Capstone Two – Final Project Report

Housing Price Prediction Using Machine Learning

Student Name: Neelam Pandey
Course: Capstone Two – Data Science
Instructor: Himanshu Agarwal
Date: January 2026
GitHub Repository: https://github.com/npandey472/CapstoneTwoProject

---

Abstract

This project applies machine learning techniques to predict residential housing prices using a structured dataset containing numerical and categorical property features. The workflow includes data wrangling, feature engineering, exploratory data analysis (EDA), model training, hyperparameter tuning, and evaluation. Multiple regression models were implemented, including Linear Regression, Support Vector Regression (SVR), Random Forest, and CatBoost. The Random Forest model was selected as the final model due to superior performance and interpretability. Results show that model accuracy is constrained primarily by dataset limitations rather than algorithm complexity.

---

1. Problem Identification

Accurately predicting house prices is an important problem in real estate analytics. Given historical housing data with property attributes, the objective is to build a regression model capable of predicting the sale price of a house.

Target Variable:

- SalePrice

Problem Type:

- Supervised learning – Regression

---

2. Dataset Description

HousePricePrediction.xlsx

The dataset contains approximately **2,900 housing records** with both numerical and categorical features describing structural and zoning characteristics.

Key Features:

- MSSubClass – Dwelling type
- MSZoning – Zoning classification
- LotArea – Lot size (sq ft)
- LotConfig – Lot configuration
- BldgType – Building type
- OverallCond – Overall condition rating
- Exterior1st – Exterior material
- YearBuilt, YearRemodAdd – Construction and renovation year
- BsmtFinSF2, TotalBsmtSF – Basement attributes

---

3. Data Wrangling & Feature Engineering

3.1 Feature Engineering

HouseAge

HouseAge = CurrentYear − YearRemodAdd

This captures the effective age of the property after remodeling.

Basement Finished Portion

BsmtFinishedPortion = BsmtFinSF2 / TotalBsmtSF

Original columns were dropped after feature creation.

3.2 Handling Missing Values

- BsmtFinishedPortion missing values were filled using the **mean**.
- No missing values remained in the target variable.

3.3 Encoding Categorical Variables

Four categorical variables were converted to dummy variables using **One-Hot Encoding**:

- MSZoning
- LotConfig
- BldgType
- Exterior1st

3.4 Feature Scaling

- **StandardScaler** was applied where required (SVR).
- Tree-based models were trained on unscaled features.

3.5 Train–Test Split

- **80% training**
- **20% testing**
- random_state = 42

---

4. Exploratory Data Analysis (EDA)

EDA was conducted to understand feature distributions and relationships.

Included Visualizations

- Sale price distribution
- Correlation heatmap
- Actual vs predicted plots (per model)
- Feature importance plots

The distribution of housing prices reveals a strong right skew with several high-value outliers, as shown in **Figure 1**. These extreme values increase prediction difficulty and disproportionately affect error-based metrics such as RMSE. Additionally, the correlation heatmap in **Figure 2** indicates weak linear relationships between most numerical features

and *SalePrice*, suggesting that linear models alone may be insufficient for capturing underlying patterns.

Figure Captions

Figure 1: *Distribution of Sale Prices.*
This figure shows the distribution of housing sale prices in the dataset. The distribution is right-skewed, with a concentration of properties in the lower-to-mid price range and a small number of high-value outliers.

Figure 2: *Correlation Heatmap of Numerical Features.*
This heatmap illustrates pairwise correlations among numerical features and the target variable, *SalePrice*. Most features exhibit weak linear correlations with the target.

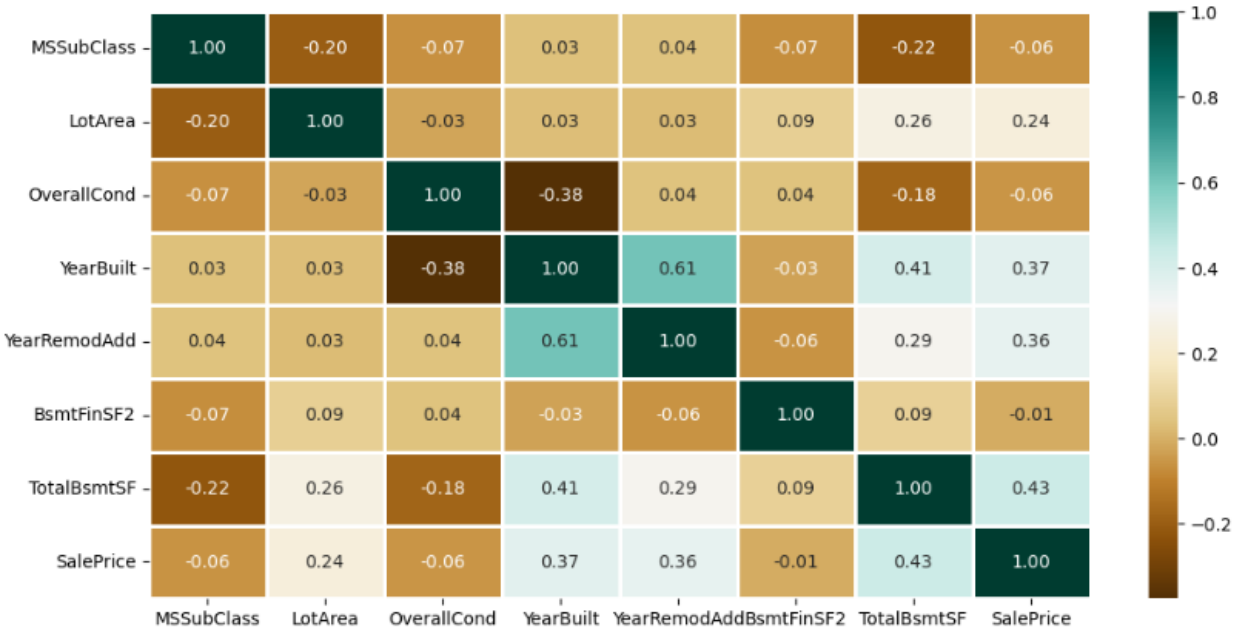


Figure 3: *Actual vs. Predicted Sale Prices Using the Random Forest Model.*
This scatter plot compares actual housing prices with predictions generated by the Random Forest model on the test dataset. The dashed diagonal line represents perfect predictions.

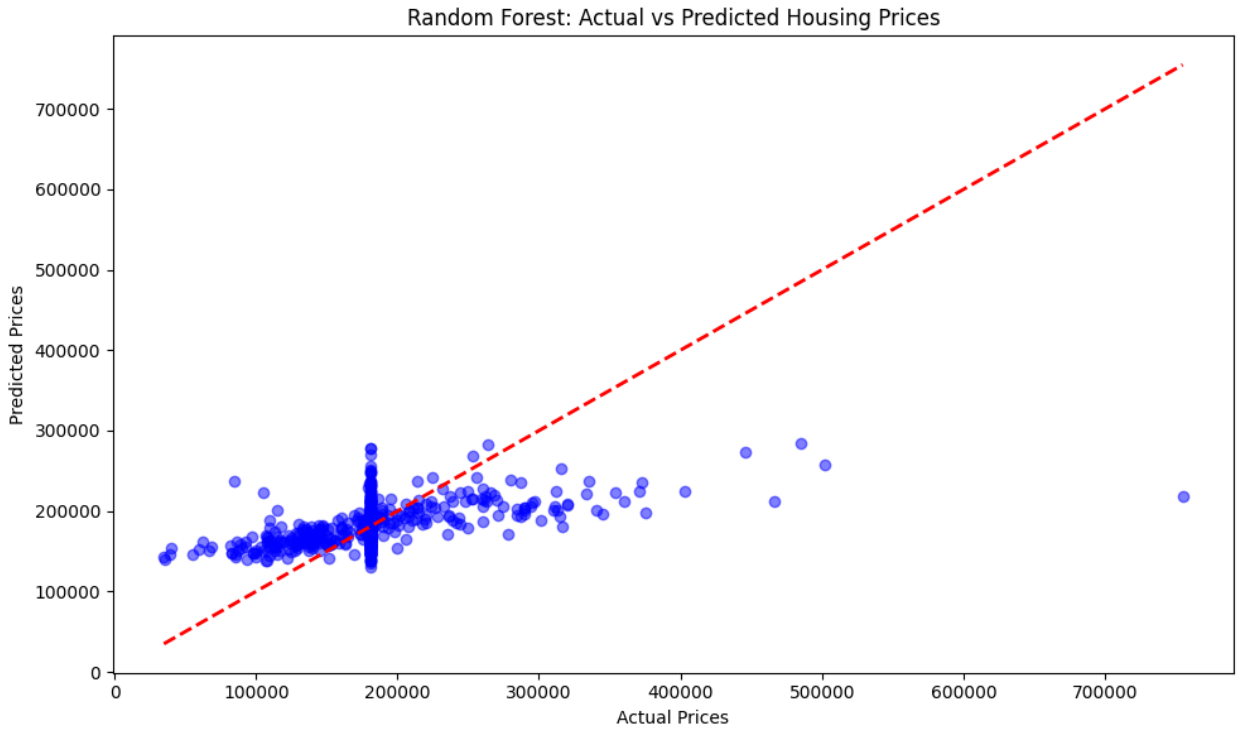
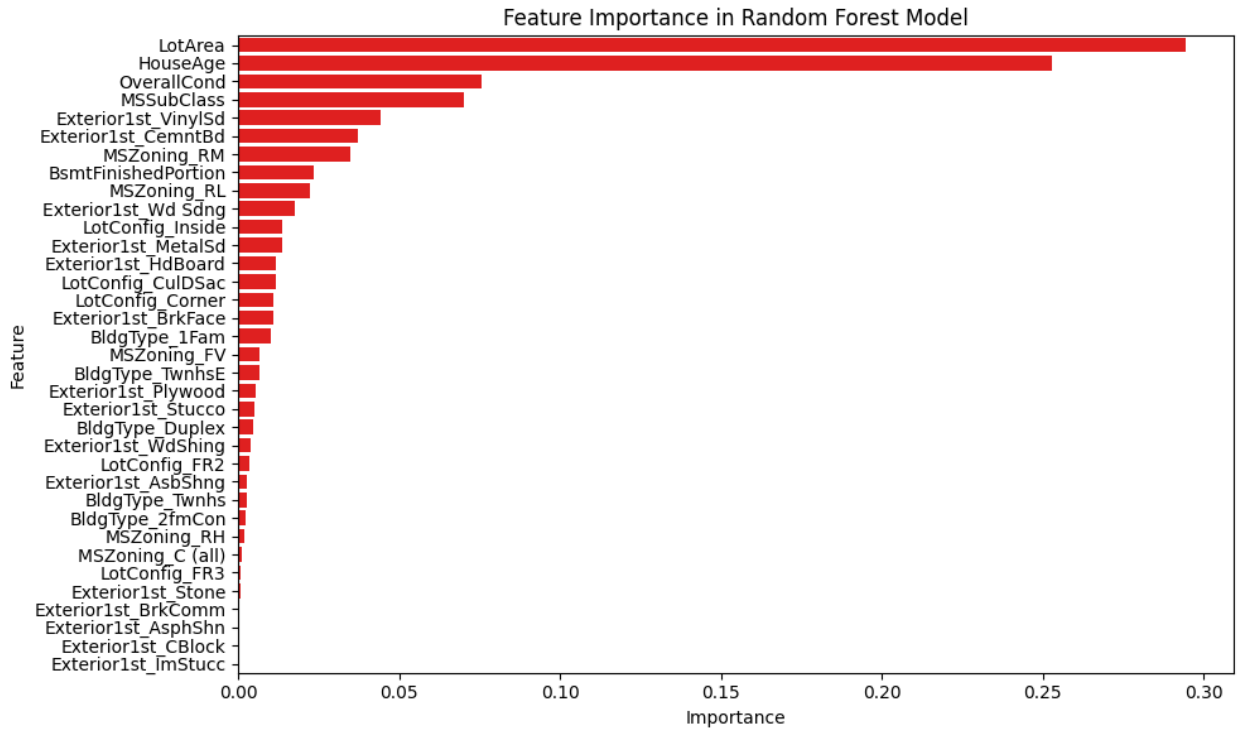


Figure 4: *Feature Importance Derived from the Random Forest Model.*
This bar chart displays feature importance scores obtained from the Random Forest model, highlighting LotArea and HouseAge as the most influential predictors.



---

5. Models Implemented

At least **three different models** were built and evaluated:

1. **Linear Regression (Baseline)**
2. **Support Vector Regression (SVR)**
3. **Random Forest Regression**
4. **CatBoost Regression**

---

6. Linear Regression (Baseline Model)

Metric Used: Mean Absolute Percentage Error (MAPE)

Result:

- **MAPE ≈ 20.13%**

Interpretation:
Linear Regression performs similarly to SVR, indicating limited linear explanatory power in the available features.

---

7. Support Vector Regression (SVR)

Best Parameters:

- Kernel: RBF
- C: 1000
- Epsilon: 1

5-Fold Cross-Validation Results

Metric	Value
MAE	~$28,200
RMSE	~$55,000
R^2	~0.00
MAPE	~17–18%

Observation:
SVR failed to learn meaningful patterns ($R^2 \approx 0$).

---

8. Random Forest Regression

Tuned Model Performance

Metric	Value
MAE	~$33,700
RMSE	~$51,500
R^2	**~0.30**
MAPE	~19–21%

Top Features

- LotArea
- HouseAge
- OverallCond
- MSSubClass

As illustrated in **Figure 3**, the Random Forest model captures the general trend between actual and predicted sale prices, with most predictions clustering around the ideal diagonal line. However, the spread of points for higher-priced properties indicates larger residual errors, highlighting the difficulty of modeling luxury homes with the available feature set.

Feature importance results from the Random Forest model are visualized in **Figure 4**. LotArea and HouseAge are the most influential predictors, followed by OverallCond and MSSubClass. The relatively low importance of many categorical features suggests that these variables contribute limited predictive value in the current dataset.

Random Forest was selected as the **final model**.

---

9. Random Forest with Log-Transformed Target

Log transformation of SalePrice was tested to reduce skewness.

Result:

- No significant improvement over standard Random Forest.

---

10. CatBoost Regression

Performance:

Metric	Value
MAE	~$35,200
RMSE	~$52,900
R^2	~0.23
MAPE	~22%

CatBoost did not outperform Random Forest.

---

11. Model Performance Comparison Table

Model	R^2	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%

---

12. Final Model Selection

Final Model: Random Forest Regressor

Reasoning:

- Highest R^2 score
- Meaningful feature importance
- Robust to nonlinearity and noise

---

13. Results & Interpretation

- Model explains ~30% of variance in house prices
- Predictions are ~20% off on average
- Missing critical features (living area, neighborhood, room count) limit performance

---

14. Conclusion

This project demonstrates a complete machine learning workflow from data preprocessing to model selection. While advanced models were applied, results confirm that **data**

quality and feature richness are more important than model complexity. The Random Forest model provided the best balance of performance and interpretability.

---

15. Future Work

- Incorporate neighborhood and geospatial data
- Add living area and room counts
- Try Gradient Boosting / XGBoost
- Perform residual analysis
- Deploy as a web application

---

16. References

1. Scikit-learn Documentation
2. CatBoost Documentation