Project Report: House Price Prediction using Machine Learning

1. Project Objective

The primary objective of this project is to develop an end-to-end machine learning regression pipeline to predict house sale prices. The workflow encompasses data cleaning, exploratory data analysis (EDA), feature engineering, encoding, scaling, and model evaluation to achieve high prediction accuracy on real-world housing data.

2. Dataset Description

The project utilizes a comprehensive housing dataset (Ames Housing–style) featuring a diverse mix of variables:

- **Numerical Features:** Square footage, room counts, and year built.
- **Ordinal Features:** Quality and condition ratings (e.g., "Excellent" to "Poor").
- **Nominal Features:** Neighborhoods, roof styles, and sale types.
- **Target Variable:** `SalePrice` (The final transaction price).

3. Methodology

3.1 Data Cleaning

To ensure data integrity, the following steps were taken:

- **Numerical Imputation:** Missing values were filled using the **median** to minimize the influence of outliers.
- **Categorical Imputation:** Features were filled using the **mode** or a specific "Missing" label where the absence of a feature (e.g., "No Basement") provided semantic value.
- **Error Correction:** Inconsistent data points and invalid entries were manually corrected.

3.2 Exploratory Data Analysis (EDA)

EDA focused on understanding the underlying patterns within the data:

- **Correlation Analysis:** A strong positive correlation was identified between `OverallQual`, `GrLivArea` (Living Area), and the `SalePrice`.

- **Distribution:** Analysis of skewness in the target variable led to the decision to apply transformations.
- **Seasonality:** Trends were observed based on the month of sale, suggesting peak periods in the real estate market.

3.3 Feature Engineering

Raw data was transformed to better represent the underlying problem:

- **Age Calculation:** Year-based features (Year Built/Remodelled) were converted into "Age at Sale" to better capture depreciation.
- **Ordinal Mapping:** Features like `ExterQual` were mapped to numerical scales (1–5) based on their impact on price.
- **Log Transformation:** Applied to highly skewed numerical variables to normalize distribution and stabilize variance.

3.4 Encoding and Scaling

To prepare the data for the regression models:

- **Encoding:** One-Hot Encoding was used for nominal data, while Ordinal Mapping was used for ranked categories.
- **Scaling:** Standard scaling was applied to numerical features **after** the train-test split to ensure no data leakage occurred from the test set into the training process.

---

4. Modeling and Evaluation

4.1 Modeling Approach

The project utilized a tiered modeling strategy:

1. **Baseline:** A simple Linear Regression model to establish a performance floor.
2. **Regularization:** Implementation of Ridge or Lasso regression to handle multicollinearity and prevent overfitting.

4.2 Evaluation Metrics

The models were evaluated using the following metrics:

- **RMSE (Root Mean Squared Error):** To measure the average magnitude of the error.
- **R2 Score:** To determine the proportion of variance in house prices explained by the model.

Results: Feature engineering and the application of regularization significantly reduced the RMSE compared to the baseline model.

---

5. Key Learnings & Conclusion

Key Learnings

- **Feature Engineering:** This stage is the primary driver of performance in tabular machine learning.
- **Data Integrity:** Domain-aware handling of missing values prevents the model from learning biased patterns.
- **Pipeline Structure:** Building a modular pipeline is essential for reproducibility and model maintenance.

Conclusion

This project successfully demonstrates a robust workflow for real estate valuation. By meticulously cleaning data and engineering domain-specific features, the model achieved a high degree of predictive power, proving the effectiveness of structured machine learning pipelines in solving complex regression problems.
