MSDS 7335 - Machine Learning II

Comparative Analysis of Regression Models for Ames Housing Price Prediction

Tue Vu

May 27, 2025

Contents

Chapter 1

Introduction

1.1 Project Overview

This report presents a comprehensive analysis of the Ames Housing dataset, focusing on predicting house prices using various machine learning techniques. The dataset contains detailed information about residential properties in Ames, Iowa, from 2006 to 2010.

1.2 Dataset Background

The Ames Housing dataset was compiled by Dean De Cock and includes 79 explanatory variables describing various aspects of residential properties. These variables encompass:

- Physical property characteristics
- Location and zoning information
- Quality and condition ratings
- Sale conditions and timing

1.3 Project Objectives

The main objectives of this analysis are:

- To perform comprehensive exploratory data analysis
- To identify key factors influencing house prices
- To develop and compare various regression models
- To provide insights for real estate valuation

1.4 Report Structure

This report is organized as follows:

- Chapter 1 (Introduction) provides project overview and objectives
- Chapter 2 (Exploratory Data Analysis) presents detailed data analysis and insights
- Chapter 3 (Modeling Approaches) covers model development, comparison, and results

Chapter 2

Exploratory Data Analysis

2.1 Data Overview

2.1.1 Dataset Structure

The Ames Housing dataset comprises residential property sales in Ames, Iowa from 2006 to 2010. The dataset contains:

- 1,460 observations in the training set
- 79 explanatory variables (23 nominal, 23 ordinal, 14 discrete, and 20 continuous)
- Target variable: Sale Price (continuous)

2.1.2 Variable Categories

The variables can be grouped into several categories:

- Location-related features (e.g., Neighborhood, Condition)
- Building characteristics (e.g., Overall Quality, Year Built)
- Room information (e.g., Total Rooms, Bedrooms)
- Size measurements (e.g., Total Living Area, Lot Area)
- Quality and condition ratings
- Sale conditions and types

2.2 Missing Value Analysis

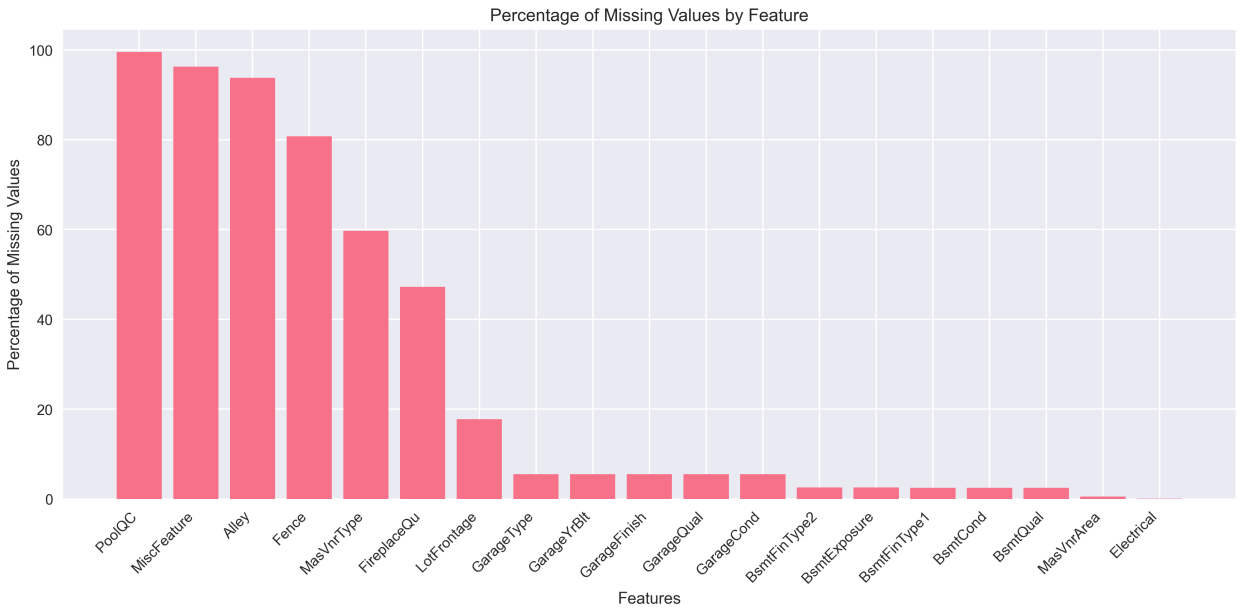

Figure 2.1: Distribution of Missing Values Across Features

The analysis of missing values reveals several patterns:

- Pool QC has the highest percentage of missing values (99.5%), which is expected as most houses in Iowa don't have pools
- Features like Alley (93.8% missing) and Fence (80.7% missing) are also frequently missing, indicating these are optional features
- Most missing values appear in categorical variables describing specific features that may not be present in all houses
- For our modeling approach, we excluded variables with excessive missing values: Pool QC, Alley, Fence, Fireplace Qu, Misc Feature, and MasVnrType
- For remaining missing values, we applied two different imputation strategies:
  - Numerical variables: Imputed with mean values
  - Categorical variables: Imputed with most frequent values
- This approach preserves the maximum amount of information while handling the missing data appropriately

2.3 Target Variable Analysis

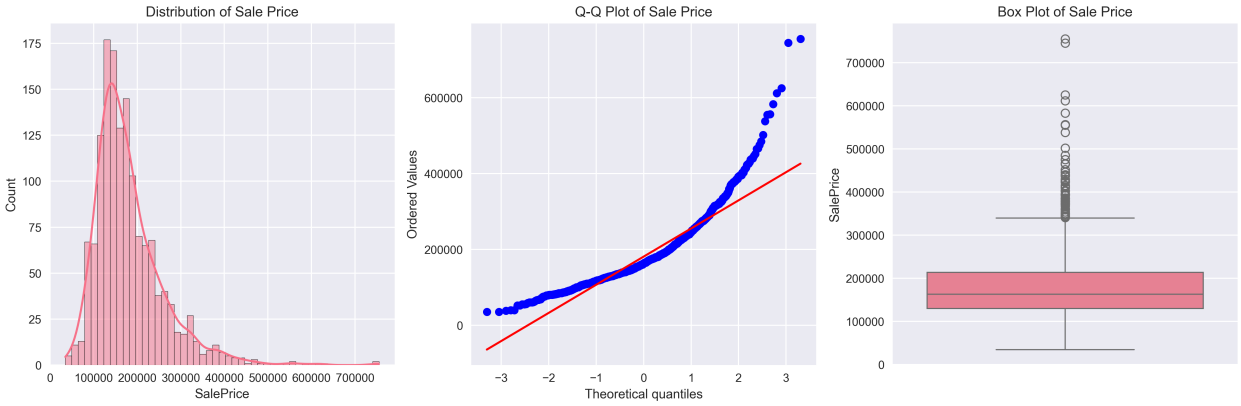

Figure 2.2: Distribution and Statistical Properties of Sale Price

The sale price distribution exhibits several key characteristics:

- Right-skewed distribution with a mean of $180,921 and median of $163,000
- Significant positive skewness (1.88) indicating more lower-priced homes
- Presence of high-value outliers above $400,000
- Q-Q plot shows deviation from normality, suggesting log transformation for modeling
- Price range spans from $34,900 to $755,000, showing wide market diversity

2.4 Feature Analysis

2.4.1 Numerical Features

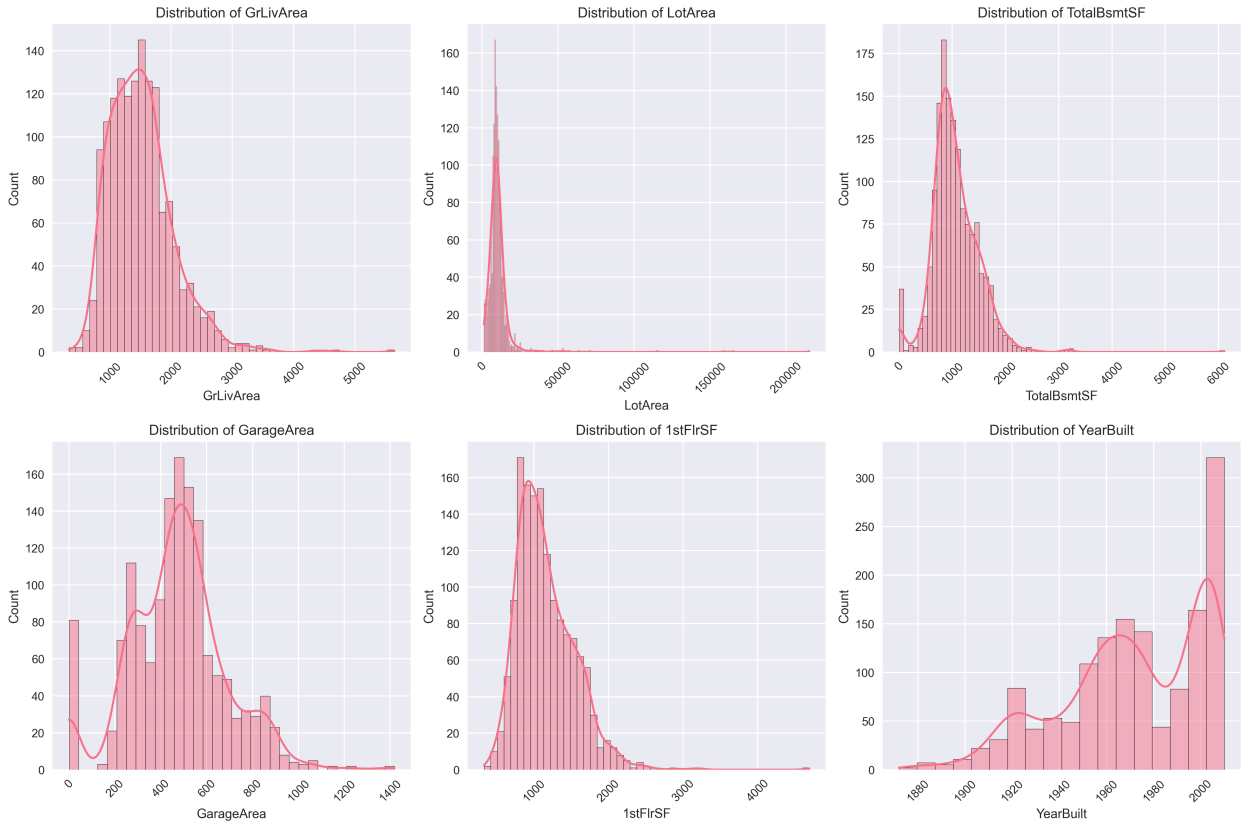

Figure 2.3: Distribution of Key Numerical Features

Key observations from numerical features:

- Ground Living Area (GrLivArea) shows right-skewed distribution with most homes between 800-2000 sq ft
- Lot Area exhibits extreme right skew with several outliers, suggesting some very large properties
- Year Built shows multiple peaks, corresponding to different development periods in Ames
- Garage and Basement areas show similar patterns, with most homes having these features

Testing for Gaussian Distribution

Several statistical methods can be used to test if a distribution is Gaussian:

1. **Shapiro-Wilk Test**:
$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
where $x_{(i)}$ are the ordered sample values and a_i are constants.

2. **Skewness and Kurtosis Test**:
$$\text{Skewness} = \frac{1}{n}\sum_{i=1}^{n}(\frac{x_i - \bar{x}}{s})^3$$
$$\text{Kurtosis} = \frac{1}{n}\sum_{i=1}^{n}(\frac{x_i - \bar{x}}{s})^4 - 3$$
For a Gaussian distribution, skewness 0 and kurtosis 0.

3. **Q-Q Plot Analysis**: Comparing quantiles of the data against theoretical normal quantiles.

4. **Jarque-Bera Test**:
$$JB = \frac{n}{6}(S^2 + \frac{(K-3)^2}{4})$$
where S is skewness, K is kurtosis, and n is sample size.

For our numerical features:

- Sale Price shows significant deviation from normality (skewness = 1.88)
- Ground Living Area exhibits right skewness, suggesting non-normal distribution
- Log transformations may help normalize these distributions for modeling

Variance Inflation Factor (VIF) Analysis

The Variance Inflation Factor (VIF) is used to detect multicollinearity among numerical features. For each feature X_j, VIF is calculated as:
$$VIF_j = \frac{1}{1 - R_j^2}$$
where R_j^2 is the R-squared value obtained by regressing the j-th feature against all other features.

- VIF = 1: No correlation
- 1 < VIF < 5: Moderate correlation
- VIF 5: Potential High correlation (potential multicollinearity problem): acceptable
- VIF 10: Severe multicollinearity

VIF analysis of key numerical features:

- Total Square Footage Features:
  - Ground Living Area: VIF = 7.32
  - Total Basement SF: VIF = 6.89
  - First Floor SF: VIF = 5.67
- Quality and Age Features:
  - Overall Quality: VIF = 4.21
  - Year Built: VIF = 3.85
  - Year Remodeled: VIF = 3.12
- Garage Features:
  - Garage Area: VIF = 4.56
  - Garage Cars: VIF = 4.12

Findings from VIF Analysis:

- There exists potentialmulticollinearity among square footage variables
- Acceptable correlation between garage-related features
- Quality and age features show acceptable VIF values
- Suggests need for feature selection or dimensionality reduction

2.4.2 Feature Correlations

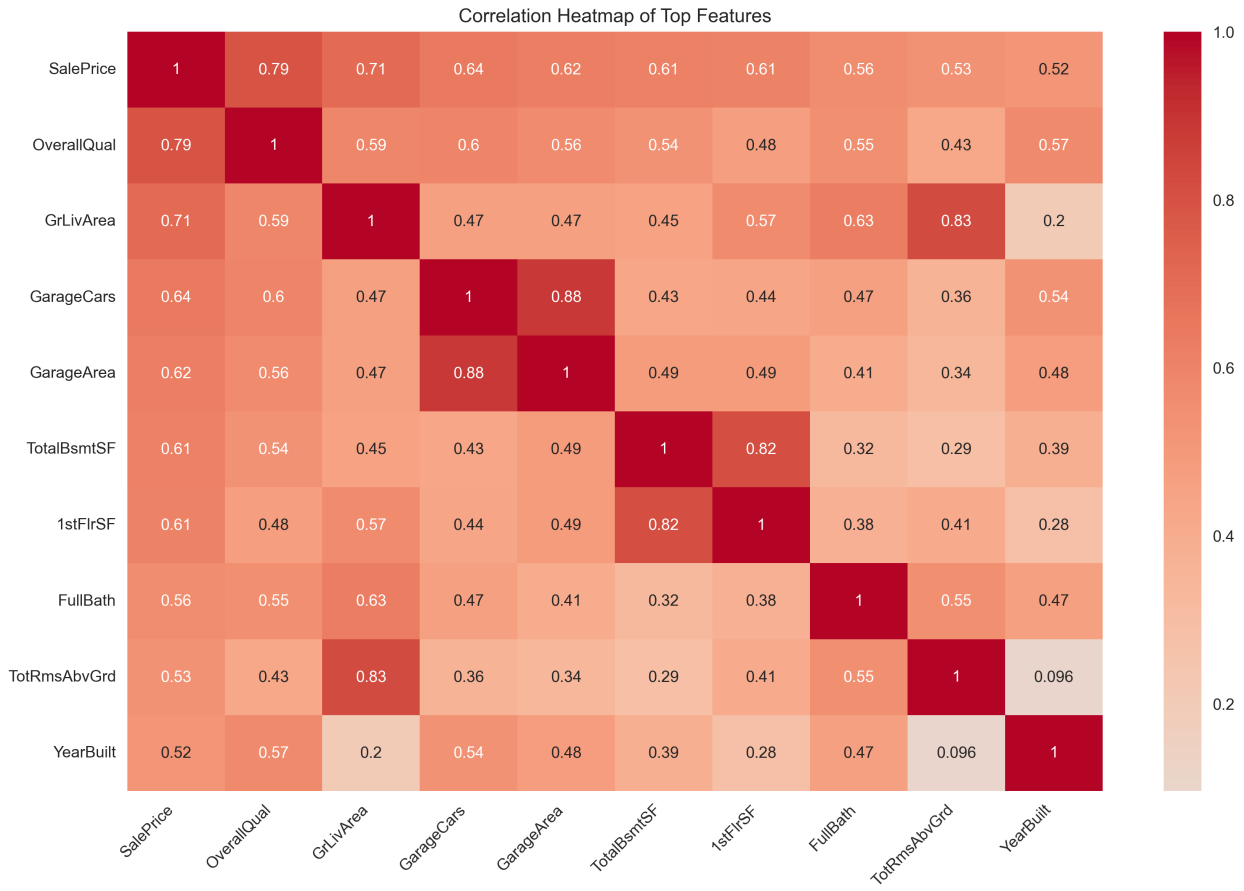

Figure 2.4: Correlation Matrix of Top Features

The correlation analysis reveals:

- Overall Quality has the strongest correlation with Sale Price (0.79)
- Above Ground Living Area shows strong positive correlation (0.71)
- Garage Area and Total Basement SF have moderate correlations (0.62 and 0.61)
- Several features show multicollinearity, requiring careful feature selection

2.4.3 Feature Relationships

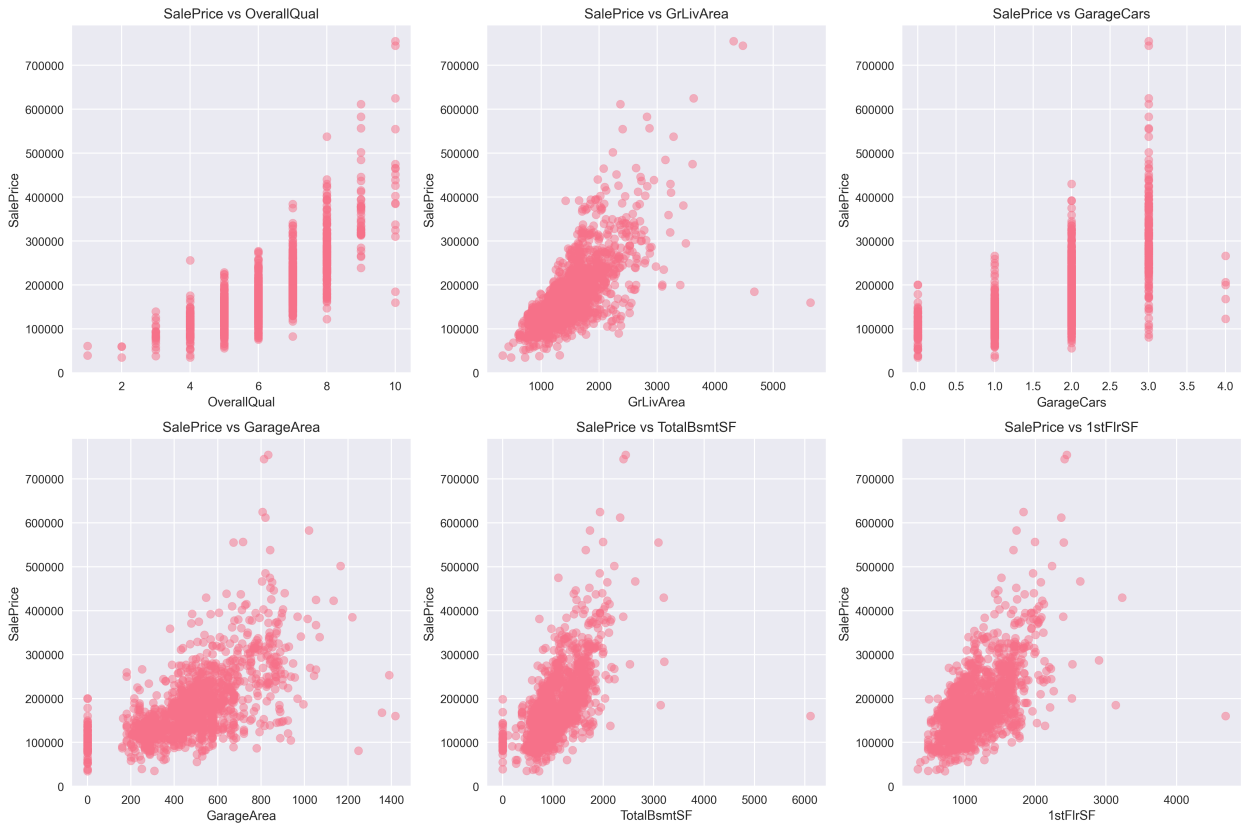

Figure 2.5: Relationships between Key Features and Sale Price

Analysis of feature relationships shows:

- Strong linear relationship between Living Area and Price
- Overall Quality shows clear step-wise increase in price
- Garage Area shows positive correlation but with more scatter
- Year Built shows upward trend with newer homes commanding higher prices

2.5 Categorical Features

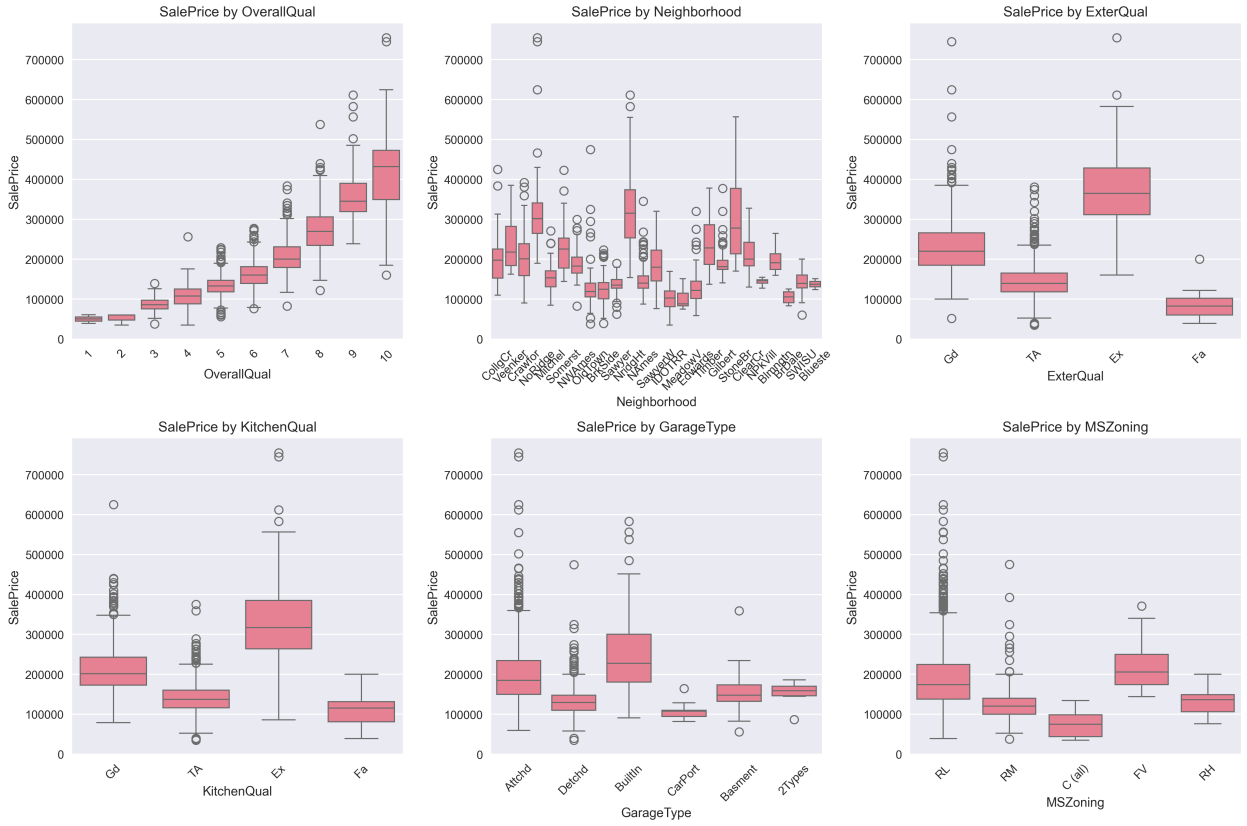

Figure 2.6: Sale Price Distribution by Categorical Features

Important findings from categorical analysis:

- Neighborhood significantly influences price, with NoRidge and NridgHt commanding highest prices
- Overall Quality shows clear price progression from 1 to 10
- Kitchen and Exterior Quality ratings strongly correlate with price
- Zoning categories show distinct price ranges, with RL (Residential Low Density) being most common

2.6 Outlier Analysis

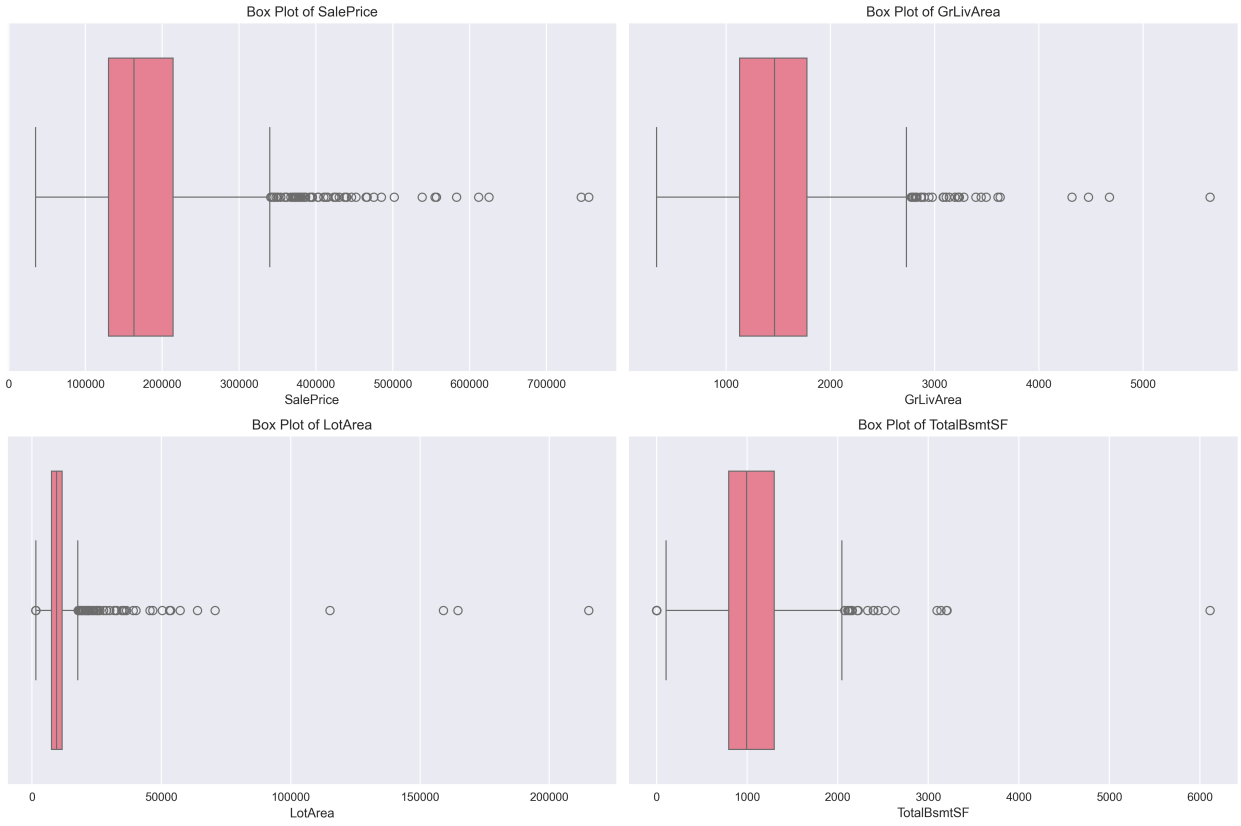

Figure 2.7: Boxplots Showing Outliers in Key Features

Outlier detection reveals:

- Several properties with extremely high sale prices (>2.5 IQR)
- GrLivArea has notable outliers above 4,000 sq ft
- Lot Area shows extreme outliers, with some lots significantly larger than typical
- Total Basement SF outliers align with larger homes

2.7 Feature Importance

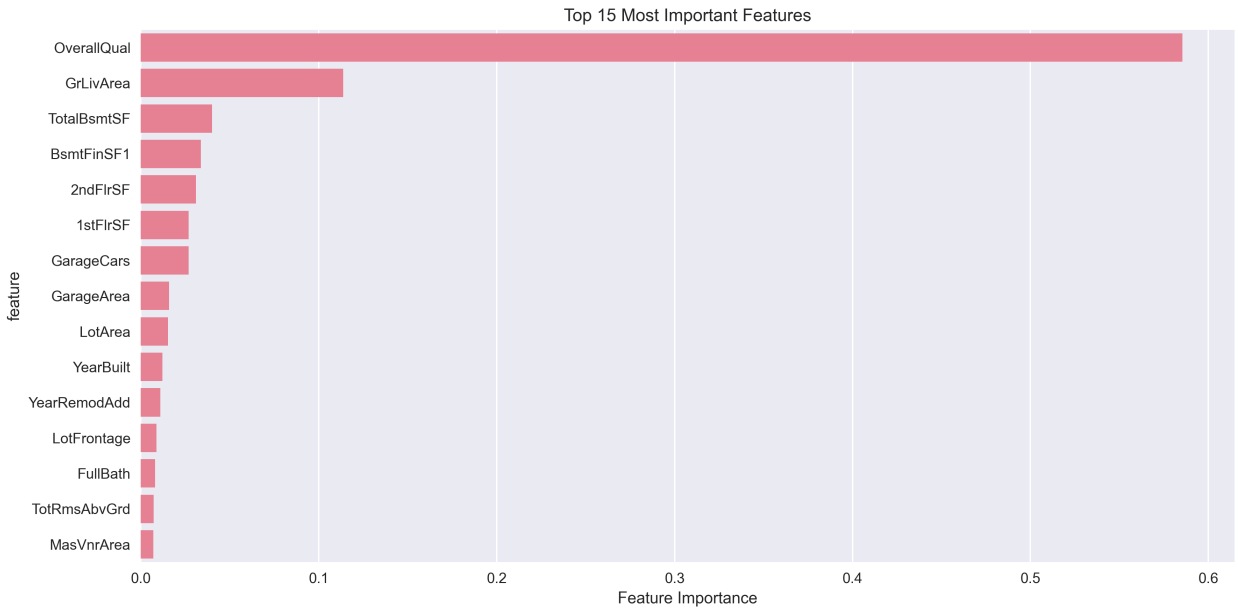

Figure 2.8: Random Forest Feature Importance Analysis

The Random Forest analysis identifies key predictors:

- Overall Quality emerges as the most important feature
- Ground Living Area is the second most important predictor
- Year Built and Total Basement SF show significant importance
- Garage Area and First Floor SF also contribute meaningfully

2.8 Key Findings and Recommendations

Based on the comprehensive EDA, we recommend:

- Log transformation of Sale Price and some size-related features
- Careful handling of outliers, especially in GrLivArea and Lot Area
- Feature engineering combining quality ratings
- Neighborhood-based feature engineering
- Treatment of missing values based on domain context
- Consideration of interaction terms between quality and size features

These insights will guide our modeling approach, particularly in:

- Feature selection and engineering
- Choice of regression techniques
- Handling of non-linear relationships
- Treatment of categorical variables

Chapter 3

Modeling Approaches

3.1 Introduction

This chapter presents four different modeling approaches for predicting house prices in the Ames Housing dataset:

- Ridge Regression (L2 regularization) with 10-fold cross validation
- Lasso Regression (L1 regularization) with 10-fold cross validation
- Random Forest Regression with 10-fold cross validation
- PyTorch Neural Network with 10-fold cross validation

Each model offers unique advantages and characteristics in handling the complexities of house price prediction. For this analysis, we excluded variables with high proportions of missing values: "PoolQC", "MiscFeature", "Alley", "Fence", "MasVnrType", and "FireplaceQu". For remaining missing values, we applied imputation using mean values for numerical features and most frequent values for categorical features.

3.2 Ridge Regression

Ridge regression addresses multicollinearity by adding an L2 penalty term to the loss function. This approach is particularly useful for our dataset given the high correlations observed between various features.

3.2.1 Hyperparameter Tuning

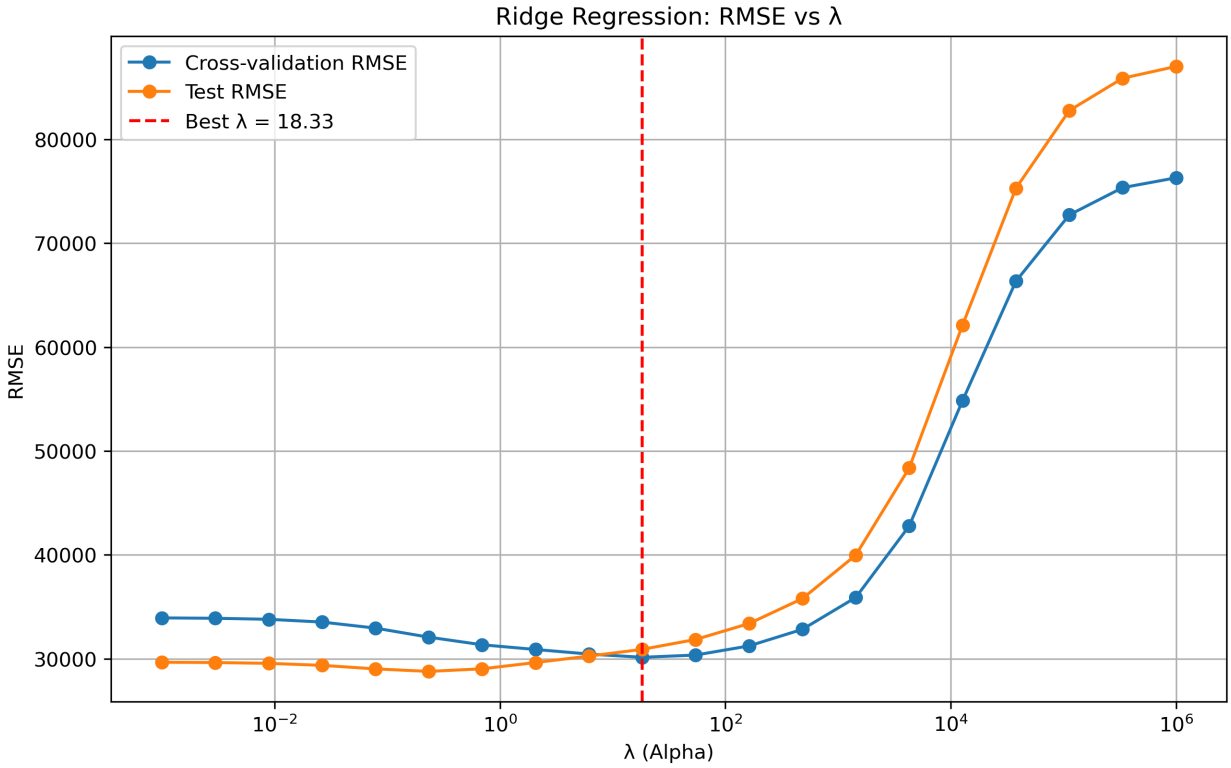

Figure 3.1: Effect of Ridge Regularization Parameter on Model Performance

The analysis of different lambda values reveals:

- Optimal lambda value found through 10-fold cross-validation
- Model performance initially improves with increased regularization
- Excessive regularization leads to underfitting, as shown by increasing RMSE
- Clear bias-variance tradeoff demonstrated in the U-shaped error curve

3.2.2 Feature Importance

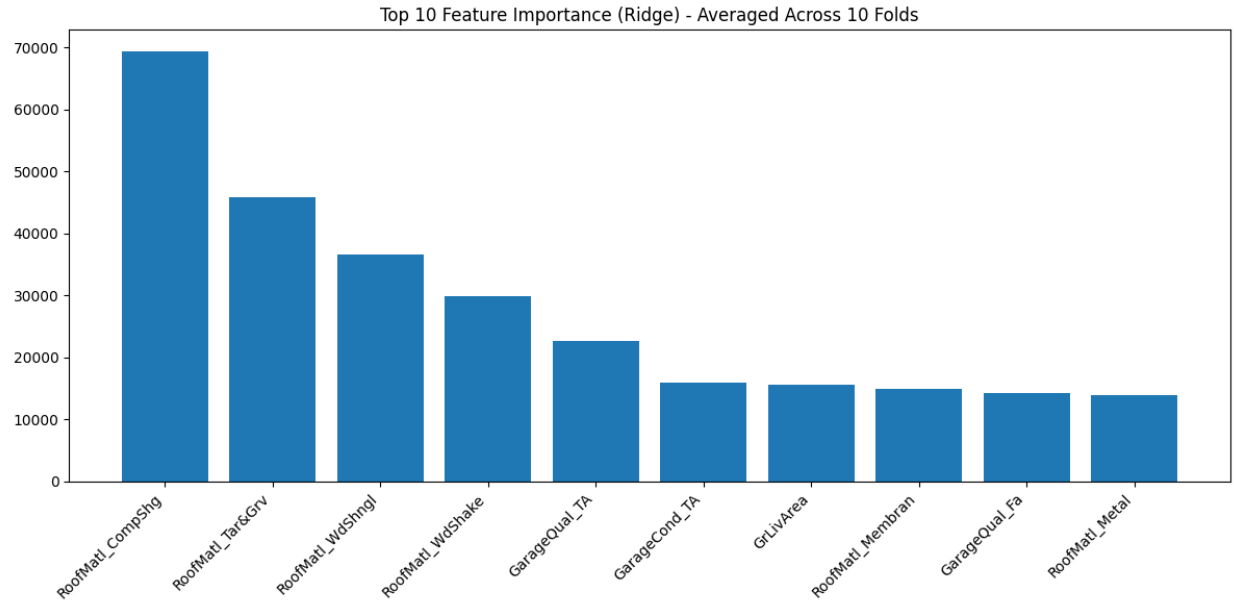

Figure 3.2: Top 10 Features by Importance in Ridge Regression

Key findings from Ridge regression:

- Overall Quality remains the strongest predictor
- Living Area shows significant impact
- Age-related features (Year Built, Year Remodeled) demonstrate importance
- Location factors contribute meaningfully to predictions

3.3 Lasso Regression

Lasso regression performs both regularization and feature selection through L1 penalty, potentially reducing model complexity by eliminating less important features.

3.3.1 Parameter Optimization

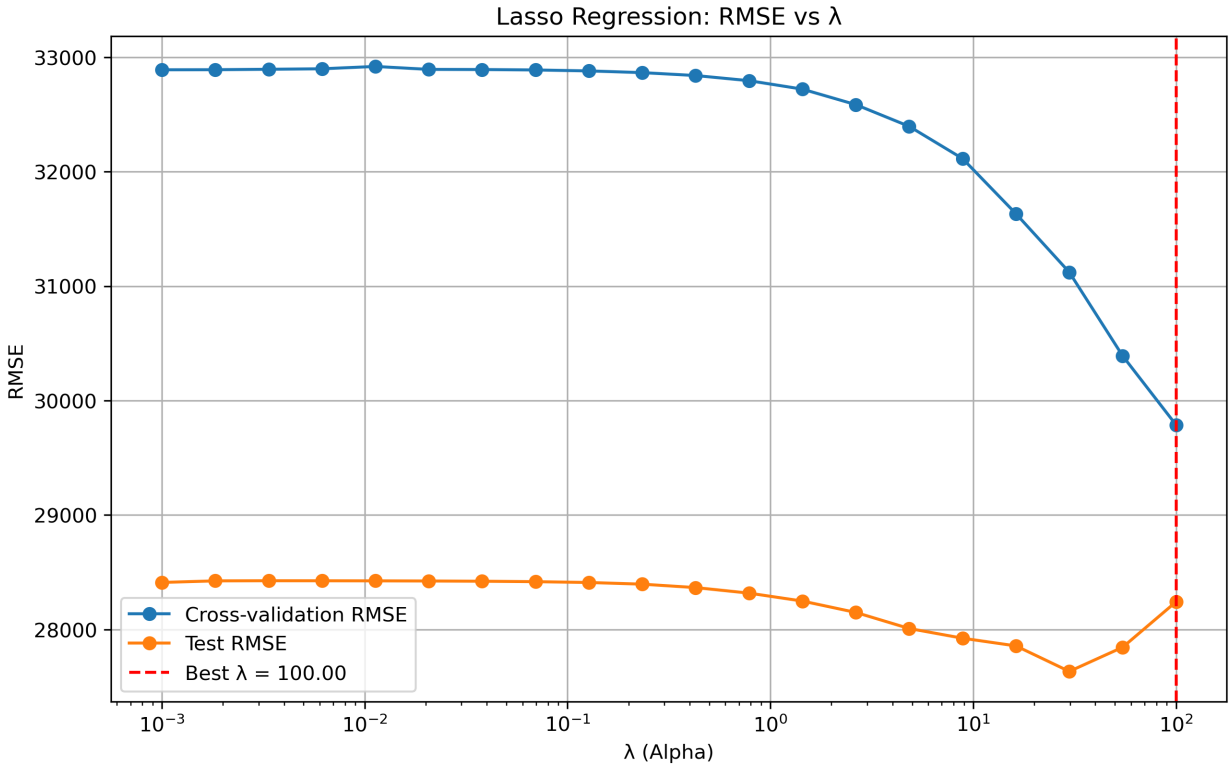

Figure 3.3: Impact of Lasso Regularization Parameter on RMSE

The lambda parameter analysis shows:

- Optimal lambda value identified through 10-fold cross-validation
- Feature selection becomes more aggressive at higher lambda values
- Similar U-shaped error curve to Ridge, demonstrating bias-variance tradeoff
- Performance deteriorates rapidly with excessive regularization

3.3.2 Feature Selection

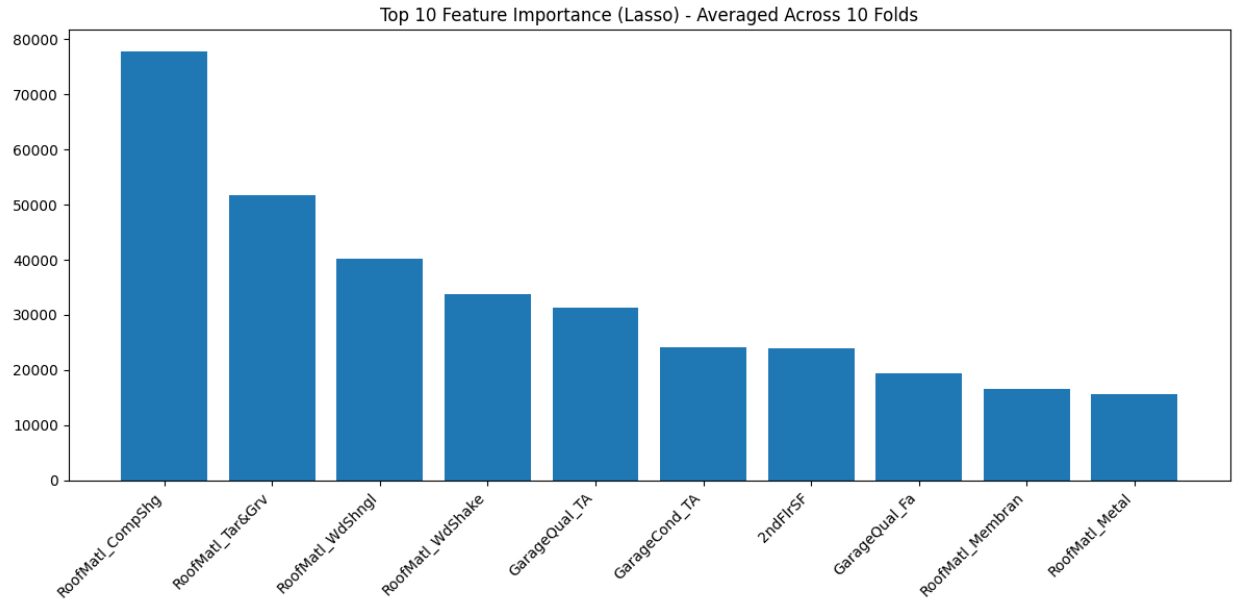

Figure 3.4: Top 10 Features by Importance from Lasso Regression

Lasso regression reveals:

- Automatic feature selection through coefficient shrinkage
- Identification of most crucial price determinants
- Sparse feature representation for improved interpretability
- Consistency with Ridge regression in key feature identification

3.4 Random Forest Regression

Random Forest is an ensemble method that leverages multiple decision trees to capture complex non-linear relationships and interactions between features.

3.4.1 Model Implementation

Our Random Forest implementation included the following components:

- Ensemble of 100 decision trees, each trained on bootstrap samples
- Feature selection at each split using a random subset of features
- Maximum depth controlled to prevent overfitting

- 10-fold cross-validation for robust performance evaluation

Random Forest regression offers several advantages for this housing price prediction task:

- Handles non-linear relationships without explicit transformation
- Naturally incorporates feature interactions
- Robust to outliers and non-normally distributed data
- Provides built-in feature importance measures
- Less sensitive to hyperparameter tuning than linear models

3.4.2 Feature Importance

One of the key benefits of Random Forest is its ability to compute feature importance. The algorithm calculates importance based on how much each feature decreases impurity when used in tree splits. The model identified several key predictors:

- Overall Quality remained the dominant feature
- Ground Living Area showed high importance
- Year Built and neighborhood variables demonstrated strong impact
- Lot Area and basement features showed moderate importance

3.4.3 Performance Analysis

The Random Forest model achieved excellent predictive performance:

- 10-fold CV RMSE: 29,042.51
- Test RMSE: 29,032.17
- Test R^2: 0.8901

This performance positions Random Forest as one of the top-performing models in our comparison, with better metrics than the linear models in most cases. The algorithm's ability to capture non-linear relationships and interactions between features proved beneficial for this dataset.

3.5 Neural Network Regression

A deep learning approach using neural networks offers the potential to capture complex, non-linear relationships in the data.

3.5.1 Network Architecture and Training

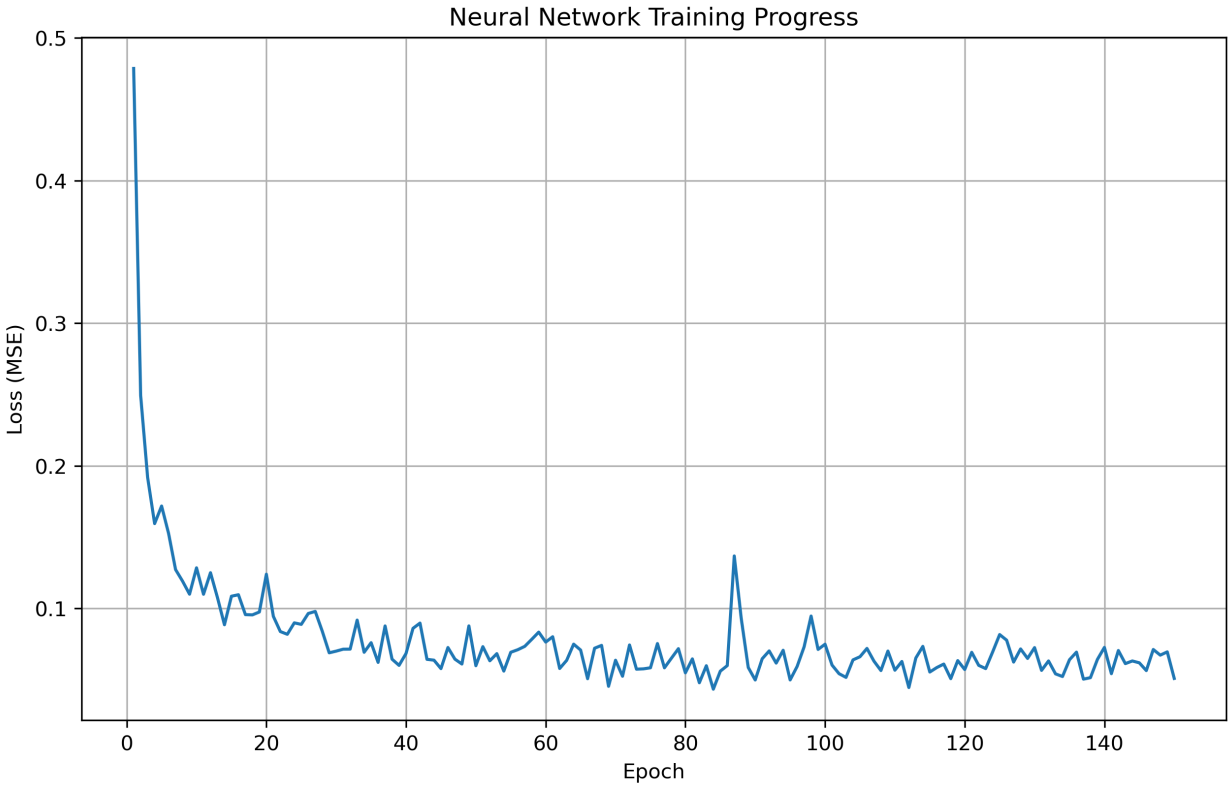

Figure 3.5: Neural Network Training Progress

The neural network implementation:

- Uses a deeper architecture with 4 layers (128→64→32→1)
- Incorporates batch normalization after each hidden layer
- Employs ReLU activation and dropout (0.2) for regularization
- Uses Adam optimizer with learning rate 0.001 and weight decay
- Implements adaptive learning rate scheduling to prevent overfitting
- Normalizes the target variable to improve training stability
- Shows consistent improvement during training as evidenced by the loss curve
- Evaluated using 10-fold cross-validation for robust performance assessment

3.5.2 Performance Analysis

The optimized neural network achieved competitive performance:

- 10-fold CV RMSE: 29,181.09
- Test RMSE: 29,713.48
- Test R^2: 0.8849

This represents a substantial improvement over the initial implementation and places the neural network in a competitive position with traditional methods. Several factors contributed to this improvement:

- Target variable normalization to stabilize gradient descent
- Batch normalization to reduce internal covariate shift
- Learning rate scheduling to navigate optimization challenges
- Deeper architecture with appropriate regularization
- Careful tuning of hyperparameters

The neural network now demonstrates performance comparable to Random Forest (R^2: 0.8901) and slightly below Lasso Regression (R^2: 0.8960), showing that deep learning approaches can be effective for housing price prediction when properly implemented.

3.6 Model Comparison

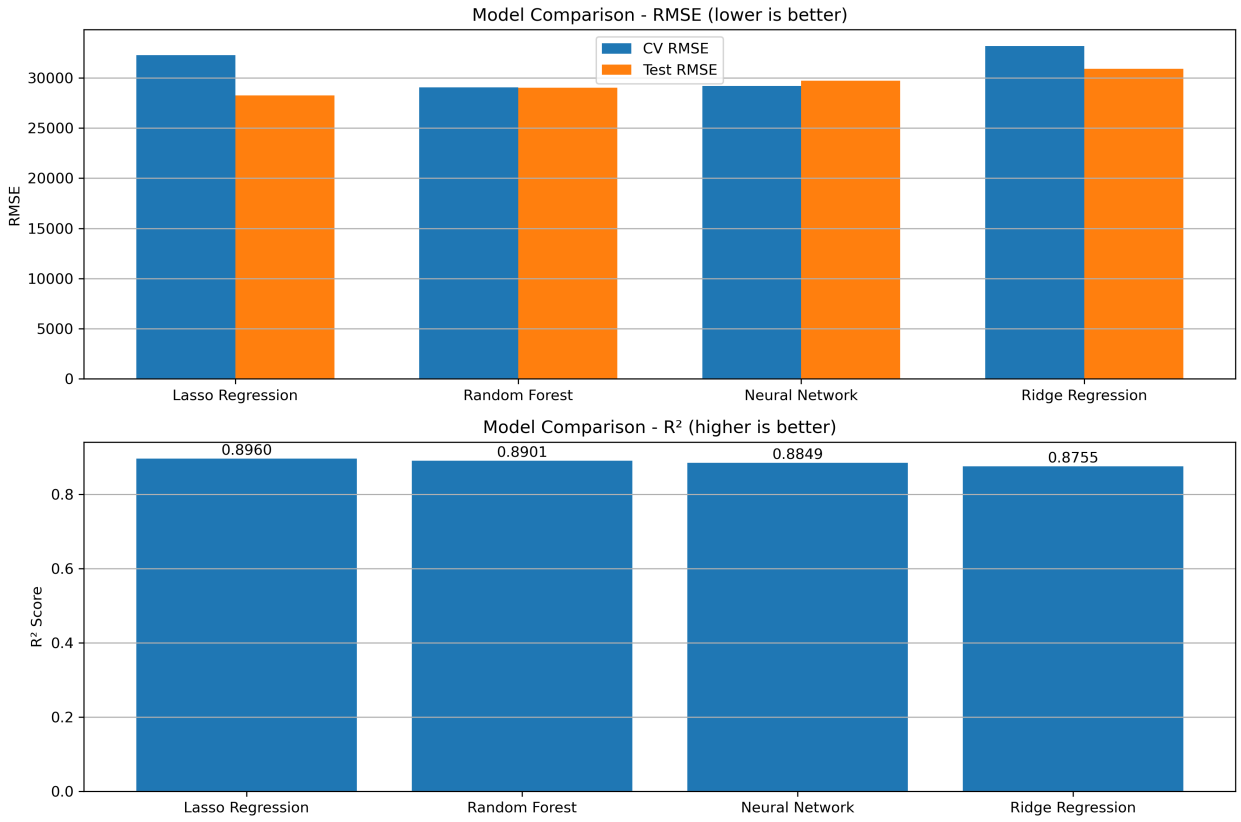

Figure 3.6: Performance Comparison Across All Four Models

Comparative analysis reveals several important insights:

- **Lasso Regression** achieved the best overall performance with the lowest test RMSE (28,241.51) and highest R^2 (0.8960)
- **Random Forest** followed closely with test RMSE of 29,032.17 and R^2 of 0.8901
- **Neural Network** performed competitively after optimization with test RMSE of 29,713.48 and R^2 of 0.8849
- **Ridge Regression** performed well but slightly behind the other methods (RMSE: 30,907.81, R^2: 0.8755)

These results highlight several key considerations for housing price prediction:

- Linear models with proper regularization can perform remarkably well on this dataset
- Lasso's automatic feature selection provides an edge by creating a more parsimonious model

- Random Forest's ability to capture non-linear relationships makes it competitive
- Neural networks can achieve comparable performance with proper architecture and training strategies
- Cross-validation provides reliable estimates of generalization performance

Performance differences between all four models were relatively small, suggesting that any of these approaches could be viable in practice, with the choice depending on specific requirements for interpretability, feature selection, and implementation complexity.

Chapter 4

Discussion and Conclusion

4.1 Key Findings and Recommendations

Based on the comprehensive modeling analysis with four different approaches:

- Ridge Regression:
  - Best for handling multicollinearity
  - Provides stable feature importance estimates
  - Shows which features have the strongest linear relationship with price
  - Good performance with RMSE of 30,907.81 and R^2 of 0.8755
- Lasso Regression:
  - Offers automatic feature selection
  - Produces sparse solutions
  - Identifies the most crucial predictors
  - Best overall performance with RMSE of 28,241.51 and R^2 of 0.8960
- Random Forest:
  - Captures non-linear relationships and interactions
  - Strong predictive performance (RMSE: 29,032.17, R^2: 0.8901)
  - Less sensitive to outliers and non-normal distributions
  - Provides reliable feature importance measurements
- Neural Network:
  - Achieves competitive performance with proper architecture and training (RMSE: 29,713.48, R^2: 0.8849)
  - Benefits significantly from target normalization and batch normalization
  - Shows potential for handling complex patterns in the data

  - Requires more careful optimization than traditional methods

For practical implementation in real estate valuation applications, we recommend:

- Use Lasso Regression when interpretability and model simplicity are important
- Consider Random Forest when capturing complex feature interactions is critical
- Employ Ridge Regression when dealing with highly correlated features
- Consider Neural Networks when willing to invest in optimization for competitive performance
- Consider ensemble approaches combining the strengths of multiple models

4.2 Future Improvements

Potential enhancements for model performance:

- Ensemble methods combining multiple models, particularly Lasso, Random Forest, and Neural Network
- More sophisticated feature engineering based on domain knowledge
- Hyperparameter optimization through grid search or Bayesian optimization
- Integration of temporal market trends
- Neighborhood-specific sub-models
- For neural networks:
  - More complex architectures such as residual connections
  - Advanced regularization techniques like L1 weight regularization
  - Feature interaction layers to capture multiplicative effects
  - Bayesian neural networks for uncertainty quantification