House Price Prediction using Machine Learning

Advanced Apex Project
Data Disruptors
Dr. Naga Janapati

The Challenge

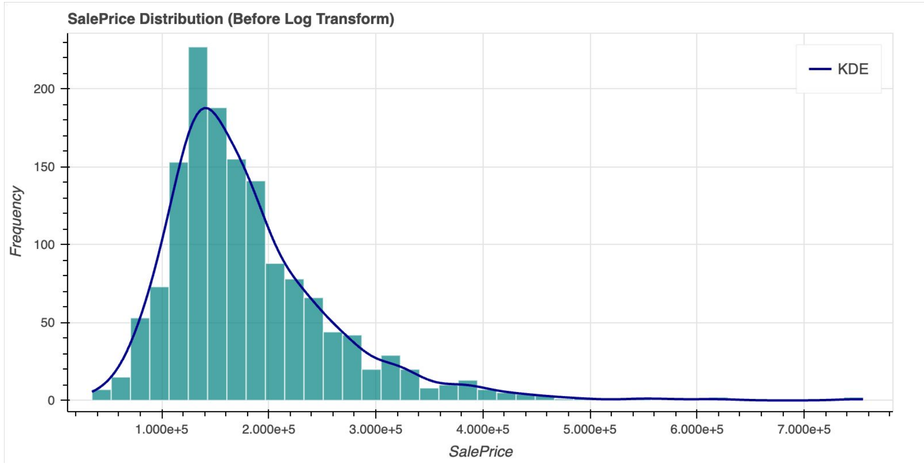
Every house sale raises the same question: **"Is this the right price?"**

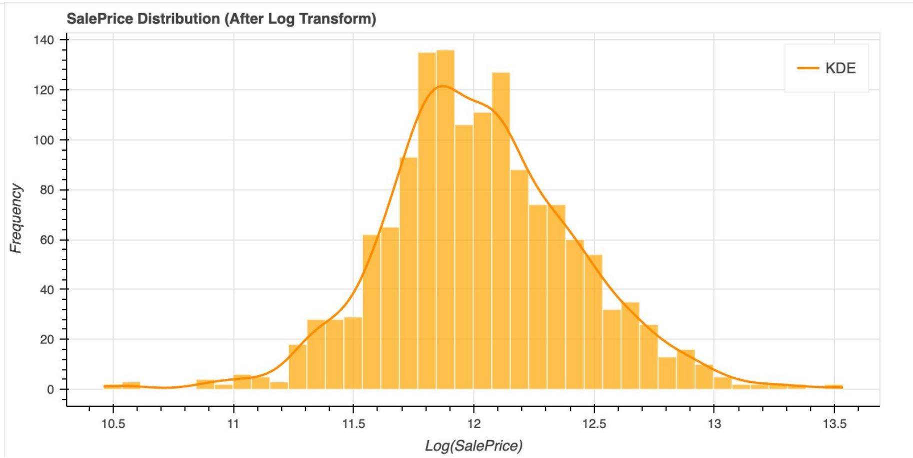
The Problem:

- ❖ Buyers want fair deals
- ❖ Sellers want accurate valuations
- ❖ Real estate agents need quick estimates
- ❖ Banks need reliable appraisals

Our Mission: Build an AI model that predicts house prices with 90%+ accuracy using property features.

Why we need prediction?





The before/after comparison visually demonstrates that log transformation converts the right-skewed price distribution into a normal distribution, which is essential for accurate regression modeling.

The Data

1,460 Houses, 81 Features, One Goal

Dataset: Kaggle House Prices Competition (Ames, Iowa)

What We Have:

- 1,460 real property sales
- 81 features: size, quality, location, age, amenities
- Target: SalePrice (ranging from $34,900 to $755,000)

The Challenge:

- Missing values in 19 columns
- Outliers in 32 numeric features
- Mix of categorical and numerical data
- - Skewed price distribution

Shape of dataset: (1460, 81)

Data types of each column:

Id	int64
MSSubClass	int64
MSZoning	object
LotFrontage	float64
LotArea	int64
...	
MoSold	int64
YrSold	int64
SaleType	object
SaleCondition	object
SalePrice	int64

Length: 81, dtype: object

The Discovery Journey

What Drives House Prices?

Key Insights from EDA:

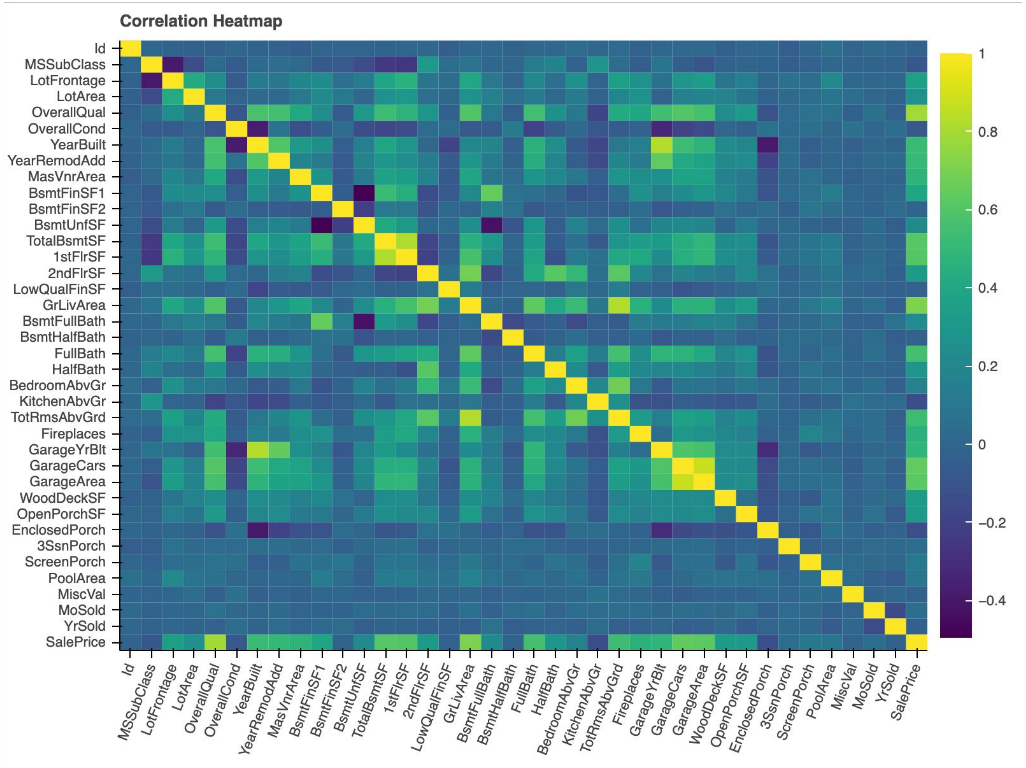
- Size Matters Most
- Total Square Footage is the #1 predictor (68% mutual information score)
- Living area, basement, and floors all correlate strongly

2. Quality Over Location

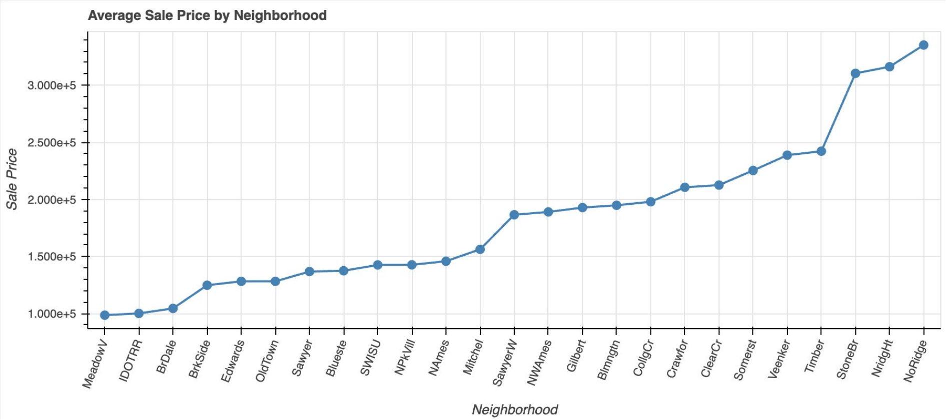
- Overall Quality (58% MI score) beats neighborhood
- Quality Score (quality × condition) is a powerful predictor

3. Price Patterns

- Average price varies 3.4x across neighborhoods ($98K to $335K)
- Quality rating (1-10) shows clear price progression
- Log transformation needed (skewness: 1.88)



The correlation heatmap reveals feature relationships and identifies the strongest price predictors (like OverallQual and GrLivArea) while detecting redundant features.



The three average sales graphs reveal how location (neighborhood), quality rating, and building type each impact house prices, showing that quality has the strongest linear relationship while neighborhoods show the widest price variation.

The Transformation

From Raw Data to Predictions

Our Data Pipeline:

1. Cleaning → Removed 5 columns with >50% missing data, imputed strategically, capped outliers (1st-99th percentile)

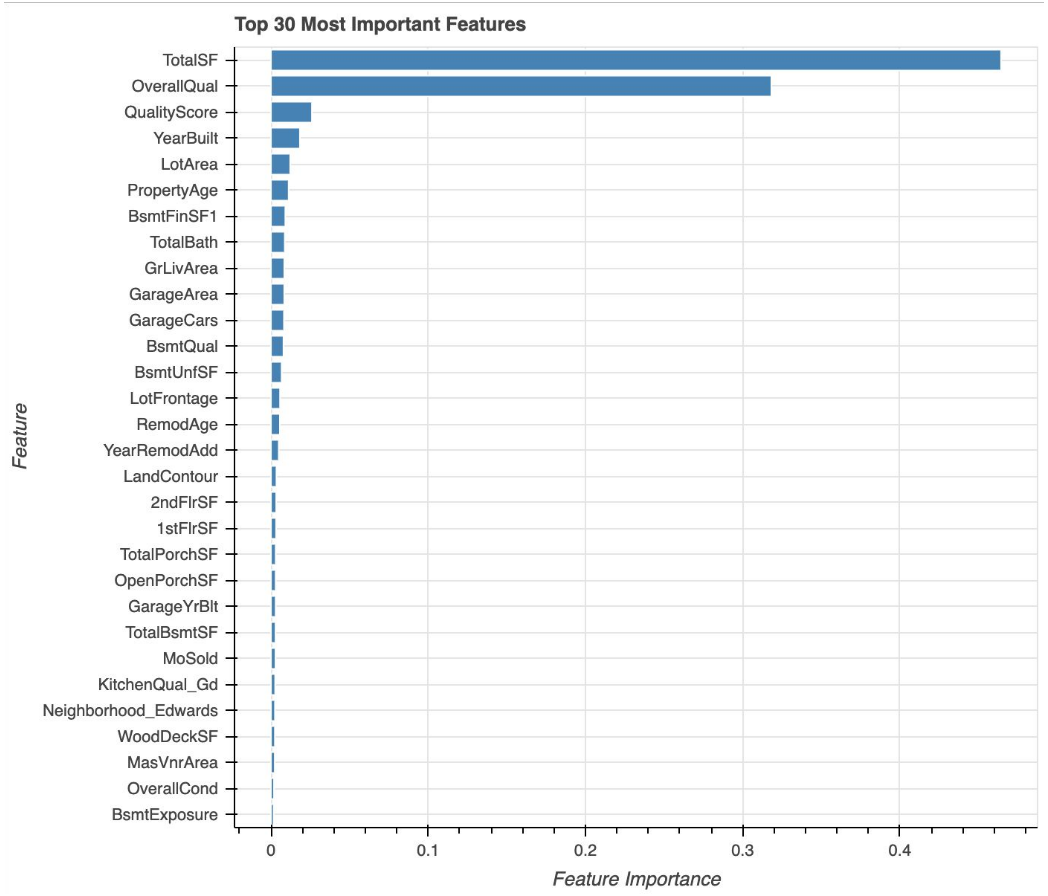
2. Feature Engineering → Created 10 powerful new features:

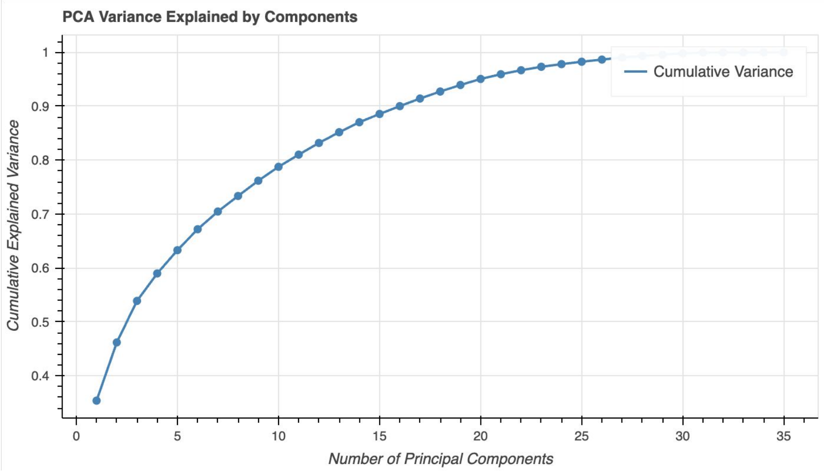
- TotalSF (total square footage)
- PropertyAge, RemodAge
- QualityScore (quality × condition)
- TotalBath, TotalPorchSF
- Binary flags (HasGarage, HasBasement, etc.)

3. Feature Selection → Multi-method approach:

- Started with 81 features
- Filtered to 35 using Mutual Information, F-test, Random Forest
- Refined with RFECV (24 features) and Lasso (23 features)
- Final: 35 best features selected

4. Dimensionality Reduction → PCA reduced to 20 components (95% variance retained)





PCA

Number of components to retain **95%**

Variance: **20**

Reduced shape: (1460, 20)

The Solution

Four Models, One Winner

Models Tested:

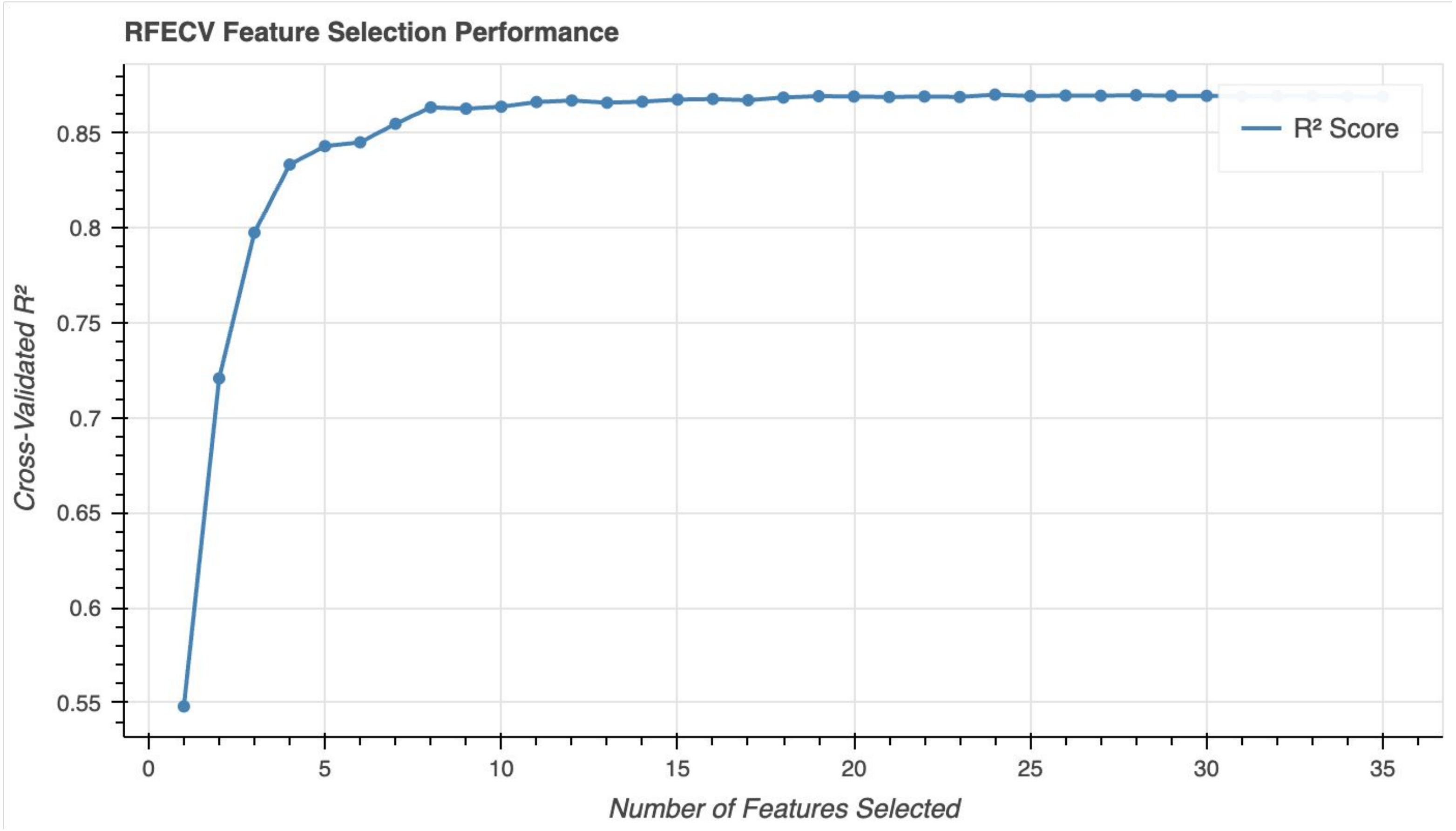
- Gradient Boosting ⭐ (Best)
- LightGBM
- Linear Regression
- Tuned SVR

Why Gradient Boosting Won:

- Handles non-linear relationships
- Captures feature interactions
- Robust to outliers
- Best performance on PCA-reduced features

Training Strategy:

- Log-transformed target (normalized distribution)
- 80/20 train-test split
- PCA-reduced features (20 components)
- Hyperparameter tuning for SVR



RFECV Feature Selection Performance
R² Score
Cross-Validated R²
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0
5
10
15
20
25
30
35
Number of Features Selected

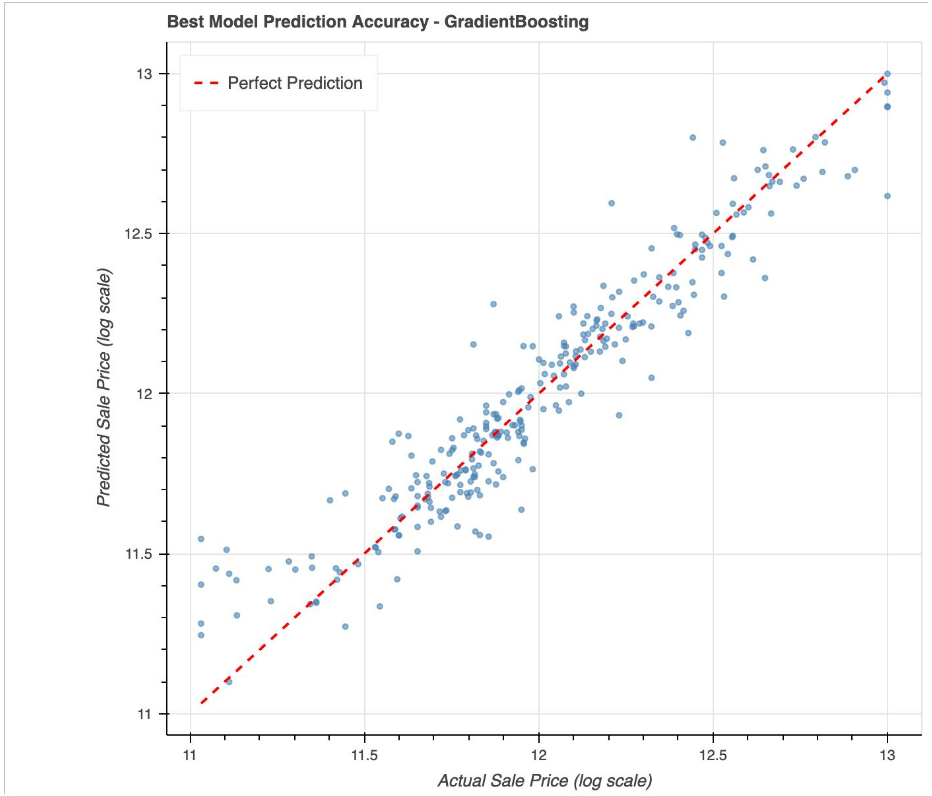
The Results

90.4% Accuracy Achieved

Model	R^2 Score	RMSE (log)	Performance
Gradient Boosting	0.9043	0.1267	Best
LightGBM	0.9001	0.1295	Excellent
Linear Regression	0.8942	0.1333	Good
Tuned SVR	0.8801	0.1419	Good

What This Means:

- Model explains 90.4% of price variance
- Predictions are within 12.7% error (log scale)
- Top 3 models all exceed 89% accuracy
- Robust and reliable predictions



R^2 Chart: Gradient Boosting leads with 90.4% R^2, outperforming all other models.

Scatter Plot: The tight clustering of points around the diagonal line confirms the model accurately predicts house prices across all price ranges.

The Insights

What We Learned About House Prices

Top 5 Price Drivers:

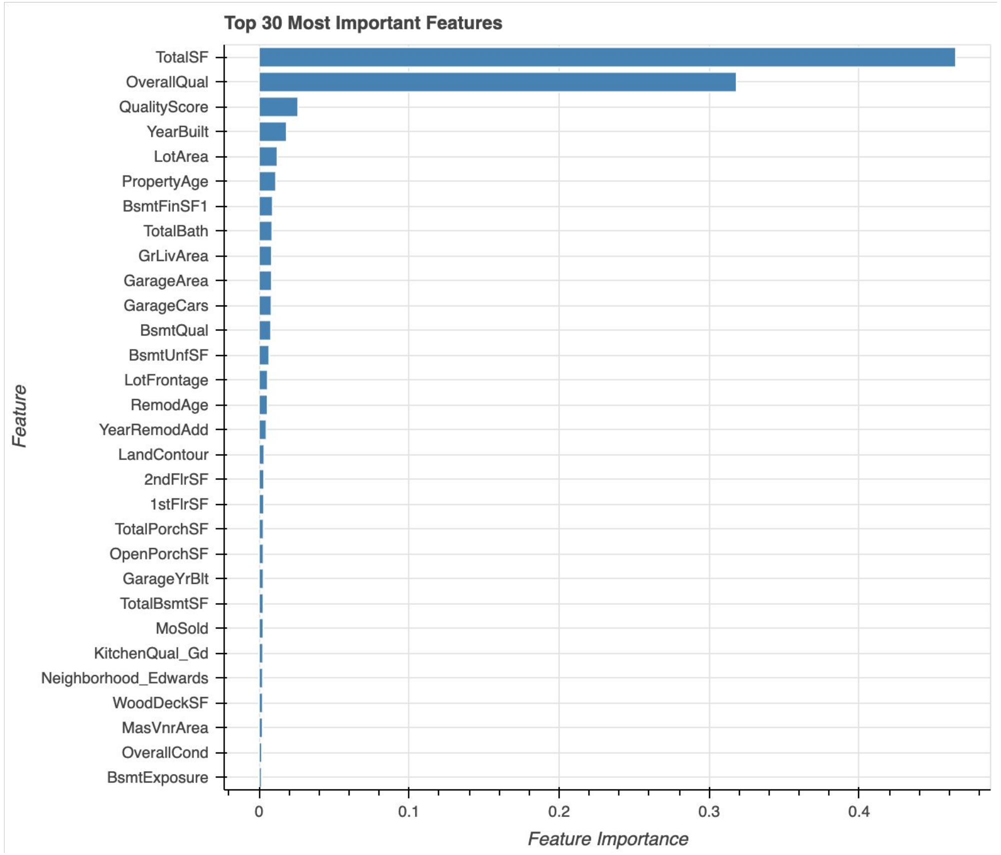
1. Total Square Footage (46.4% importance)
2. Overall Quality (31.8% importance)
3. Quality Score (2.6% importance)
4. Year Built (1.8% importance)
5. Lot Area (1.2% importance)

Business Recommendations:

❖ **Size investments** yield highest ROI
❖ **Quality improvements** significantly boost value
❖ **Age matters** but can be offset by quality
❖ **Location** matters less than quality and size

15 Features consistently selected across all methods:

TotalSF, OverallQual, GrLivArea, GarageArea, TotalBath, and 10 more



Total unique selected features: **35**

Total features after PCA: **20**

Highly consistent across all methods: **15**

Common features across ALL methods:

```
['BsmtFinSF1', 'BsmtQual', 'Fireplaces',
'GarageArea', 'GarageCars', 'GrLivArea',
'KitchenQual_TA', 'LotArea', 'MSSubClass',
'OverallQual', 'PropertyAge',
'QualityScore', 'RemodAge', 'TotalBath',
'TotalBsmtSF']
```

The Impact

From Data to Decisions

What We Built:

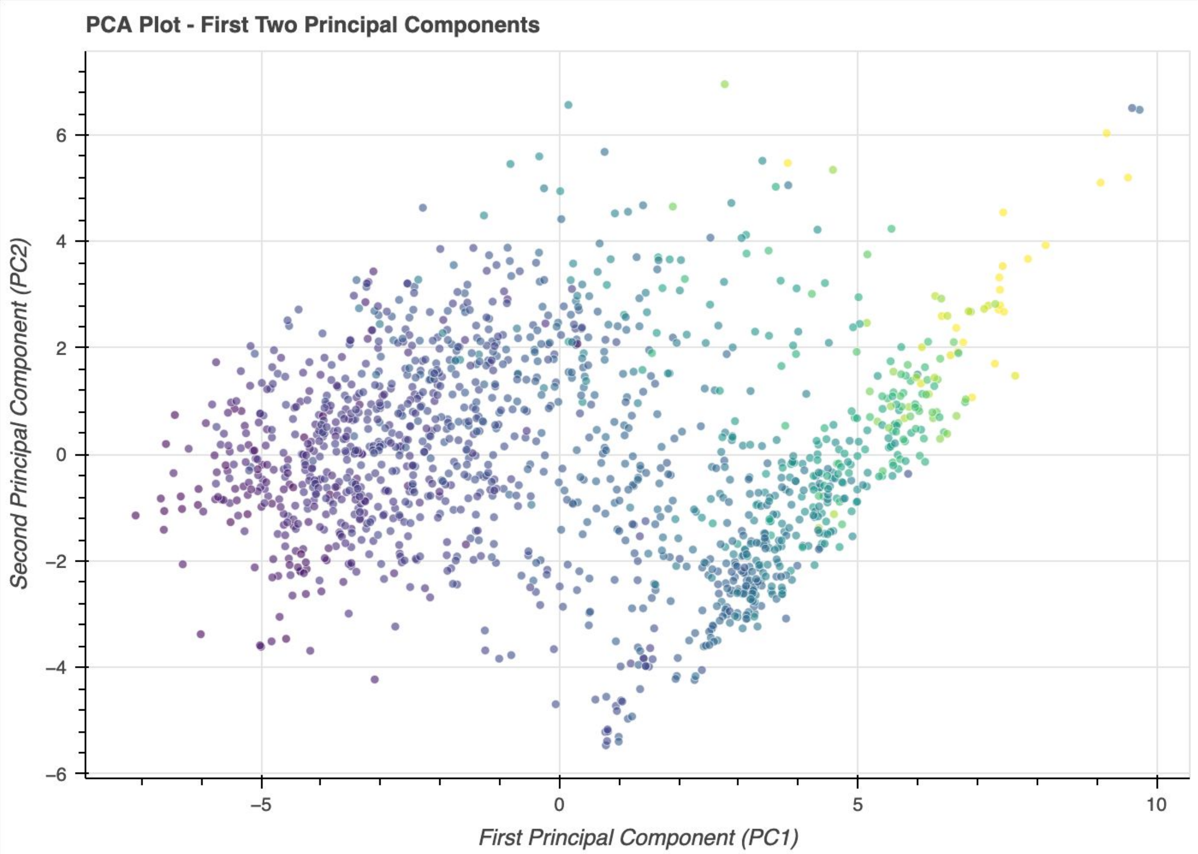
- Complete ML pipeline (data → predictions)
- 90.4% accurate price prediction model
- Multi-method feature selection approach
- Production-ready model (saved as .pkl)

Real-World Applications:

- Real Estate: Instant property valuations
- Banking: Automated loan assessments
- Investors: Identify undervalued properties
- Market Analysis: Understand price drivers

Key Achievement:

Transformed 81 messy features **into 35 powerful predictors**, achieving **90.4% accuracy** with Gradient Boosting.


PCA Plot - First Two Principal Components
Second Principal Component (PC2)
First Principal Component (PC1)
6
4
2
0
−2
−4
−6
−5
0
5
10

Validation Results Export (house_price_validation_results.csv)

Detailed Performance Analysis

What We Exported:

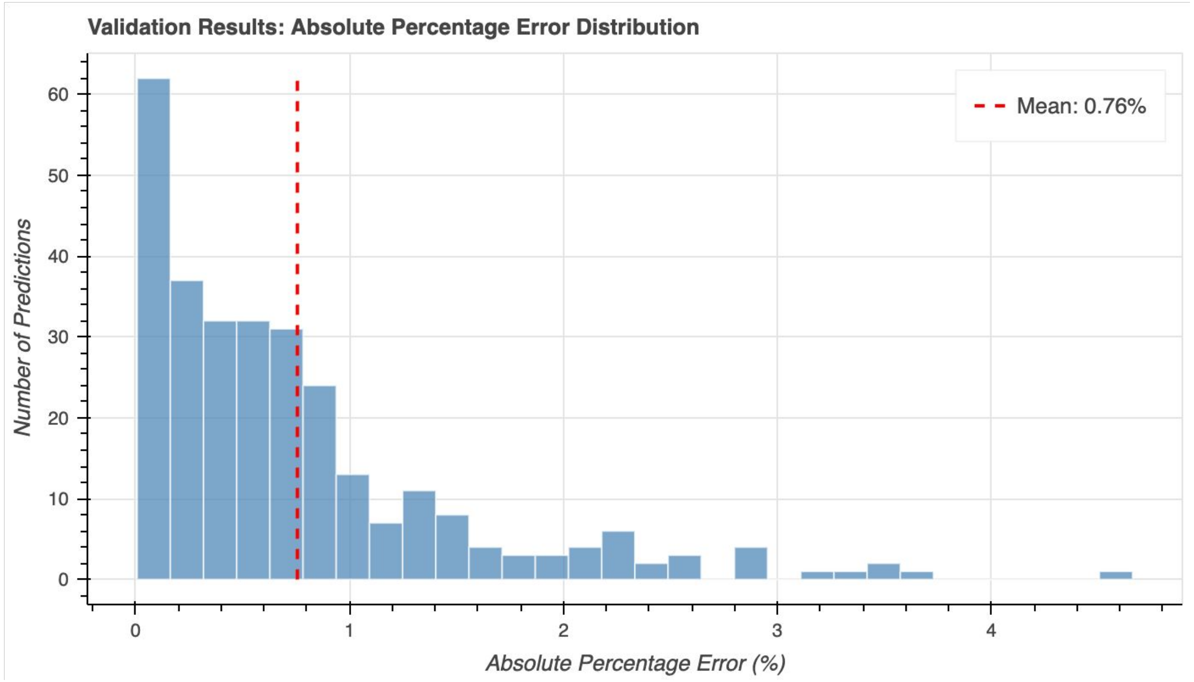
- ❖ 292 validation samples analyzed in detail
- ❖ Per-sample predictions with comprehensive error metrics
- ❖ Complete validation results saved to CSV for further analysis

Key Metrics in Export:

- ❖ Actual vs Predicted prices (log scale)
- ❖ Error distribution (raw and absolute)
- ❖ Percentage errors for interpretability
- ❖ Absolute percentage errors for magnitude assessment

Validation Insights:

- ❖ Model performance validated on 20% hold-out test set
- ❖ Each prediction includes error breakdown
- ❖ Enables detailed analysis of model behavior
- ❖ Supports production deployment validation



Key Findings from Validation:

- ❖ Mean Absolute Percentage Error: **0.76%**: Excellent average accuracy
- ❖ Median Absolute Percentage Error: **0.54%**: Most predictions are highly accurate
- ❖ **95th** Percentile: **2.35% - 95%** of predictions **within 2.35% error**
- ❖ Max Error: **4.66%**: Worst case scenario still **under 5% error**

Conclusion

The Story in Numbers

Project Summary:

- ❖ 460 houses analyzed
- ❖ 81 features ➔ 35 selected ➔ 20 PCA components
- ❖ 4 models tested
- ❖ 90.4% accuracy achieved
- ❖ Gradient Boosting as best model

Thank you!

Team Details

Team Members

1. Shivansh Tiwari
2. Pathaneni Gangotri
3. Sonkar Vedant Rajesh Ranjeeta
4. Himanshu Soni

Team Supervisor

- Utkarsh Khare