# House Price Prediction - Kaggle Competition

## 1. Introduction

This project focuses on predicting house prices using the Kaggle dataset. We applied data preprocessing, feature engineering, and advanced regression models such as XGBoost. The goal was to minimize prediction error and achieve a high R² score.

## 2. Dataset

The dataset contains housing attributes with the target column `SalePrice`.

|  |  |
| --- | --- |
| Feature Type | Description |
| Numerical Features | LotArea, OverallQual, YearBuilt, GrLivArea, etc. |
| Categorical Features | Neighborhood, HouseStyle, Exterior1st, etc. |
| Target | SalePrice - the final house price to predict. |

## 3. Preprocessing & Feature Engineering

- Handled missing values using median/mode imputation.  
- Applied Label Encoding for categorical variables.  
- Created new features such as TotalSF = TotalBsmtSF + 1stFlrSF + 2ndFlrSF.  
- Normalized skewed numerical features.

## 4. Model Training

We used XGBoost Regressor with optimized hyperparameters. The model was trained on 80% training data and validated on 20% validation data.

## 5. Evaluation Metrics

|  |  |
| --- | --- |
| Metric | Score |
| R² Score | 0.9182 |
| RMSE | 25053.77 |
| RMSLE | 0.1347 |

## 6. Conclusion

The model achieved high predictive performance with R² above 0.91. XGBoost proved to be effective in handling both categorical and numerical data with feature engineering.