MSDS 7335 - Machine Learning II

Comparative Analysis of Regression Models for Ames Housing Price Prediction

Your Name

May 16, 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 2 presents detailed exploratory data analysis

- Chapter 3 will cover feature engineering and selection

- Chapter 4 will present modeling approaches and results

- Chapter 5 will provide conclusions and recommendations

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

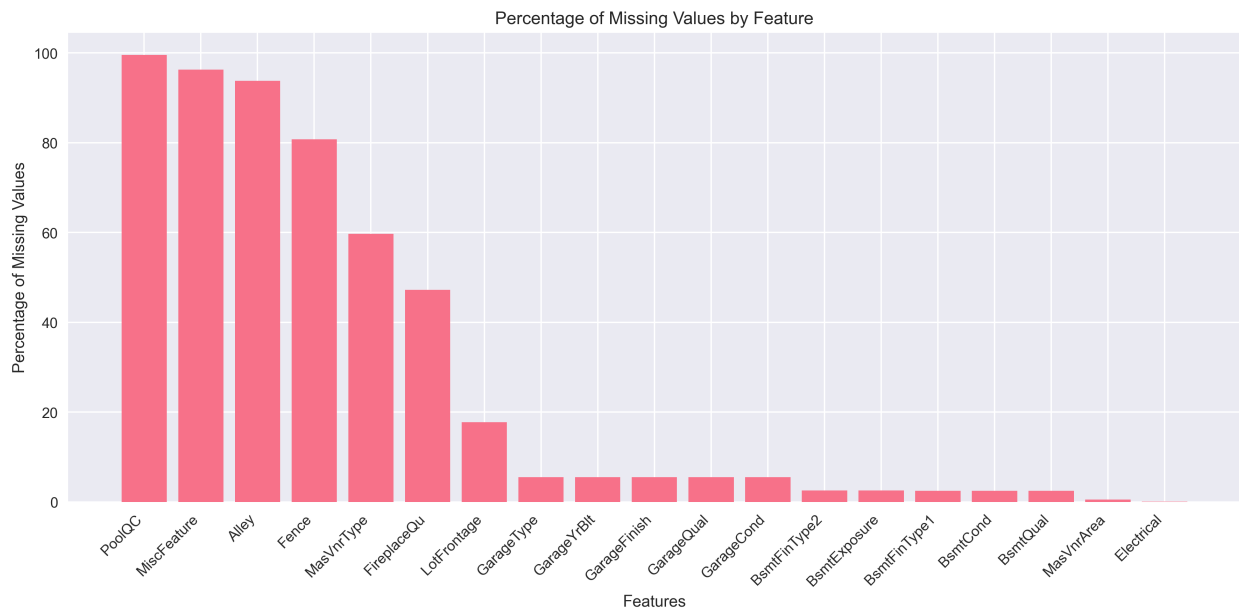## 2.2 Missing Value Analysis

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 modeling purposes, missing values in categorical variables will be treated as 'None' or 'Not Available'
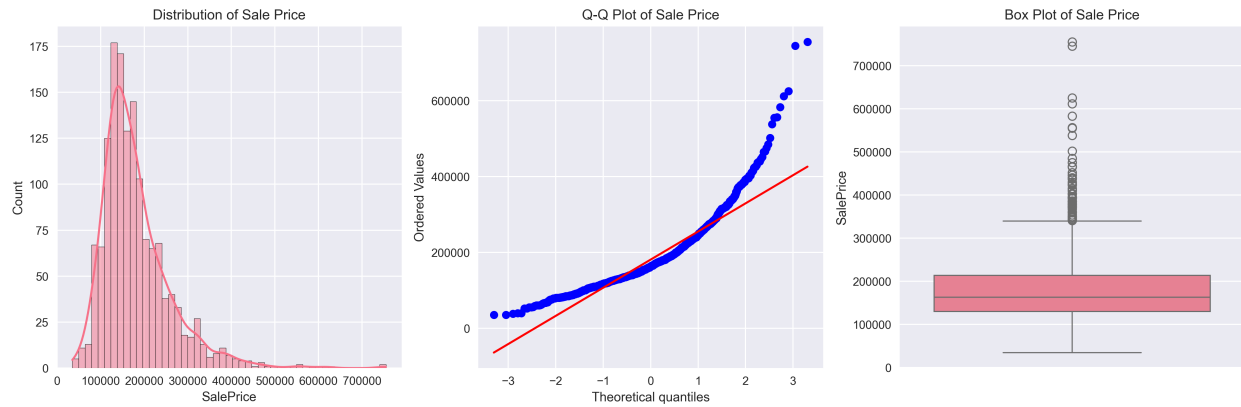
2.3 Target Variable Analysis

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

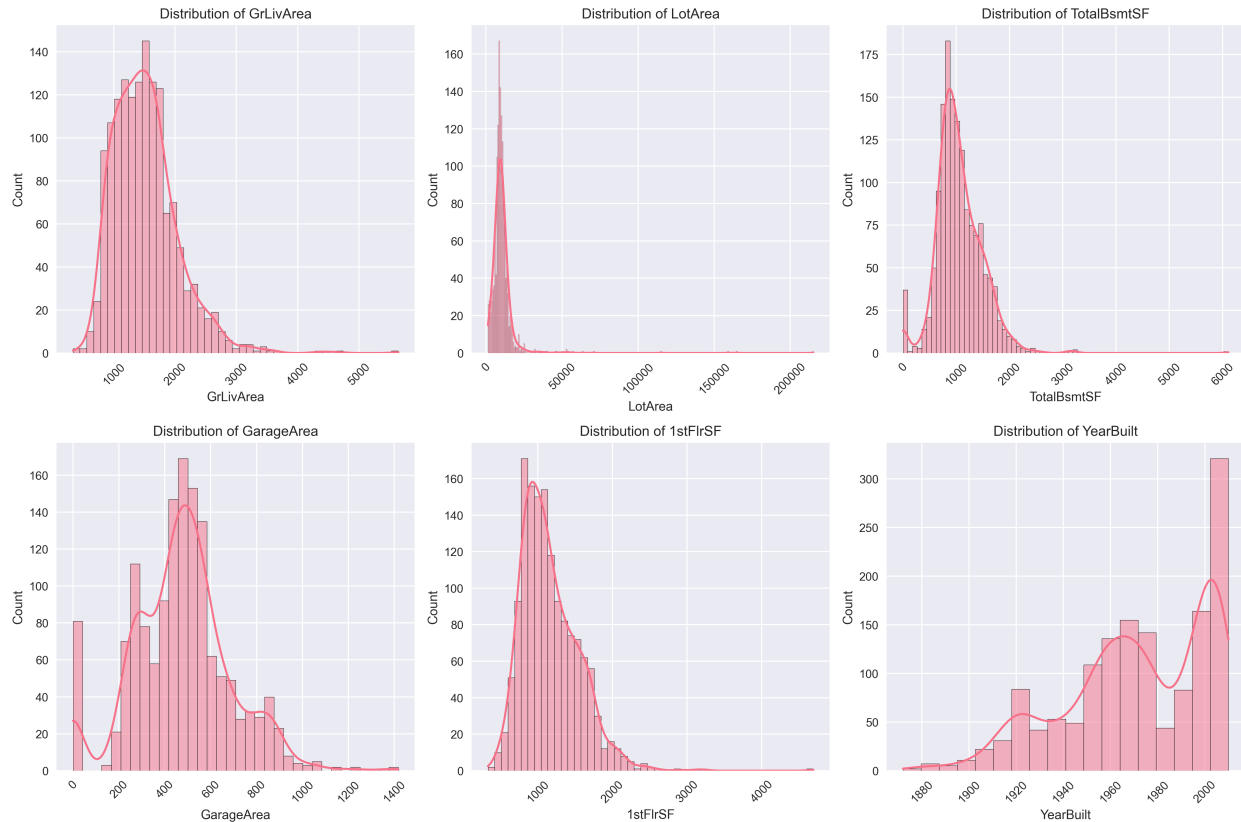### 2.4.1 Numerical Features

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
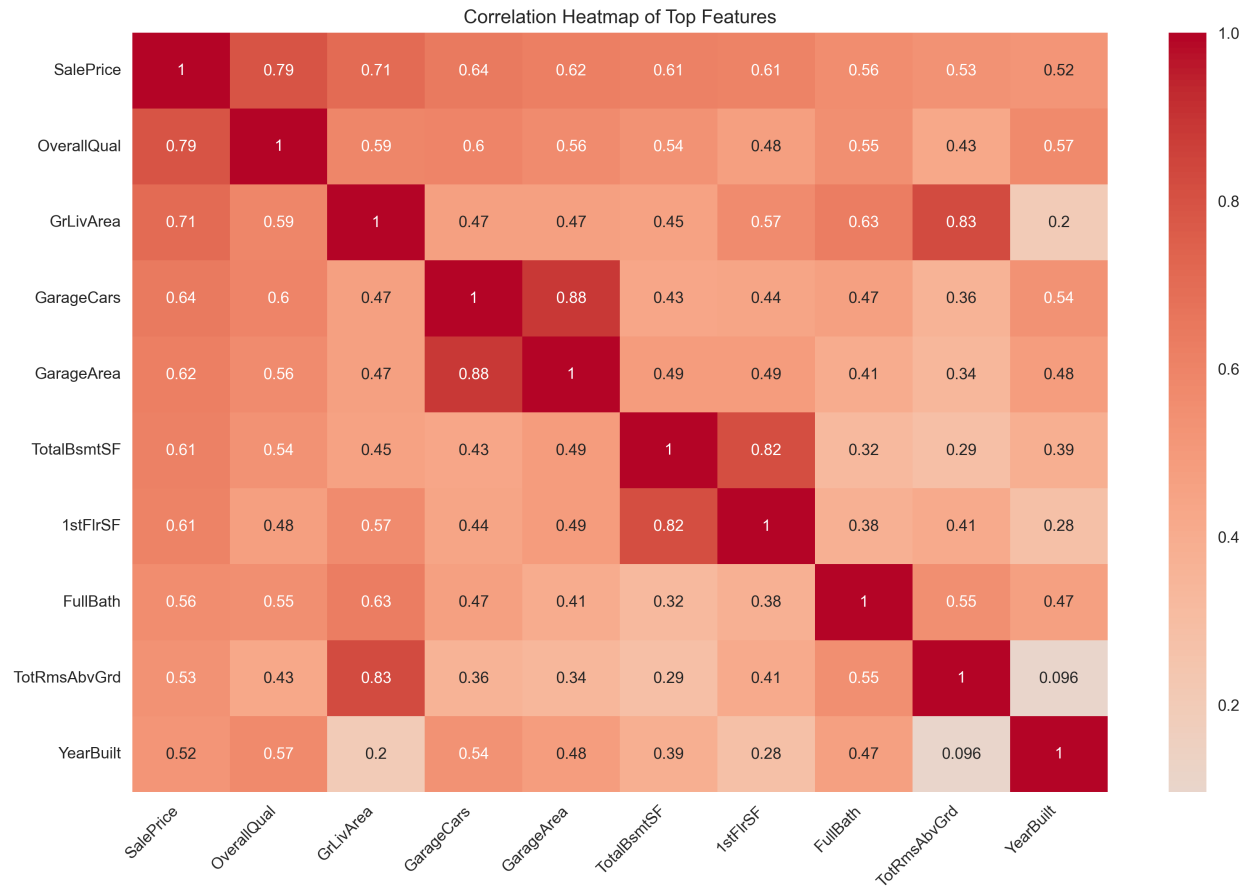
2.4.2 Feature Correlations

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
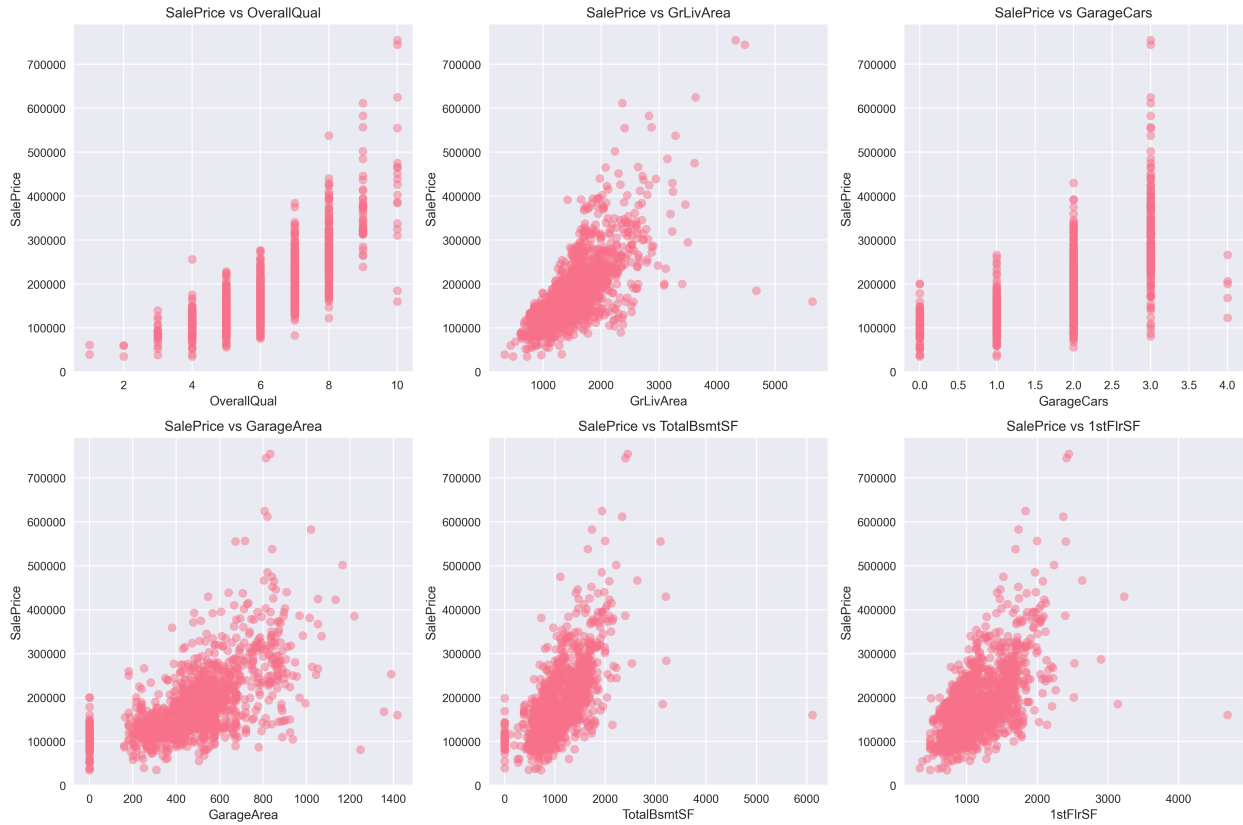
2.4.3 Feature Relationships

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
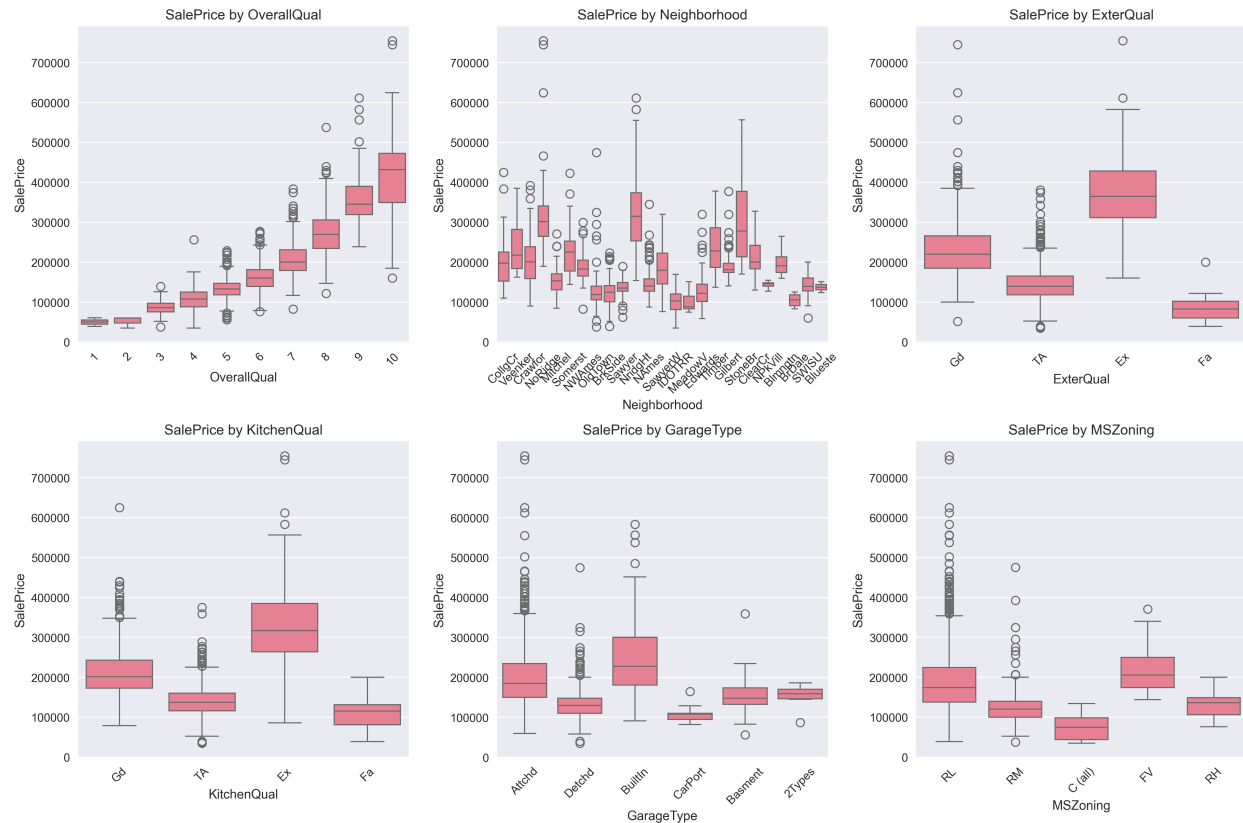
2.5 Categorical Features

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

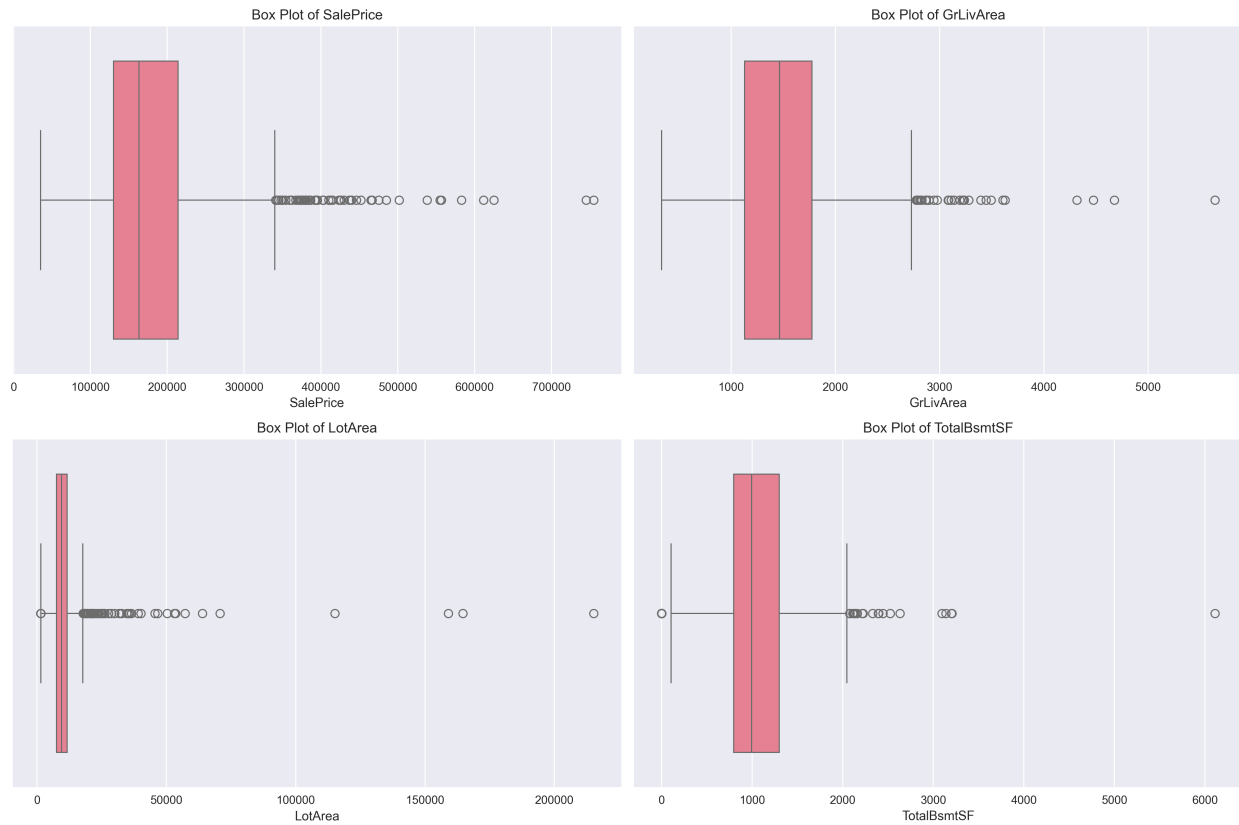## 2.6 Outlier Analysis

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
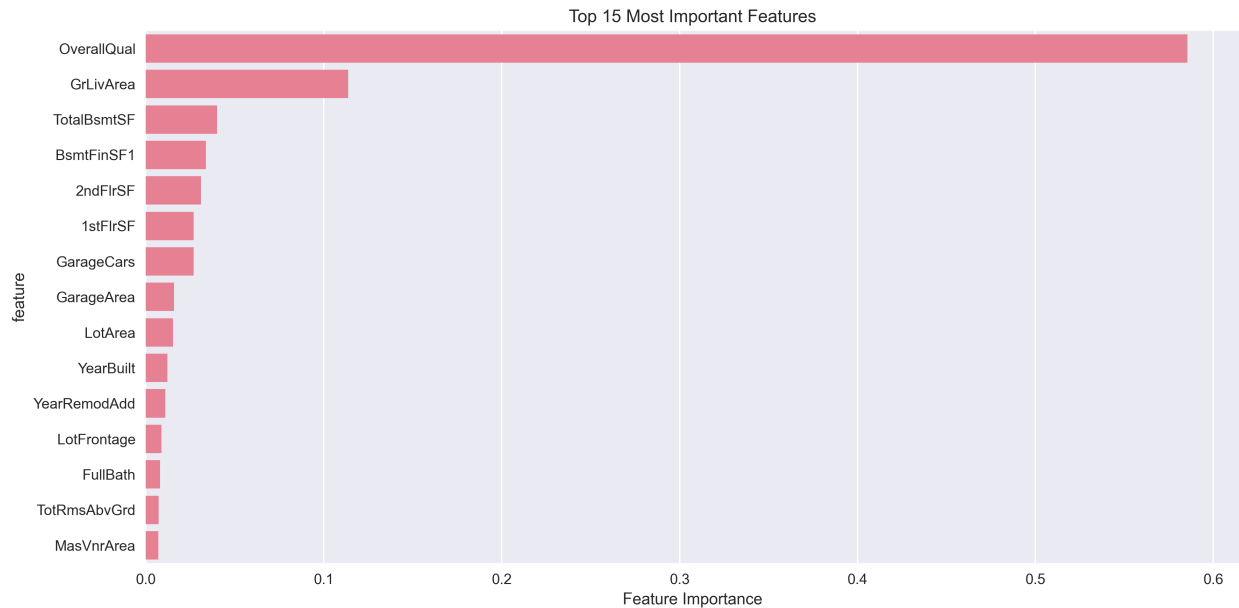
2.7   Feature Importance

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 three different modeling approaches for predicting house prices in the Ames Housing dataset:

- Ridge Regression (L2 regularization)

- Lasso Regression (L1 regularization)

- Neural Network Regression

Each model offers unique advantages and characteristics in handling the complexities of house price prediction.

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.
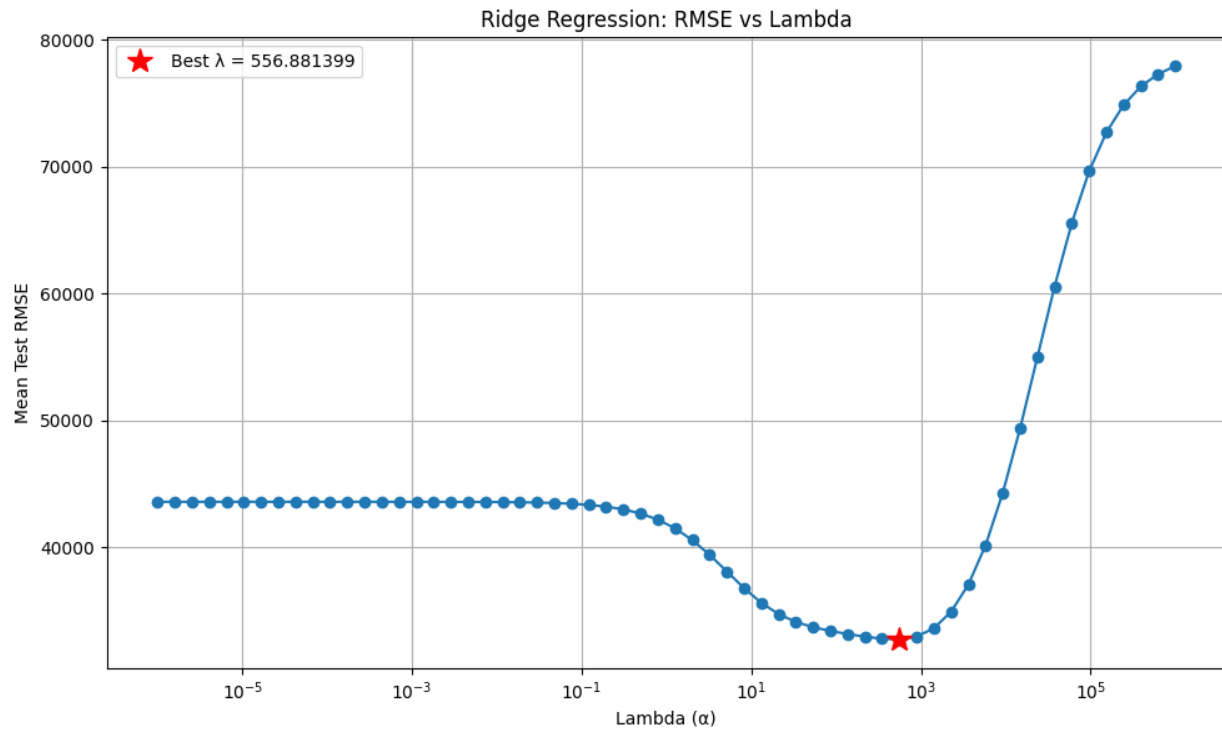
3.2.1 Hyperparameter Tuning

Figure 3.1: Effect of Ridge Regularization Parameter on Model Performance

The analysis of different lambda values reveals:

- Optimal lambda value found at approximately 1048.11

- Best RMSE achieved: 33,839.38

- Model performance degrades significantly with lambda > 2000

- Excessive regularization leads to underfitting, as shown by increasing RMSE
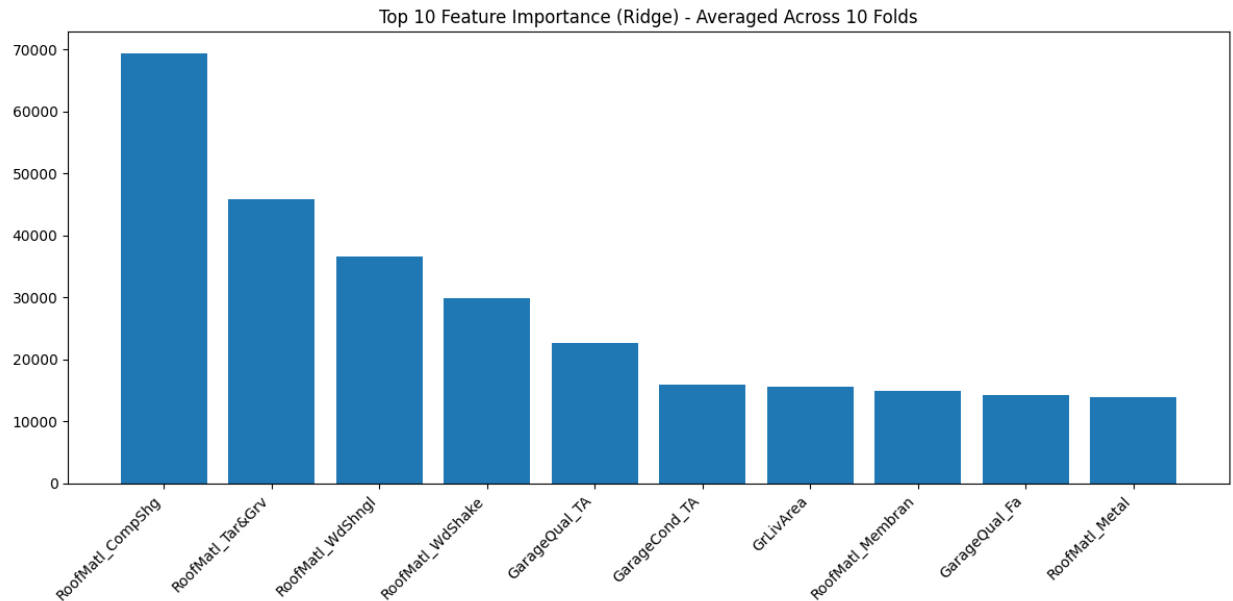
3.2.2 Feature Importance

Figure 3.2: Feature 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.
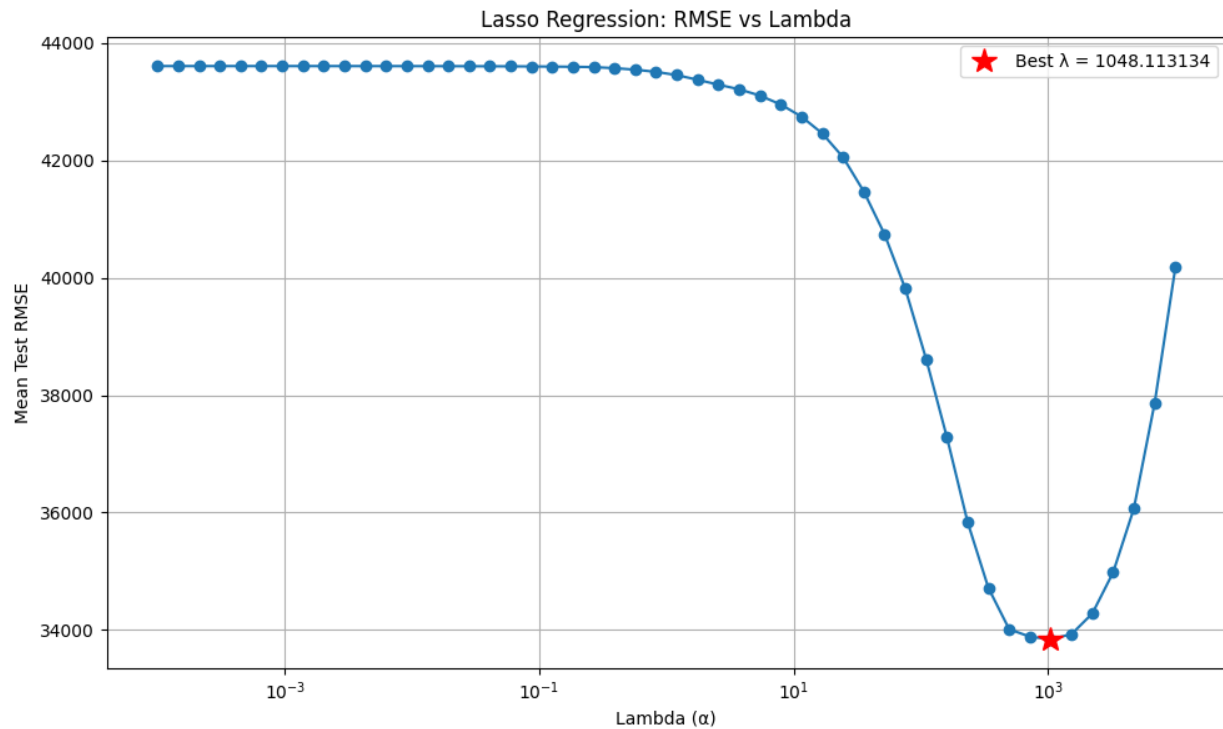
3.3.1 Parameter Optimization

Figure 3.3: Impact of Lasso Regularization Parameter on RMSE

The lambda parameter analysis shows:

- Optimal lambda value identified at 1048.11

- Minimum RMSE achieved: 33,839.38

- Performance deteriorates rapidly with lambda > 2000

- Feature selection becomes more aggressive at higher lambda values
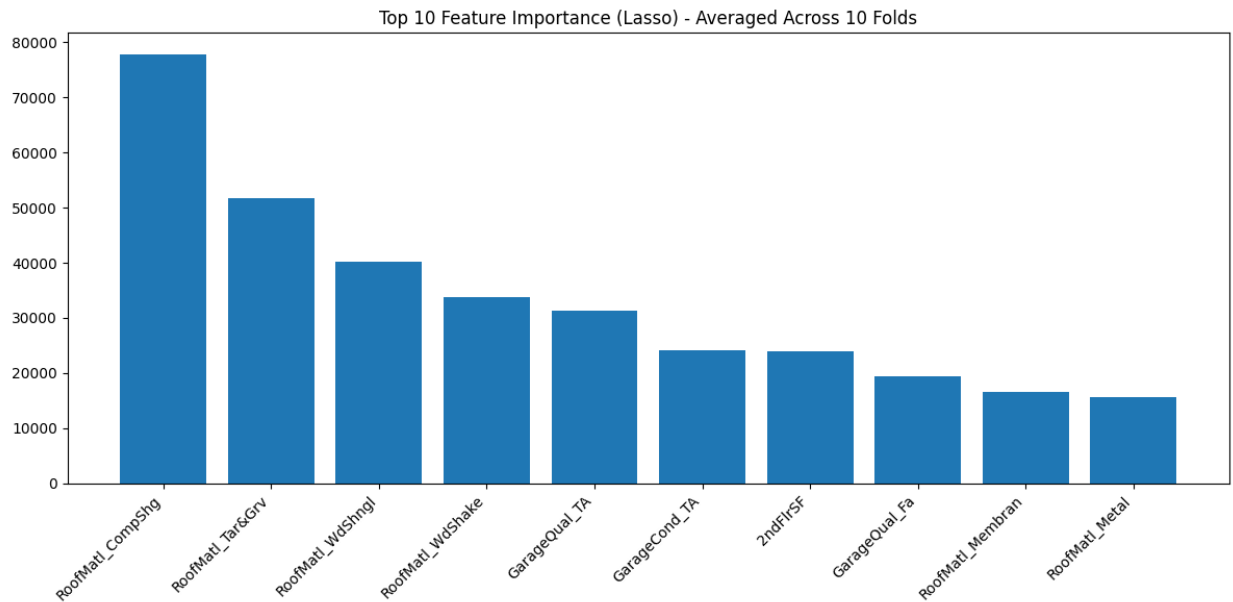
3.3.2 Feature Selection

Figure 3.4: Feature 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 Neural Network Regression

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

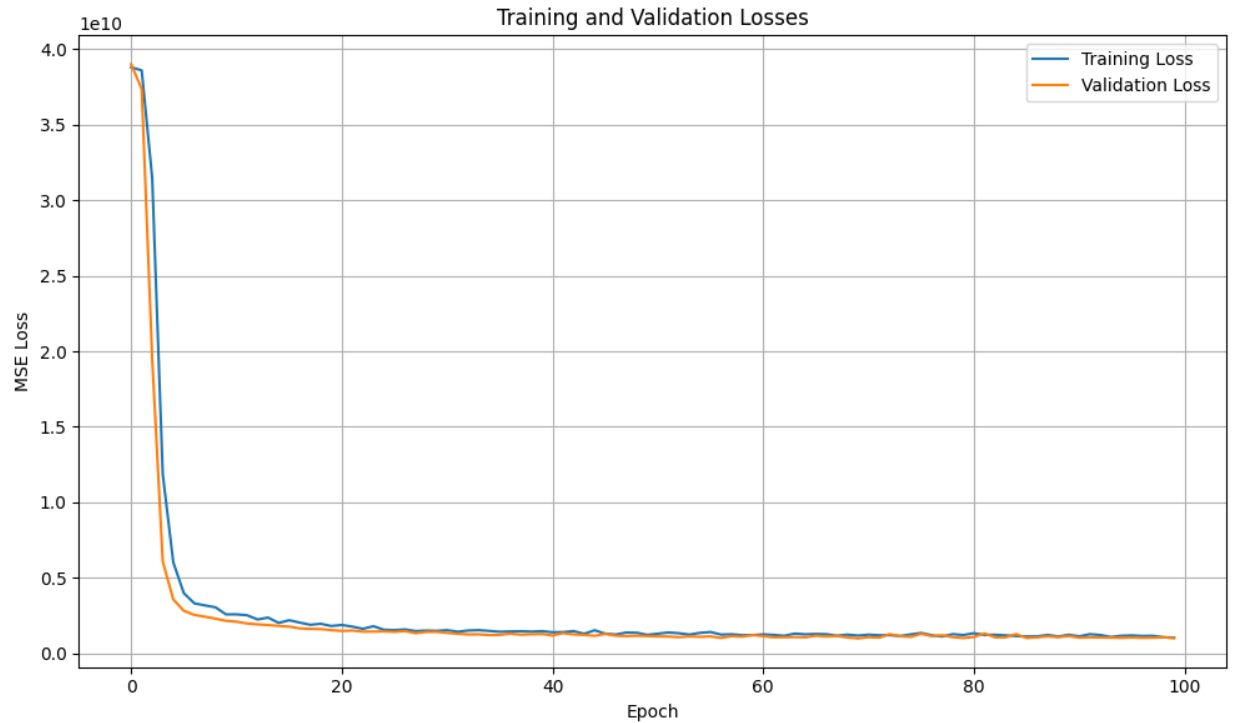### 3.4.1   Network Architecture and Training

Figure 3.5: Neural Network Training Progress

The neural network implementation:

- Utilizes multiple hidden layers for complex pattern recognition

- Shows consistent improvement during training

- Employs dropout for regularization

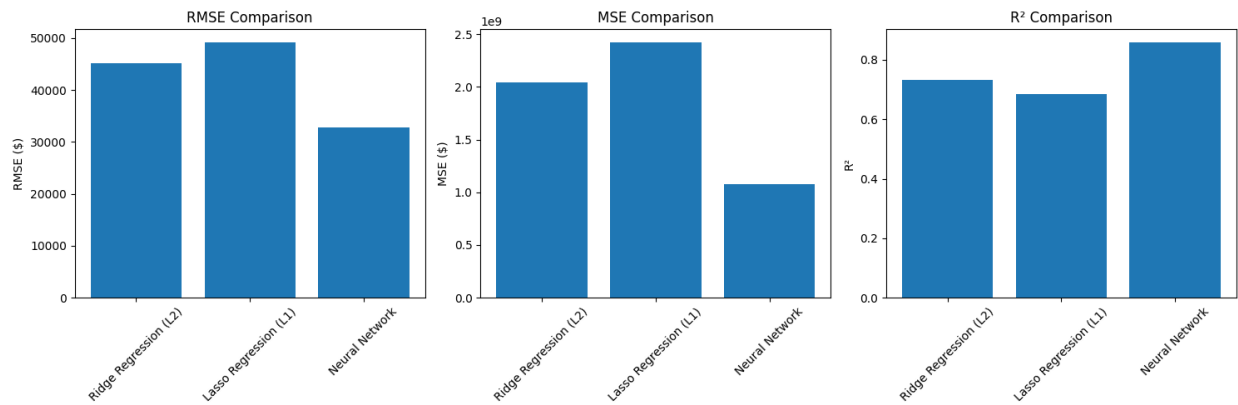- Demonstrates good convergence characteristics

3.5 Model Comparison

Figure 3.6: Performance Comparison Across Models

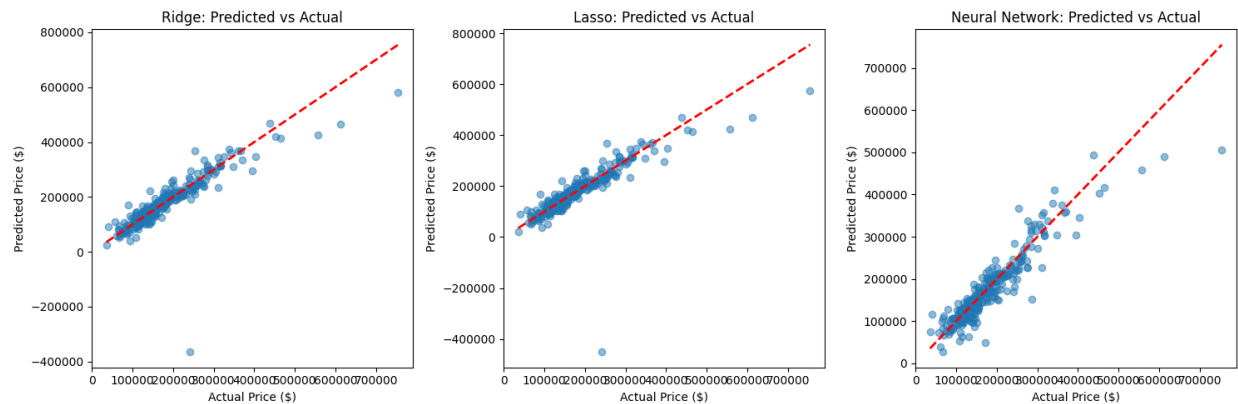### 3.5.1 Prediction Analysis

Figure 3.7: Prediction Comparison Between Models

Comparative analysis reveals:

- Both Ridge and Lasso achieve similar optimal performance (RMSE 33,839)

- Linear models (Ridge and Lasso) provide good interpretability

- Neural network captures complex non-linear relationships

- Model ensemble potential for improved predictions

3.6   Key Findings and Recommendations

Based on the comprehensive modeling analysis:

- Ridge Regression:

    - Best for handling multicollinearity
    - Provides stable feature importance estimates
    - Achieves optimal performance at lambda 1048

- Lasso Regression:

    - Offers automatic feature selection
    - Produces sparse solutions
    - Shows similar optimal lambda value to Ridge

- Neural Network:

    - Captures complex non-linear relationships
    - Shows potential for high accuracy
    - Requires more data for optimal performance

3.7   Future Improvements

Potential enhancements for model performance:

- Ensemble methods combining multiple models

- Feature engineering based on domain knowledge

- Hyperparameter optimization through cross-validation

- Integration of temporal market trends

- Neighborhood-specific sub-models