

# Housing Price Prediction Using Machine Learning

**Student:** Neelam Pandey

**Project:** Capstone Two – Data Science

**Mentor:** Himanshu Agarwal

**Date:** January 2026

**GitHub:**

<https://github.com/npandey472/CapstoneTwoProject>

# Project Overview

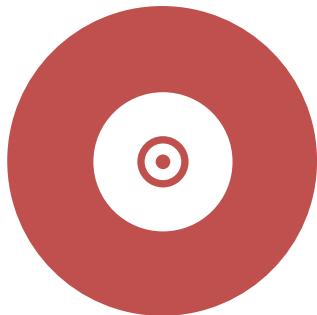
- Predict housing prices using ML

- Complete ML workflow

- Multiple regression models

- Random Forest selected

# Problem Identification



- **TARGET VARIABLE:**  
SALEPRICE



- **SUPERVISED LEARNING**  
– REGRESSION



- **REAL ESTATE PRICE  
PREDICTION**

# Dataset Description

- ~2,900 records

- Numerical & categorical features

- Structural & zoning attributes

# Feature Engineering

- HouseAge = CurrentYear – YearRemodAdd

- Basement Finished Portion

- Dropped original columns

# Handling Missing Values

- Mean imputation for basement feature

- No missing values in target

# Encoding & Scaling

- One-Hot Encoding for categorical features

- StandardScaler used for SVR

MSSubClass	1.00	-0.20	-0.07	0.03	0.04	-0.07	-0.22	-0.06
LotArea	-0.20	1.00	-0.03	0.03	0.03	0.09	0.26	0.24
OverallCond	-0.07	-0.03	1.00	-0.38	0.04	0.04	-0.18	-0.06
YearBuilt	0.03	0.03	-0.38	1.00	0.61	-0.03	0.41	0.37
YearRemodAdd	0.04	0.03	0.04	0.61	1.00	-0.06	0.29	0.36
BsmtFinSF2	-0.07	0.09	0.04	-0.03	-0.06	1.00	0.09	-0.01
TotalBsmtSF	-0.22	0.26	-0.18	0.41	0.29	0.09	1.00	0.43
SalePrice	-0.06	0.24	-0.06	0.37	0.36	-0.01	0.43	1.00



# Exploratory Data Analysis

- Right-skewed SalePrice distribution
- Weak linear correlations
- Presence of outliers

# Models Implemented



- LINEAR  
REGRESSION



- SVR

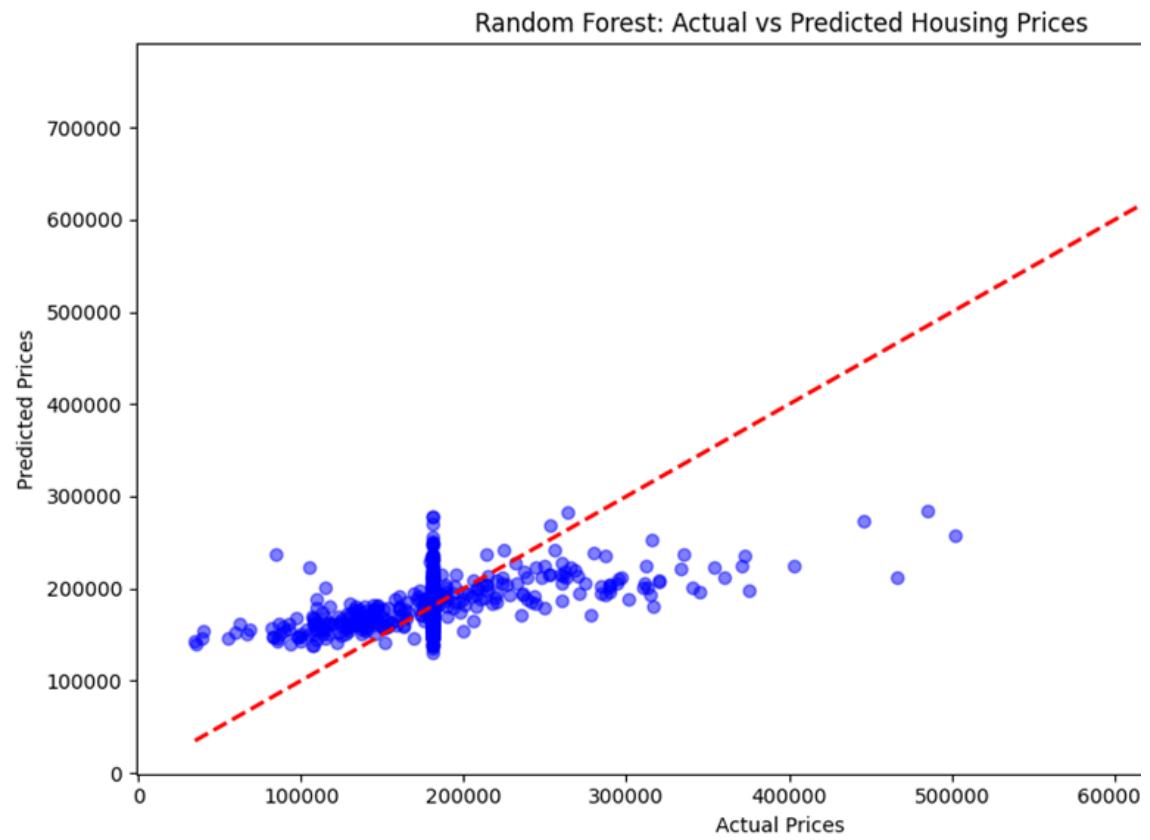


- RANDOM  
FOREST

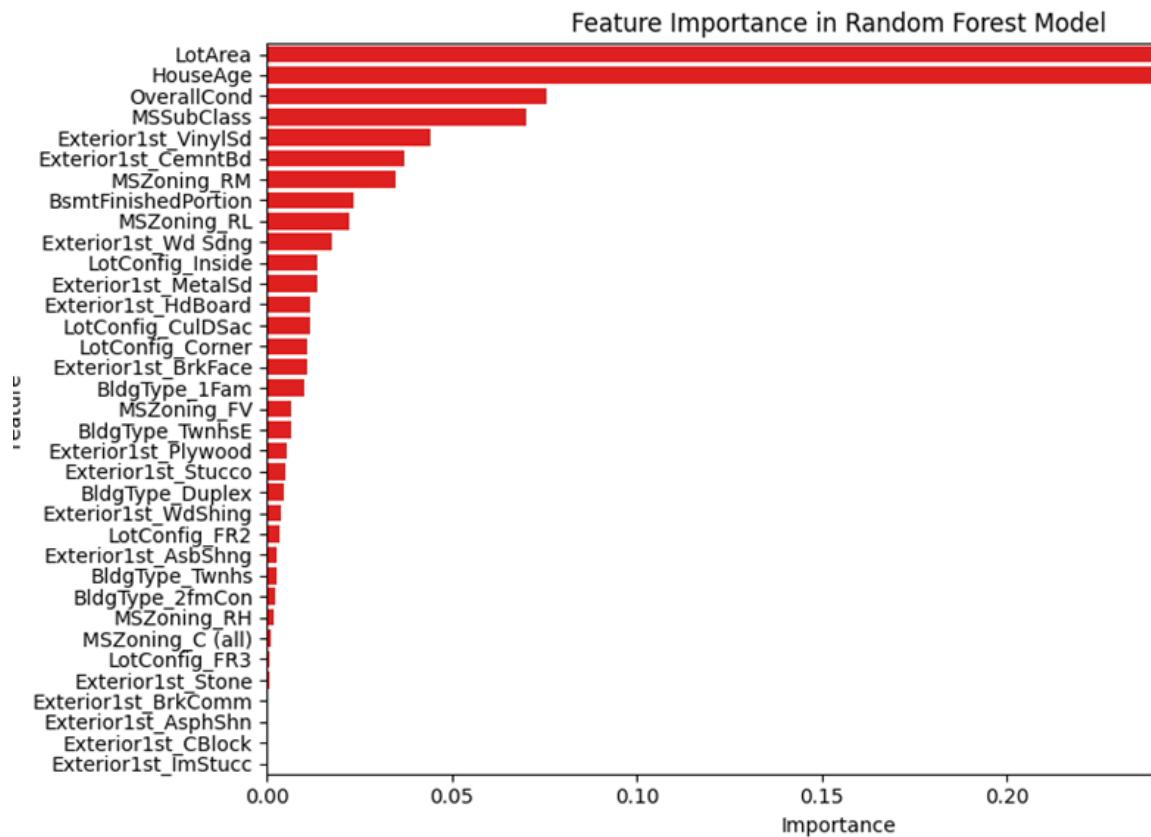


- CATBOOST

# Random Forest Performance



# Random Forest Performance



## Model Comparison

- Random Forest outperformed all other models

Model	R <sup>2</sup>	MAE	MAPE
Linear Regression	—	—	20.13%
SVR	~0.00	~\$28k	~18%
Random Forest	~0.30	~\$33k	~20%
RF (Log Target)	~0.26	~\$33k	~20%
CatBoost	~0.23	~\$35k	~22%

# Conclusion



- DATA QUALITY MORE IMPORTANT THAN MODEL COMPLEXITY
- RF PROVIDES BEST BALANCE

# Future Work

- Add neighborhood & living area data

- Try XGBoost

- Deploy as web app