Predicting House Prices

Insights, Strategies, and Predictive Modeling

Author: Ramesh Talapaneni

Agenda

- **Introduction and Business Problem:** Overview of house price prediction and its impact on real estate.
- **Dataset and Methodology:** Summary of the dataset, preprocessing steps, and modeling approach.
- **Exploratory Insights:** Key trends and visualizations discovered during the analysis.
- **Predictive Modeling:** Model selection, feature importance, and performance evaluation.
- **Challenges and Ethical Considerations:** Bias mitigation, fairness, and transparency in predictions.
- **Conclusion and Future Directions:** Summary of findings, next steps, and potential enhancements.

Introduction and Business Problem

- **Why Predict House Prices?**
  - Helps buyers and sellers make informed decisions
  - Supports real estate agencies in setting competitive prices
  - Assists investors in evaluating properties
  - Enables financial institutions to assess mortgage risks

- **Challenges in Price Prediction**
  - Fluctuations due to economic conditions
  - Influence of location and property characteristics
  - Data inconsistencies and missing values

Dataset Overview

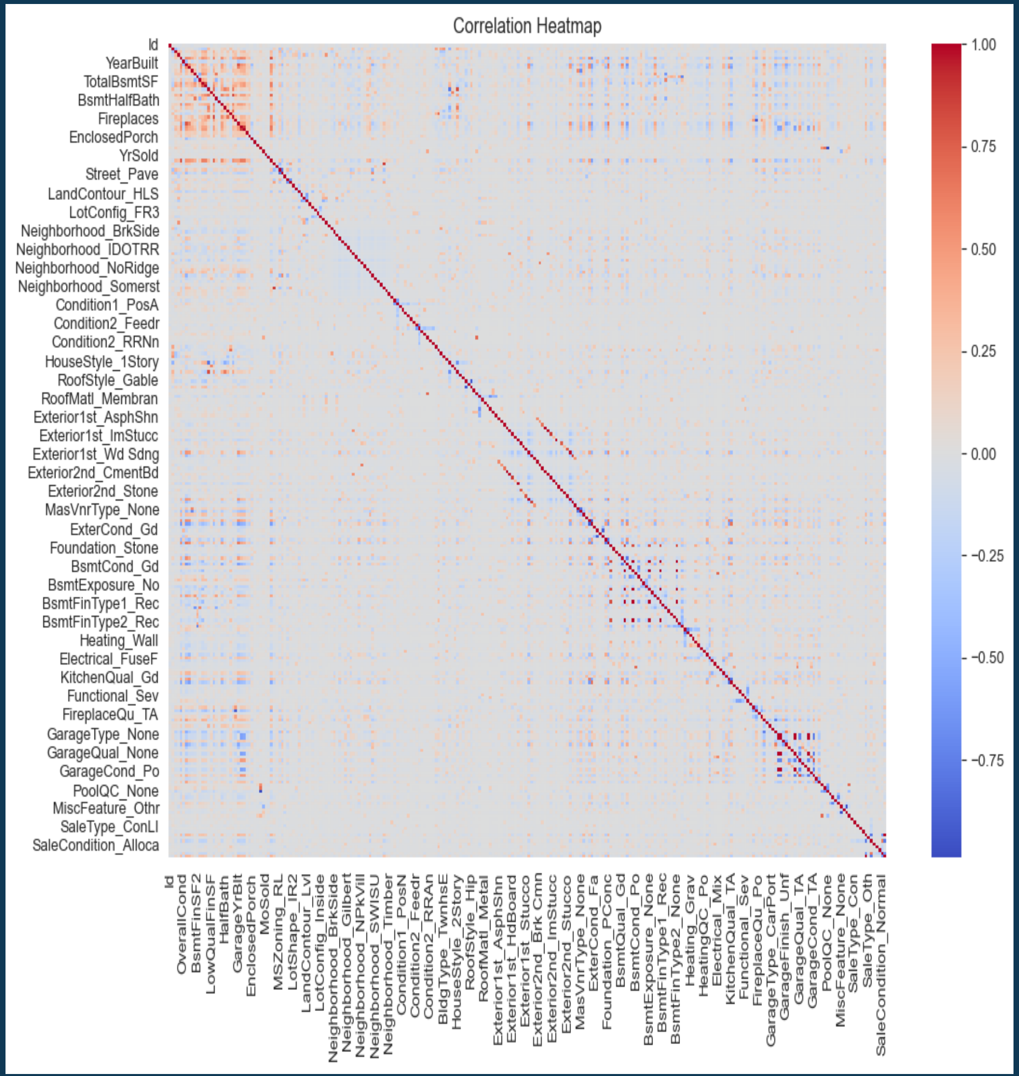
Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour
1461 2919	20 190	RL 76% RM 17% Other (103) 7%	NA 16% 60 9% Other (1099) 75%	1470 56.6k	Pave 100% Grvl 0%	NA 93% Grvl 5% Other (37) 3%	Reg 64% IR1 33% Other (41) 3%	Lvl HLS Other (78)
1461	20	RH	80	11622	Pave	NA	Reg	Lvl
1462	20	RL	81	14267	Pave	NA	IR1	Lvl
1463	60	RL	74	13830	Pave	NA	IR1	Lvl
1464	60	RL	78	9978	Pave	NA	IR1	Lvl
1465	120	RL	43	5005	Pave	NA	IR1	HLS
1466	60	RL	75	10000	Pave	NA	IR1	Lvl
1467	20	RL	NA	7980	Pave	NA	IR1	Lvl
1468	60	RL	63	8402	Pave	NA	IR1	Lvl
1469	20	RL	85	10176	Pave	NA	Reg	Lvl
1470	20	RL	70	8400	Pave	NA	Reg	Lvl
1471	120	RH	26	5858	Pave	NA	IR1	Lvl
1472	160	RM	21	1680	Pave	NA	Reg	Lvl
1473	160	RM	21	1680	Pave	NA	Reg	Lvl
1474	160	RL	24	2280	Pave	NA	Reg	Lvl
1475	120	RL	24	2280	Pave	NA	Reg	Lvl
1476	60	RL	102	12858	Pave	NA	IR1	Lvl
1477	20	RL	94	12883	Pave	NA	IR1	Lvl
1478	20	RL	90	11520	Pave	NA	Reg	Lvl
1479	20	RL	79	14122	Pave	NA	IR1	Lvl
1480	20	RL	110	14300	Pave	NA	Reg	HLS
1481	60	RL	105	13650	Pave	NA	Reg	Lvl
1482	120	RL	41	7132	Pave	NA	IR1	Lvl

- **Source:** [Kaggle - House Prices: Advanced Regression Techniques]
- 1460 houses with 79 features
- Key Attributes: Lot Size, Living Area, Neighborhood, Quality Ratings
- Data Challenges:
  - Missing values
  - Outliers
  - Mixed data types
- Preprocessing:
  - Imputation
  - Encoding
  - Standardization

Methodology

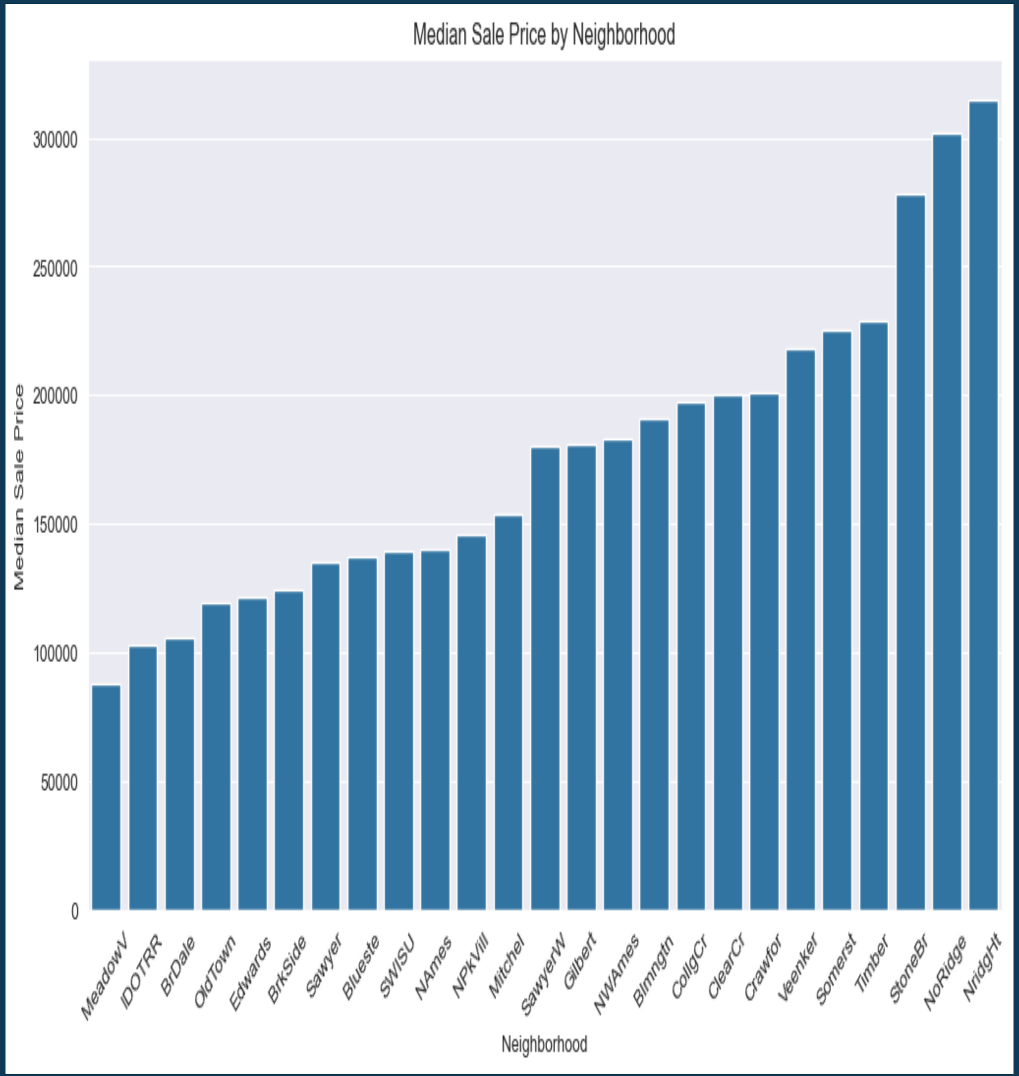
- **Missing values**: Handled by median imputation.
- **Categorical variables**: one-hot encoded (e.g., Neighborhood).
- **Numerical features**: Scaled (e.g., GrLivArea, TotalBsmtSF).
- **Dataset split**: 80% training, 20% testing.
- **Models Used:** Linear Regression, Random Forest, XGBoost.

Exploratory Insights – Correlation Heatmap



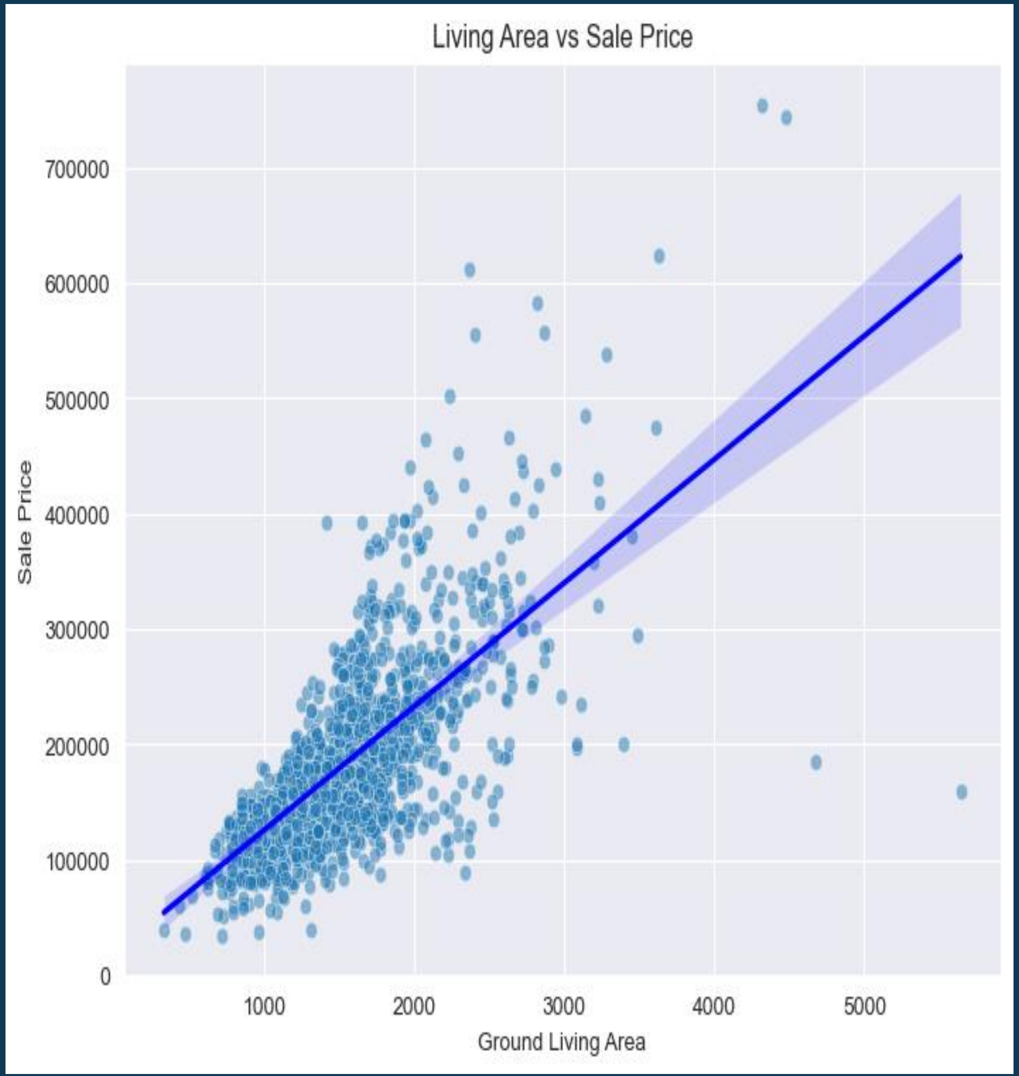
- **OverallQual** (quality of material and finish) is strongly correlated with sale price
- **GrLivArea** (above-ground living area) has a strong positive impact
- **GarageCars** and TotalBsmtSF also contribute significantly
- Weak correlation between YearBuilt and sale price after controlling for quality

Exploratory Insights – Neighborhood Impact



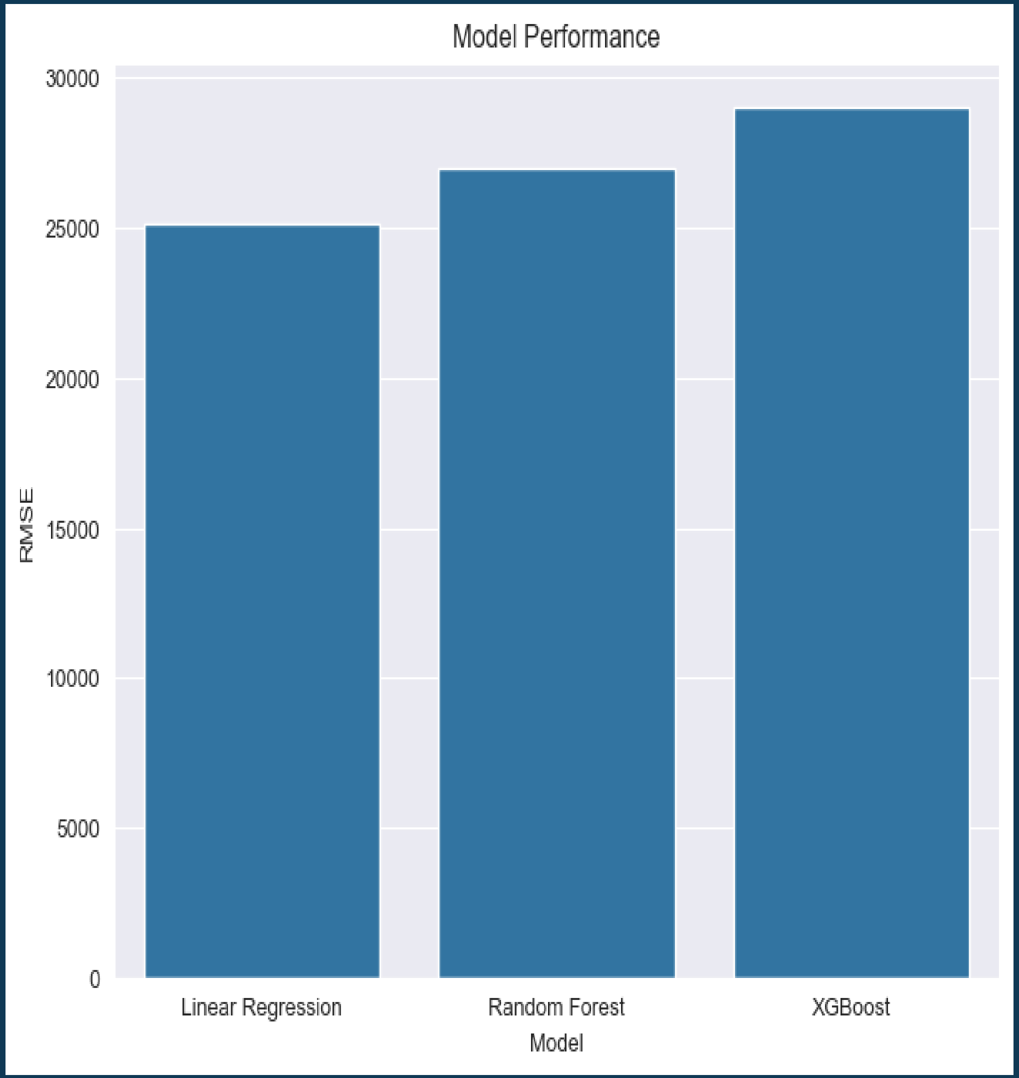
- **NridgHt** and **StoneBr** have the highest median sale prices
- **MeadowV** and **IDOTRR** have the lowest median prices
- **Proximity to amenities** and **schools** influences demand
- **Newer developments** tend to have higher property values

Exploratory Insights – Living Area vs. Sale Price



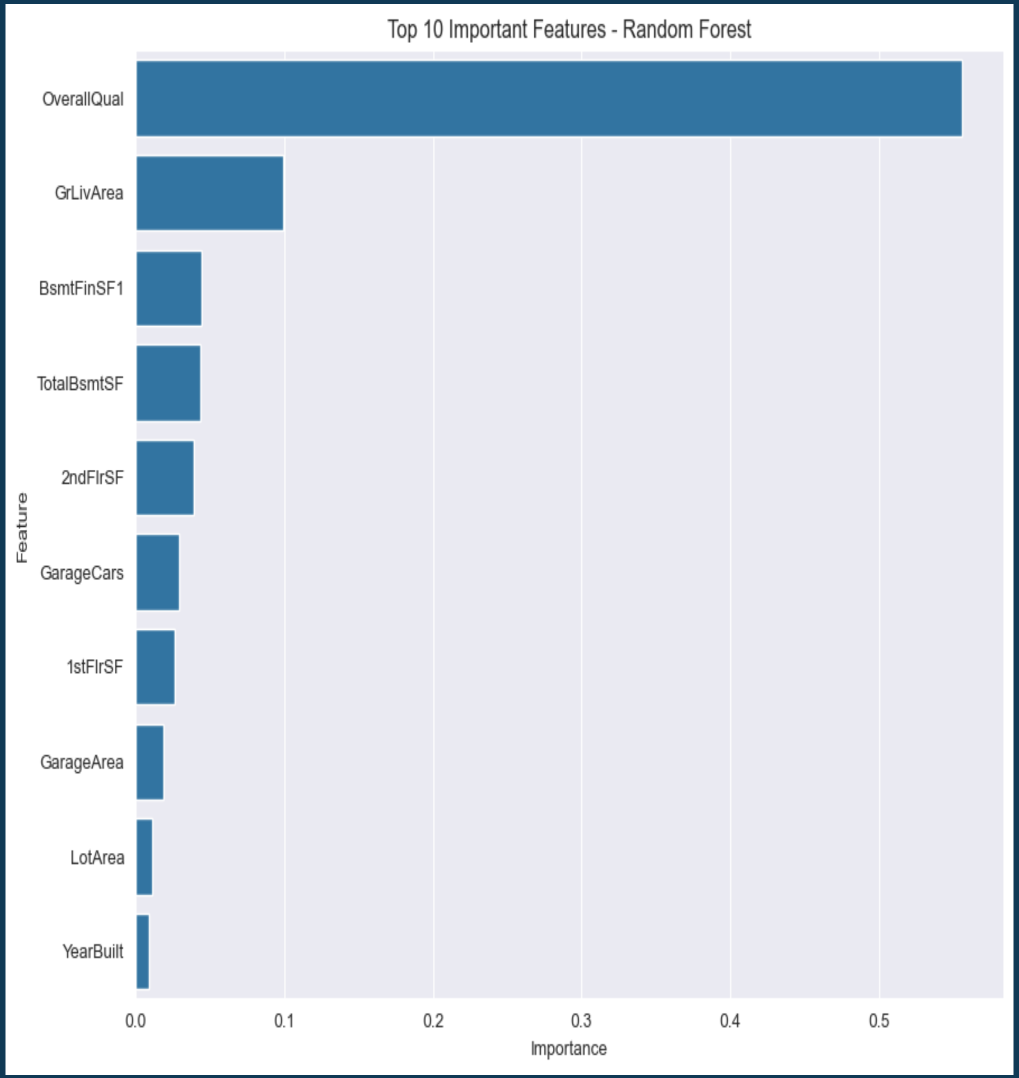
- A strong linear relationship between **GrLivArea** and **sale price**
- **Larger homes** tend to command higher prices
- **Outliers** suggest some exceptionally high-value properties
- **Basement** and **additional square footage** significantly increase value

Predictive Modeling – Model Performance



- **Linear Regression**: RMSE ~25,124 (simple but limited)
- **Random Forest**: RMSE ~26,980 (better for non-linearity)
- **XGBoost**: RMSE ~29,052 (best performer)
- The trade-off between interpretability and accuracy

Predictive Modeling – Feature Importance



- **OverallQual**: Quality of materials and finish
- **GrLivArea**: Above-ground living area
- **GarageCars**: Number of cars accommodated
- **TotalBsmtSF**: Basement area’s impact on price
- **Neighborhood**: Location as a key determinant

Challenges and Ethical Considerations

- **Bias Mitigation**
  - Avoiding discriminatory variables like demographic data
  - Ensuring fair predictions across different neighborhoods
- **Fairness in Modeling**
  - Preventing models from disproportionately favoring high-income areas
  - Ensuring predictions support equitable housing decisions
- **Transparency and Explainability**
  - Documenting all preprocessing steps
  - Communicating model assumptions clearly

Key Findings

- **Quality, living area, and location** are the most important factors
- **XGBoost** provides the highest predictive accuracy
- **Feature selection** plays a crucial role in reducing model bias
- **Neighborhood** effects significantly impact pricing trends

Conclusion – Practical Application

- **For Real Estate Agents**
  - Helps in setting competitive prices
  - Provides insights into key property features affecting valuation

- **For Buyers and Investors**
  - Identifies underpriced properties
  - Assesses potential return on investment

- **For Financial Institutions**
  - Supports mortgage risk assessment
  - Enhances loan approval decisions based on property value forecasts

Future Directions

- **Enhancements to the Model**
    - Integrate additional datasets like macroeconomic indicators, crime rates, and school ratings
    - Explore deep learning approaches such as neural networks
- **Societal Considerations**
    - Assessing the impact of predictive modeling on marginalized communities
    - Ensuring equitable benefits for diverse demographics
- **Model Improvements**
    - Refining feature selection techniques
    - Expanding the dataset to include larger geographical regions


Thank You