Enhancing Real Estate Investment Decisions with Predictive Modeling

### **Introduction**

The real estate industry has long relied on traditional methods and expert intuition to evaluate property values and make investment decisions. However, with the growing availability of data and advancements in machine learning, data science now offers powerful tools to bring accuracy, scalability, and insight into real estate analytics.

This project, *“Enhancing Real Estate Investment Decisions with Predictive Modeling,”* is undertaken for Surprise Housing, a US-based real estate company looking to expand its operations into the Australian housing market. As real estate prices in Australia vary significantly across regions, property features, and market conditions, Surprise Housing needs a data-driven approach to identify profitable investments.

The goal of this project is to develop a robust and accurate machine learning model capable of predicting the sale prices of residential properties in Australia. By analyzing a comprehensive dataset of housing attributes, the model will enable the company to:

* Accurately assess the fair market value of properties,
* Identify the key drivers of property prices,
* Optimize investment decisions, and
* Strategically enter new regions within the market.

Through thorough data preprocessing, exploratory data analysis, feature engineering, and model tuning, this project aims to provide Surprise Housing with actionable insights and a reliable prediction tool for strategic decision-making in the Australian real estate landscape.

### 

### **Problem Statement**

Surprise Housing, a US-based real estate company, plans to expand into the Australian real estate market. To support strategic investments, the company requires a robust machine learning model that predicts the actual sale price of a house. This model will aid in identifying undervalued properties, understanding market trends, and making informed investment decisions.

### **Objective**

* Develop a machine learning model that accurately forecasts house prices using Australian real estate data.
* Identify key features that influence house prices.
* Apply appropriate data preprocessing, EDA, model training, and evaluation techniques.
* Generate actionable insights for Surprise Housing’s market expansion strategy.

### **1. Data Understanding and Cleaning**

#### **Dataset Overview**

* Total observations: 1460
* Total features: 81 (including target variable SalePrice)
* Types of variables: Categorical (object) and Numerical (int/float)
* Two files provided: train.csv (with target) and test.csv (for prediction)

#### **Missing Value Treatment**

* Features with >50% missing values were dropped.
* Remaining missing values handled using:  
  + Median imputation for numerical columns.
  + Mode imputation or 'None' for categorical columns.
  + Domain knowledge used for imputing features like GarageType, BsmtQual.

### **2. Exploratory Data Analysis (EDA)**

#### **Target Variable (SalePrice)**

* Positively skewed distribution; log-transformation applied to normalize.

#### **Univariate Analysis**

* Examined individual variables using histograms, boxplots.
* Example: OverallQual has a strong positive effect on SalePrice.

#### **Bivariate Analysis**

* Correlation heatmaps used to identify features highly correlated with SalePrice.
* Top positively correlated: OverallQual, GrLivArea, GarageCars, TotalBsmtSF
* Negatively correlated: KitchenAbvGr, EnclosedPorch

#### **Outlier Detection**

* Noted outliers in GrLivArea, SalePrice. Removed extreme values to improve model performance.

### **3. Feature Engineering**

* **Created Features:**
  + TotalBathrooms = FullBath + 0.5 \* HalfBath
  + TotalSF = TotalBsmtSF + 1stFlrSF + 2ndFlrSF
  + Age = YrSold - YearBuilt
* **Encoded categorical variables** using:  
  + Label Encoding for ordinal features.
  + One-Hot Encoding for nominal features.

### **4. Model Building**

#### **Models Used**

1. **Linear Regression** (Baseline)
2. **Ridge Regression** (L2 Regularization)
3. **Lasso Regression** (L1 Regularization)
4. **Random Forest Regressor**
5. **Gradient Boosting Regressor**

#### **Hyperparameter Tuning**

* GridSearchCV used to find optimal hyperparameters.
* Example: alpha for Lasso/Ridge, n\_estimators, max\_depth for ensemble methods.

#### **Evaluation Metrics**

* **Mean Squared Error (MSE)**
* **Root Mean Squared Error (RMSE)**
* **R-squared Score (R²)**

### **5. Model Evaluation**

| **Model** | **RMSE** | **R² Score** |
| --- | --- | --- |
| Linear Regression | 0.1440 | 0.91 |
| Ridge Regression | 0.1325 | 0.93 |
| Lasso Regression | 0.1283 | 0.94 |
| Random Forest Regressor | 0.1187 | 0.96 |
| Gradient Boosting | **0.1122** | **0.97** |

Gradient Boosting outperformed others and was selected as the final model.

### **6. Feature Importance**

Top features influencing price:

* **OverallQual**
* **GrLivArea**
* **GarageCars**
* **TotalBsmtSF**
* **YearBuilt**
* **Neighborhood**
* **TotalSF**

### **7. Predictions on Test Set**

* Applied preprocessing and feature engineering on test.csv.
* Used the final model (Gradient Boosting) to generate predictions.
* Submitted results in the required format.

### 

### **8. Conclusion and Business Implications**

* The model helps identify undervalued properties based on objective metrics.
* Surprise Housing can use this predictive tool to prioritize high-return areas and property types.
* Features like Overall Quality, Location, and Living Area are critical for investment decisions.
* The model also highlights non-intuitive influences like basement area and year built, which can guide future data collection.

### **9. Recommendations**

* Invest in properties with high overall quality in well-performing neighborhoods.
* Consider newly built or renovated houses as they fetch higher prices.
* Maintain a data-driven strategy using predictive models for every new market entry.