regression models.
- Key objectives included understanding the factors affecting prices, data preprocessing, model development, and selecting the best model among Linear Regression, Ridge Regression, and Lasso Regression.

3. Dataset Overview

- The dataset included features: location, total_sqft, bath, bhk, and price (target variable).
- Preprocessing cleaned the data, making it suitable for modeling.

4. Challenges Faced

- **Missing Data**: Handled missing values in critical columns.
- **Categorical Encoding**: Location feature had over 1,000 unique values, solved with OneHotEncoder.
- **Feature Scaling**: StandardScaler was used for numerical features to ensure consistency in scale.
- **Model Selection**: Trained multiple models and compared their performance.
- **Pipeline Implementation**: Implemented preprocessing, encoding, scaling, and modeling in one streamlined pipeline.

5. Model Training

- **Linear Regression**: A baseline model that showed decent results but had some overfitting, with an R^2 score of **0.8572**.
- **Lasso Regression**: Regularization technique to reduce overfitting and multicollinearity, with an R^2 score of **0.8874**.
- **Ridge Regression**: Best performer with an R^2 score of **0.9180**, effectively handling multicollinearity and overfitting.

Ridge Regression was selected as the final model due to its superior performance.

6. Pipeline Implementation

- A complete pipeline handled data preprocessing, scaling, encoding, and model training, ensuring reproducibility and ease of integration for new data.

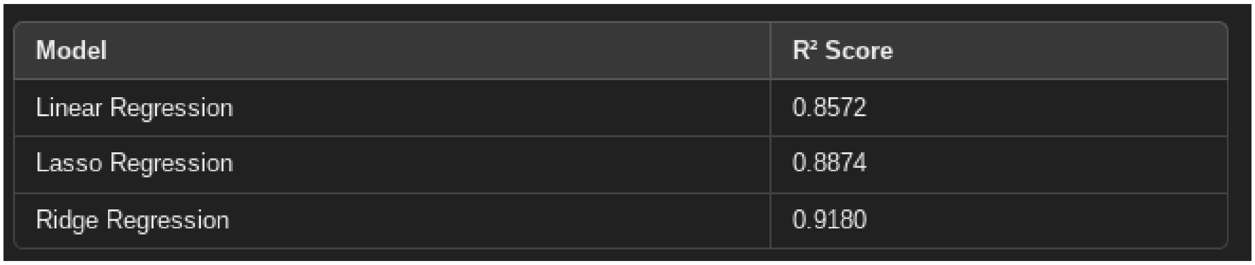
7. Conclusion and Learnings

- **Conclusion**: Ridge Regression gave the most accurate predictions with an R^2 score of 0.9180. Key features influencing house prices included location, total square footage, number of bathrooms, and BHK.
- **Learnings**:
  - High-cardinality categorical features require efficient encoding methods.
  - Pipelines improve workflow efficiency.
  - Regularization (Ridge, Lasso) enhances model performance, especially for multicollinearity.

8. Future Scope

- **Incorporating Additional Features**: Adding features like property age and proximity to landmarks could improve model accuracy.
- **Deploying the Model**: The model could be deployed as a web application for real-time predictions.
- **Advanced Techniques**: Experimenting with ensemble models (e.g., Random Forest, Gradient Boosting) could further enhance accuracy.

9. Final Metrics



Model	R^2 Score
Linear Regression	0.8572
Lasso Regression	0.8874
Ridge Regression	0.9180

Ridge Regression was chosen as the final model due to its superior performance and stability across variations.