Predicting Housing Prices in Ames Using Advanced Regression Techniques

Webster Estimé [1] and Vladzimir Kushchav [2]

wbe21@fsu.edu [1]

vk17@fsu.edu [2]

STA 5167, Statistics in Applications II

Dr. Xin Zhang

Florida State University

May 1, 2025

Introduction

Estimating home prices is of critical importance to various stakeholders in society. For individuals and families, reliable estimates of a home's market value can guide import financial decisions such as buying, selling, refinancing, or choosing where to live. Overpaying or underestimating a property's value may have serious financial consequences, particularly when considering mortgage affordability, property taxes, and future resale value.

For real estate investors, predictive models are valuable tools for identifying undervalued properties, optimizing investment portfolios, and estimating return on investment. These models enable data-driven decisions in competitive markets where timing and realistic valuation are crucial.

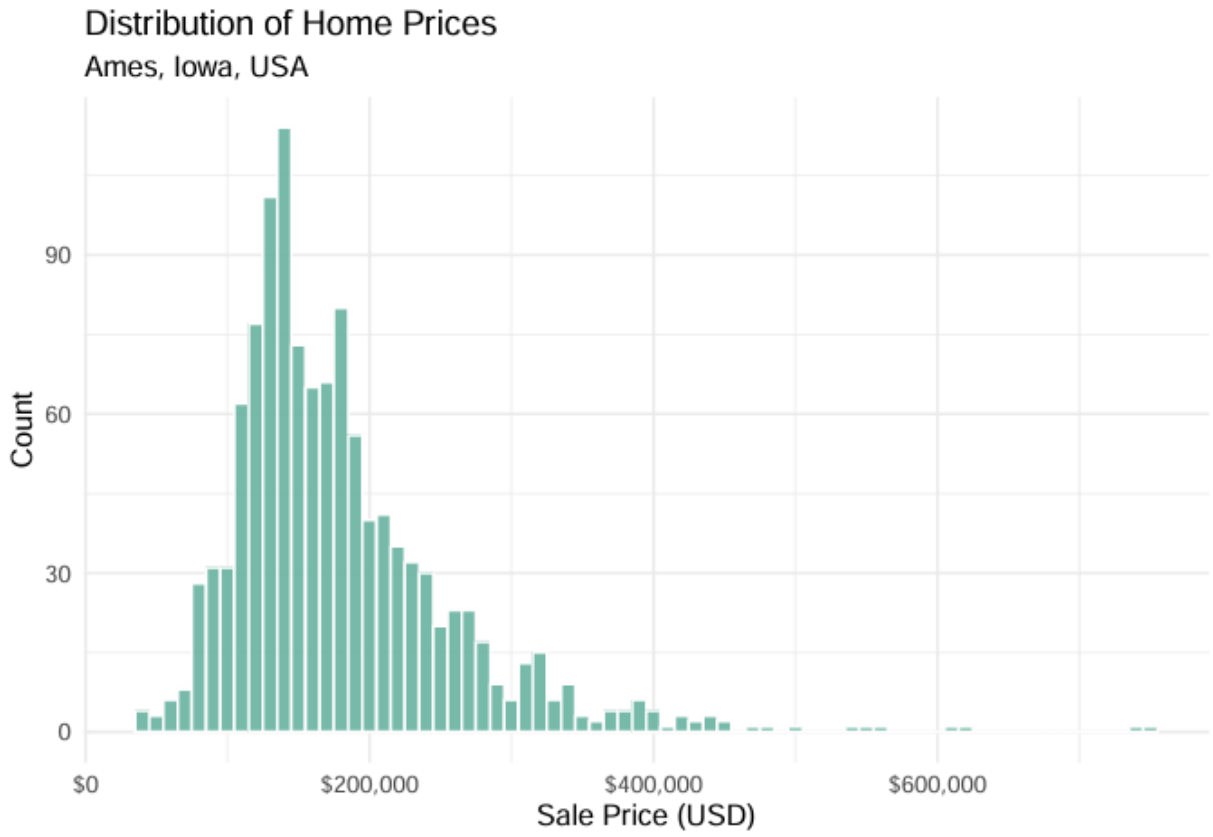
Furthermore, lenders, insurers, and policymakers also benefit from reliable price estimates, as they inform risk assessment, underwriting decisions, and housing and tax policies. By modeling sale prices based on physical characteristics and market factors, we help bring transparency and fairness to real estate transactions.

In this project, we develop and compare predictive modeling approaches, such as Principal Component Regression ("PCR"), Random Forest, Least Absolute Shrinkage and Selection Operator Regression ("LASSO"), and gradient boosted trees models - using data from Ames, Iowa housing dataset. The dataset was obtained from Kaggle. The goal was to identify a model that balances predictive accuracy with interpretability, using cross-validation to assess out-of-sample performance.

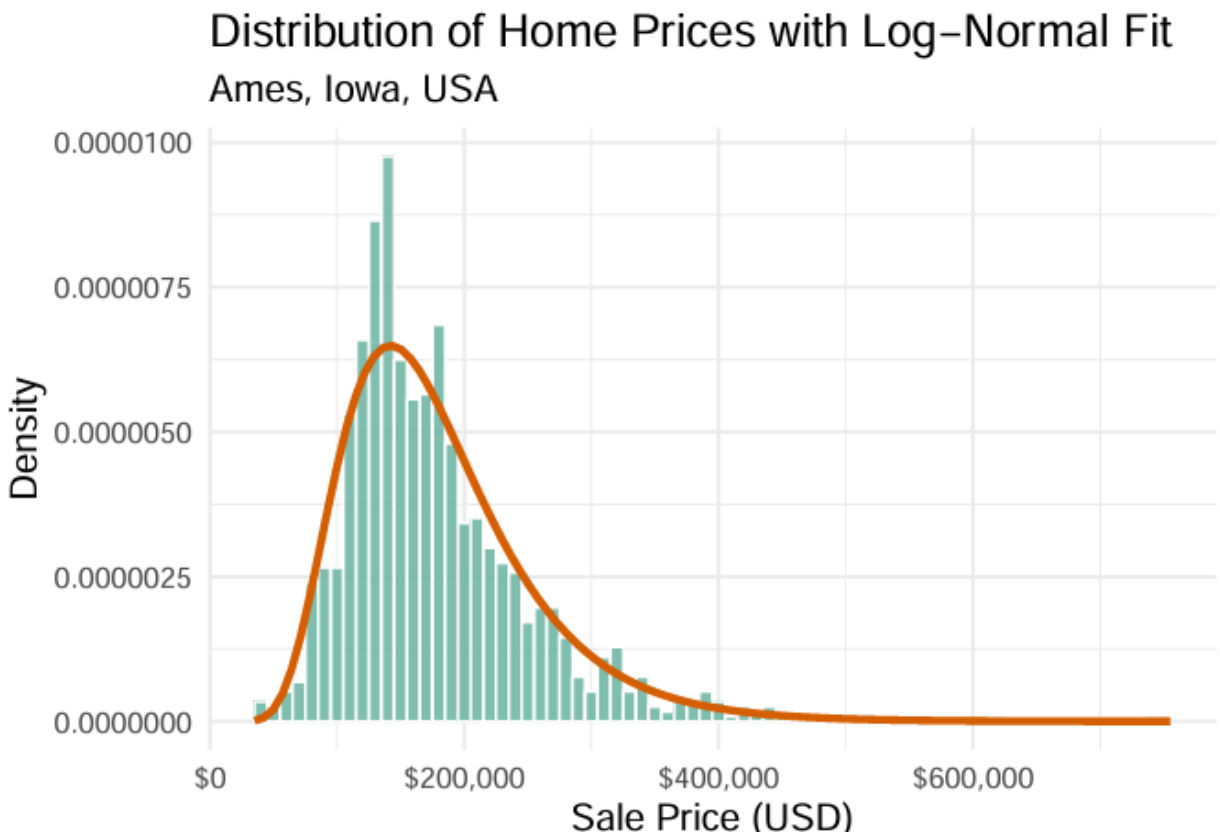
Data Preprocessing

The data set consists of 1,168 observations and 81 variables, including both numerical and categorical features. It was split into two sets: 80% being used for training and 20% for validation. Non-numeric features were removed, and missing values were imputed using either mean or median depending on the model.

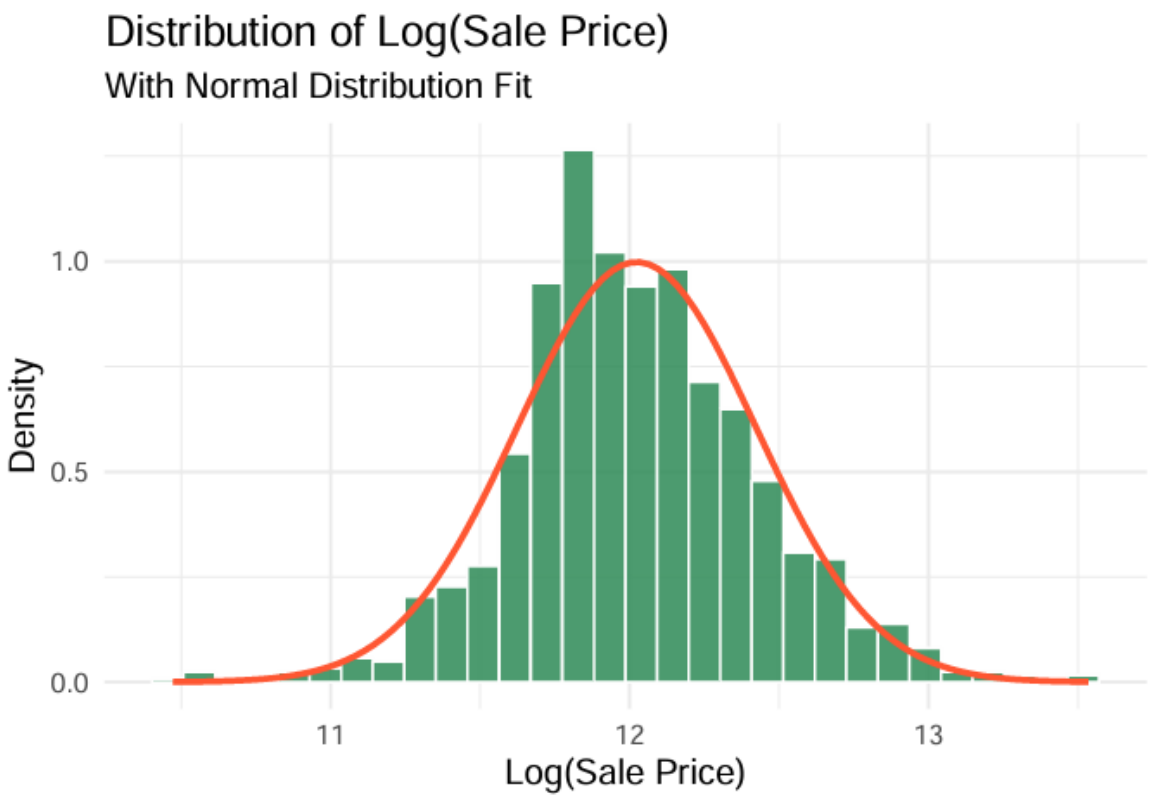
Log transformation of the response variable *SalePrice* was applied to stabilize variance and normalize the distribution. This transformation is justified by visual inspection of the skewness in the original distribution and the improved symmetry post transformation.



The response variable was approximately lognormally distributed with a mean of 12.02 and a standard deviation of 0.3996. Using these parameter estimates, a density plot fitted to the dataset as shown below.

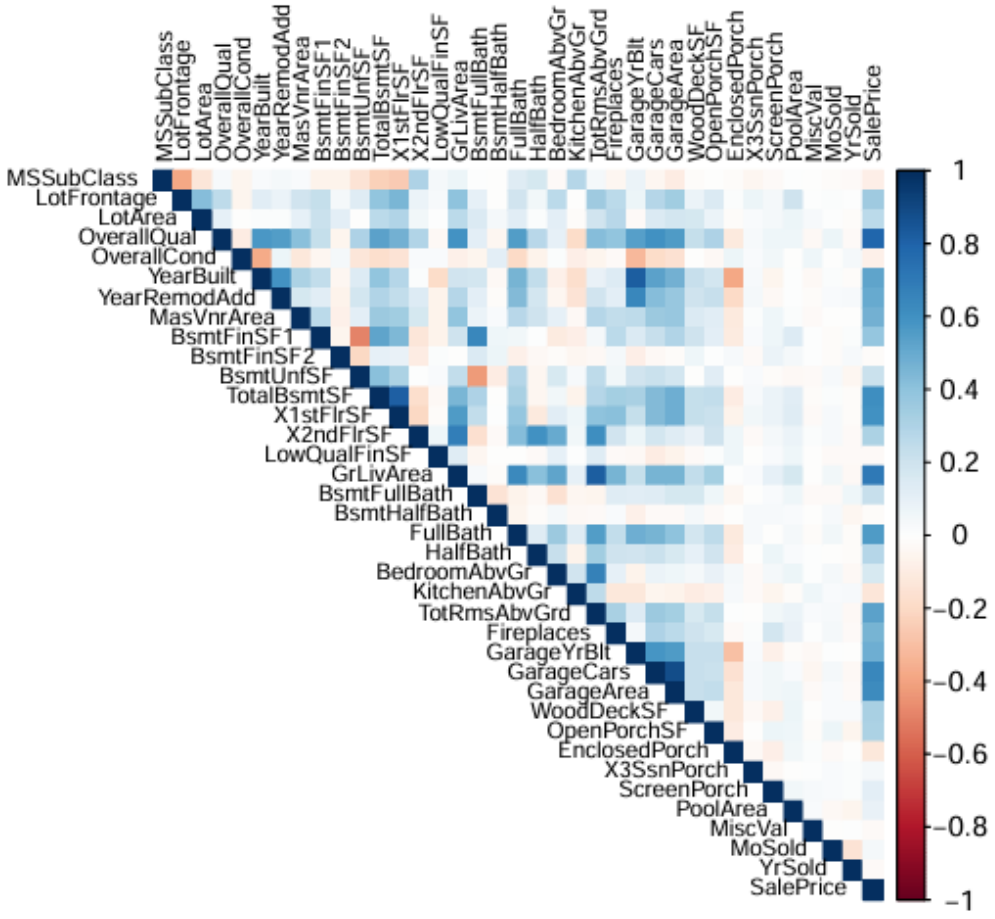


After the logarithmic transformation, the response *SalePrice* was found to be well approximated using a normal distribution as shown graphically below.

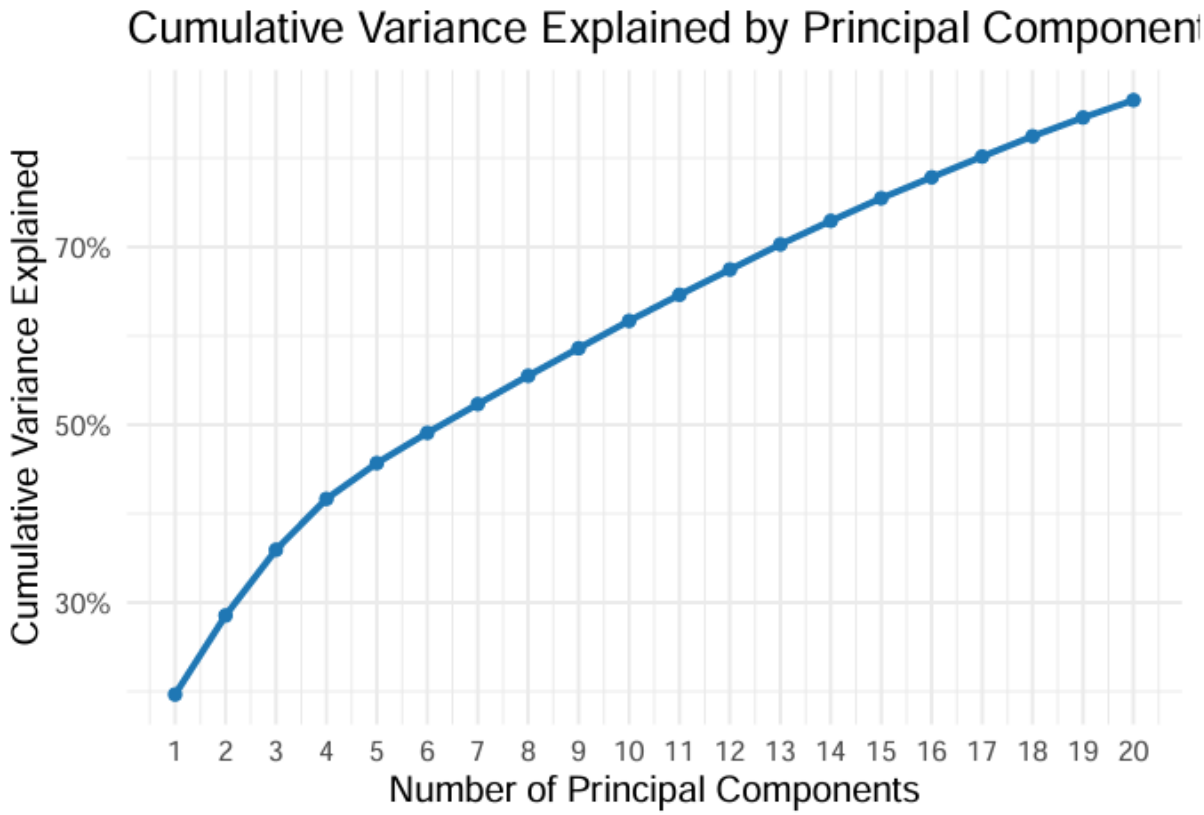


Exploratory Data Analysis

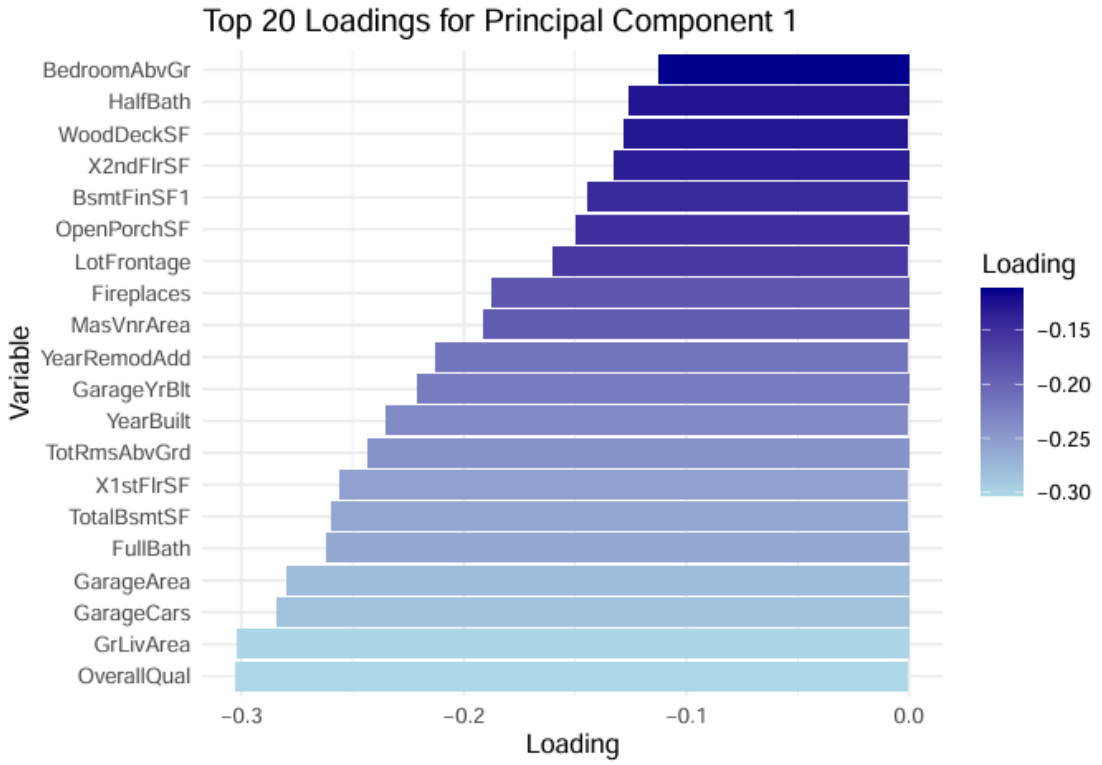
A Correlation analysis revealed several strong linear relationships between predictors and sale prices. Notably, the overall material quality and finish of the house rating ("*OverallQual*"), the above grade or ground living area square feet ("*GrLivArea*"), and the size of garage in terms of car capacity ("*GarageCars*") exhibited the highest correlation with *SalePrice*. See the correlation plot below for visualization.



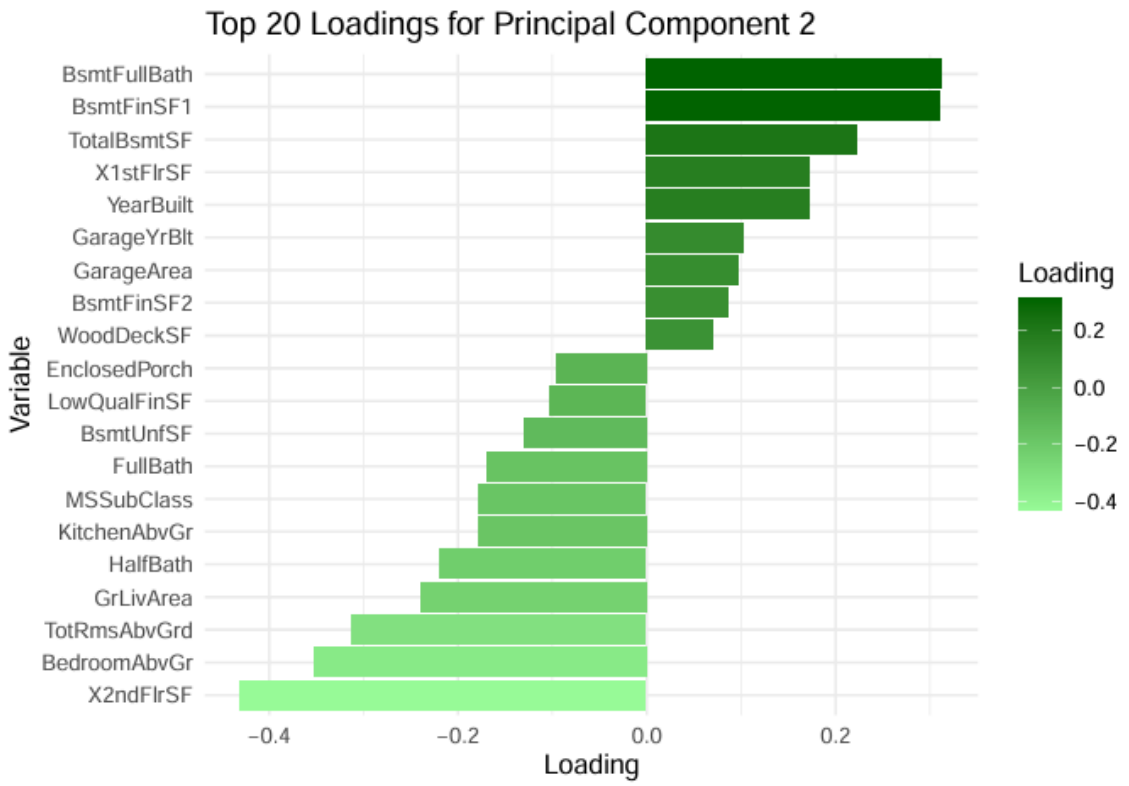
Principal Component Analys (“PCA”) was applied to the scaled numerical predictors, and scree plots indicated that the first 20 components out of 36 captured over 86% of the total variance.



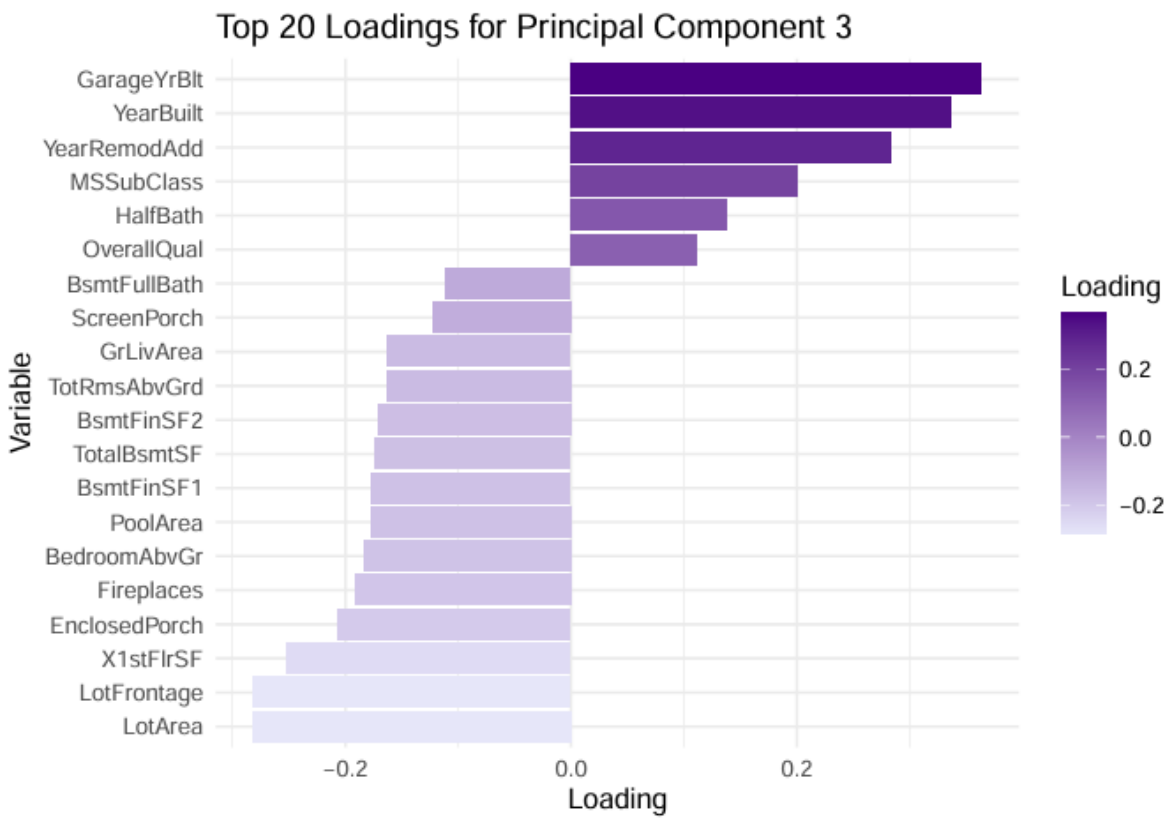
Below are plots of the first three principal components and the loadings of the most influential variables:



This bar plot above displays the top 20 variables contributing to the first principal component (PC1) in the PCA analysis. Variables with higher absolute loadings (e.g., *OverallQual*, *GrLivArea*, *GarageCars*) contribute more significantly to PC1, indicating that this component captures general indicators of home size and quality. All loadings are negative due to scaling direction, but in PCA, sign does not affect interpretability — what matters is magnitude.



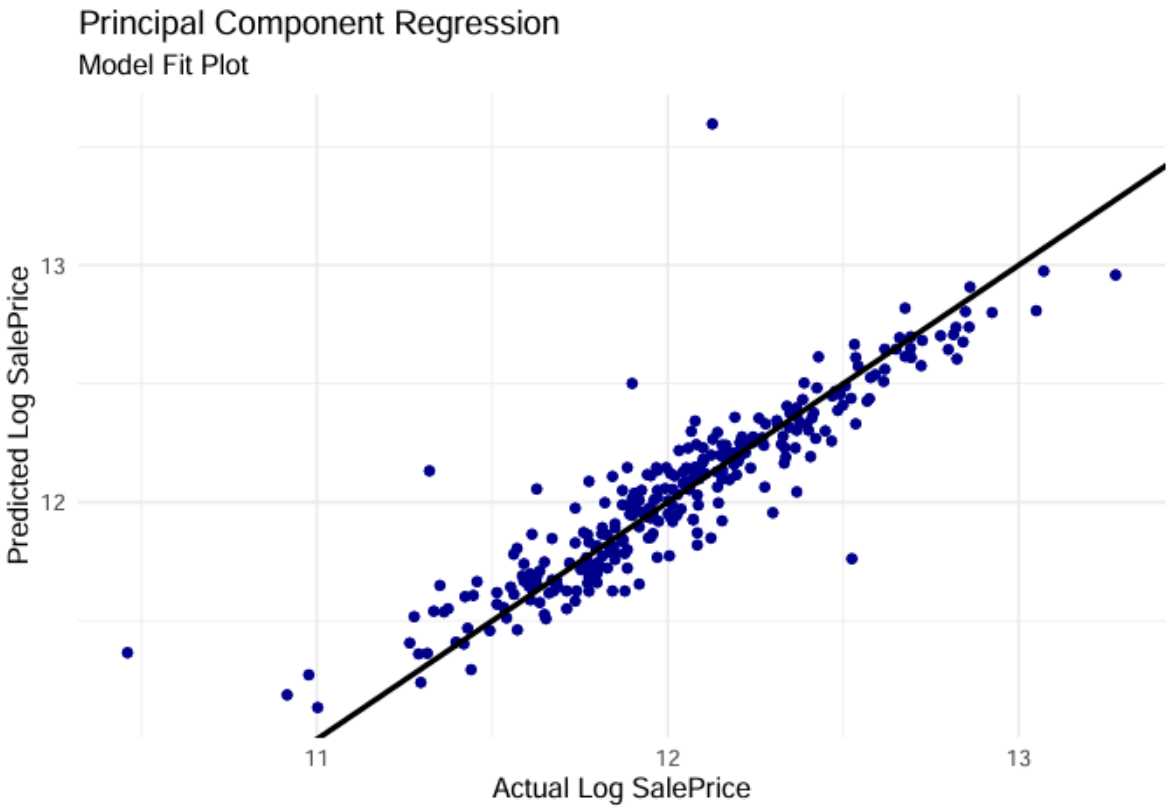
The bar plot above illustrates the top 20 variables contributing to Principal Component 2 (PC2). Variables with positive loadings (e.g., *BsmtFullBath*, *BsmtFinSF1*, *TotalBsmtSF*) are mostly related to basement features and overall home age, while those with negative loadings (e.g., *X2ndFlrSF*, *BedroomAbvGr*, *TotRmsAbvGrd*) reflect above-ground living areas and room count. This suggests that PC2 contrasts homes with finished basements against those with larger upper-level living space, capturing a structural configuration dimension in the housing data.



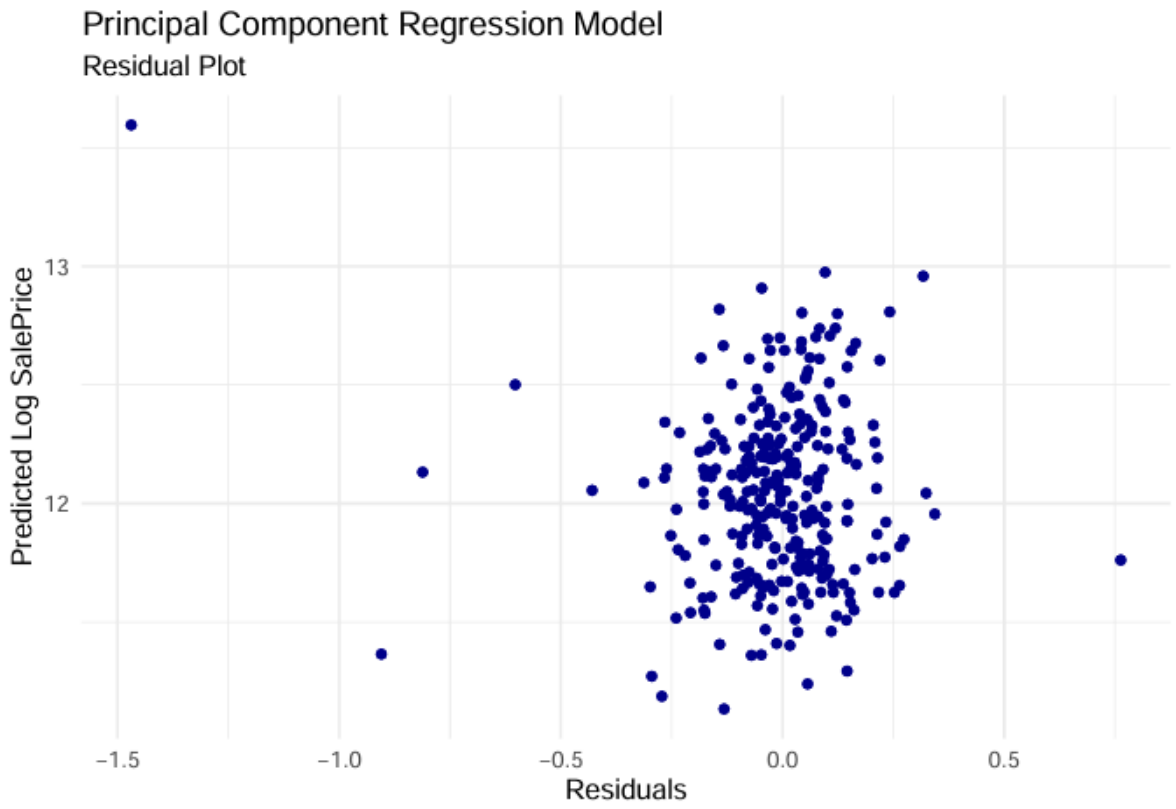
This bar chart shows the top 20 variables influencing Principal Component 3 (PC3). The component is positively associated with recent construction features such as *GarageYrBlt*, *YearBuilt*, and *YearRemodAdd*, suggesting PC3 captures a "modernity" or construction age factor. In contrast, negative loadings such as *LotArea*, *LotFrontage*, *X1stFlrSF*, and *EnclosedPorch* indicate an inverse relationship with larger lot sizes and traditional floor plans. Overall, PC3 distinguishes newer, possibly more compact homes from older or more spacious properties with traditional layouts.

Principal Component Regression (PCR)

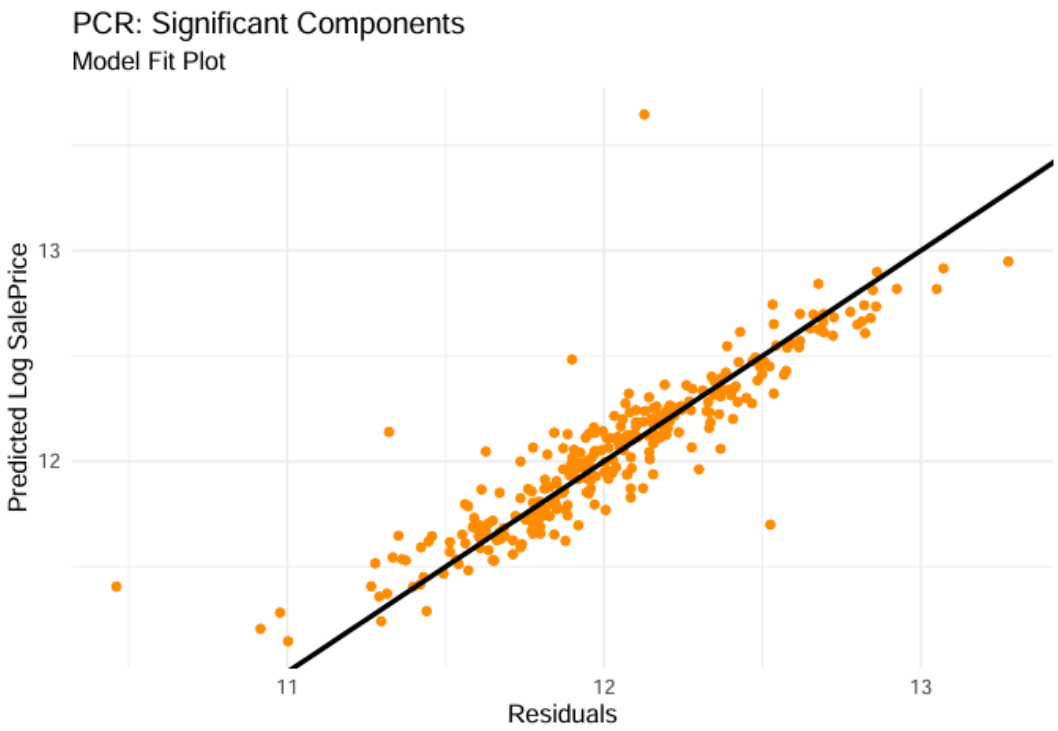
A principal component regression was fitted using the top 20 principal components. The initial full model achieved a Root Mean Squared Error ("RMSE") of 0.1738 on the validation set. Further refinement using only the components significant at $\alpha = 0.05$ reduced the model to 13 principal components, resulting in an RMSE of 0.1762. Below is a model fit plot.



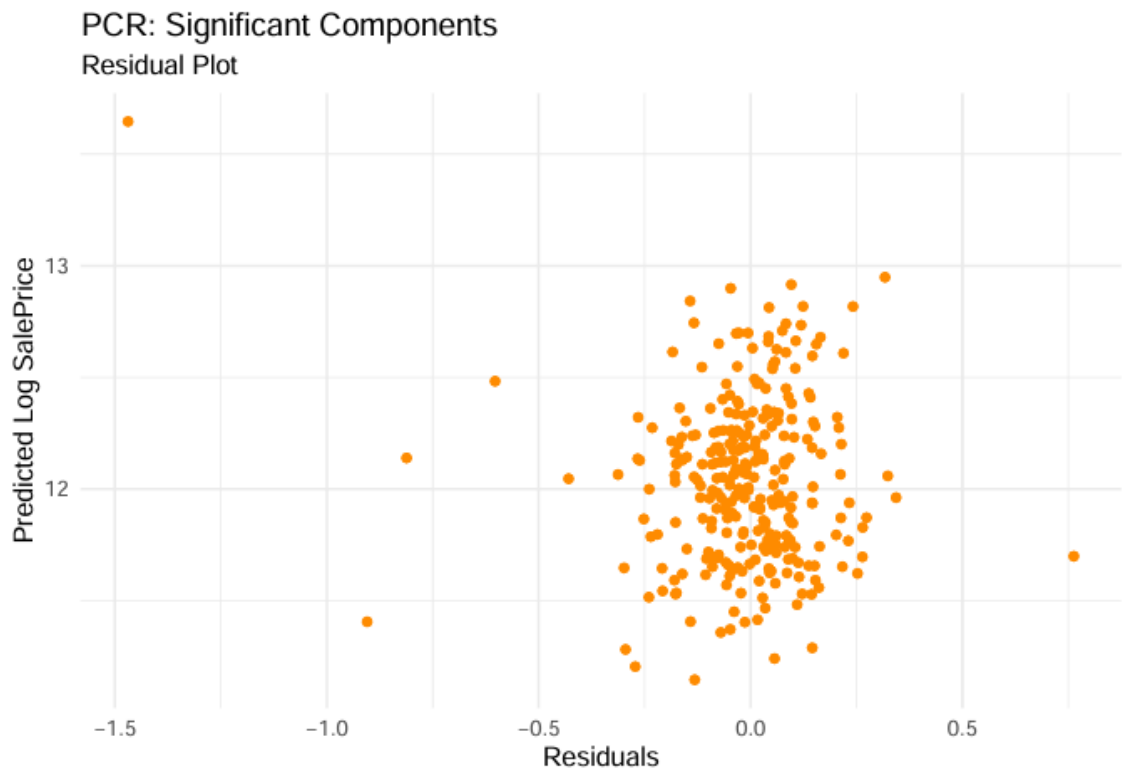
The scatter plot shows the predicted versus actual log-transformed sale prices from the Principal Component Regression (PCR) model. The points are generally aligned along the 45-degree reference line, indicating a good model fit. While most predictions are close to the actual values, some dispersion is visible at the lower and higher ends, suggesting mild under- or overestimation in those ranges.



The residual plot for the Principal Component Regression model shows residuals plotted against predicted log sale prices. The residuals are mostly centered around zero, indicating no major bias in the predictions. While a few outliers are present, especially on the left, there is no clear pattern or funnel shape, suggesting the model satisfies assumptions of homoscedasticity and is reasonably well-calibrated.



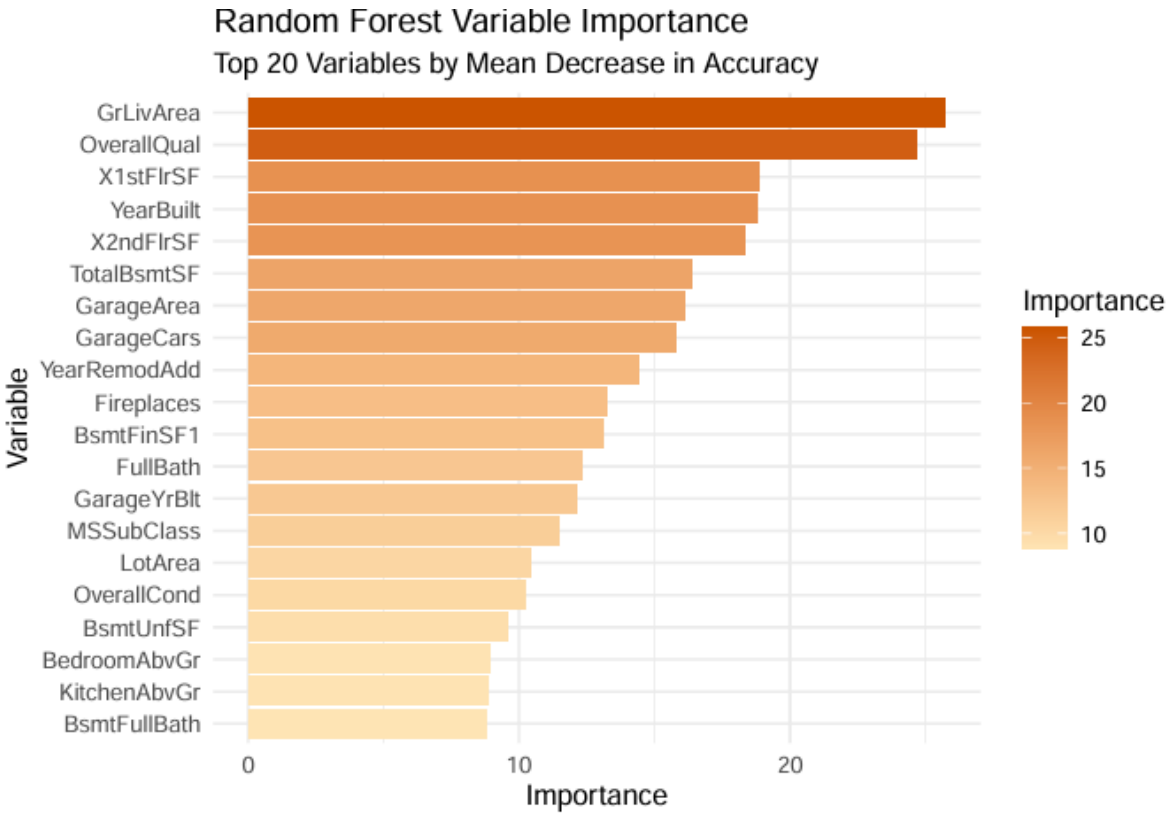
This plot shows predicted versus actual log sale prices using only the significant principal components from the PCR model. The orange points closely follow the 45-degree line, suggesting that even with a reduced set of predictors, the model retains strong predictive performance. Compared to the full PCR model, this version offers a more parsimonious model with minimal loss in accuracy.

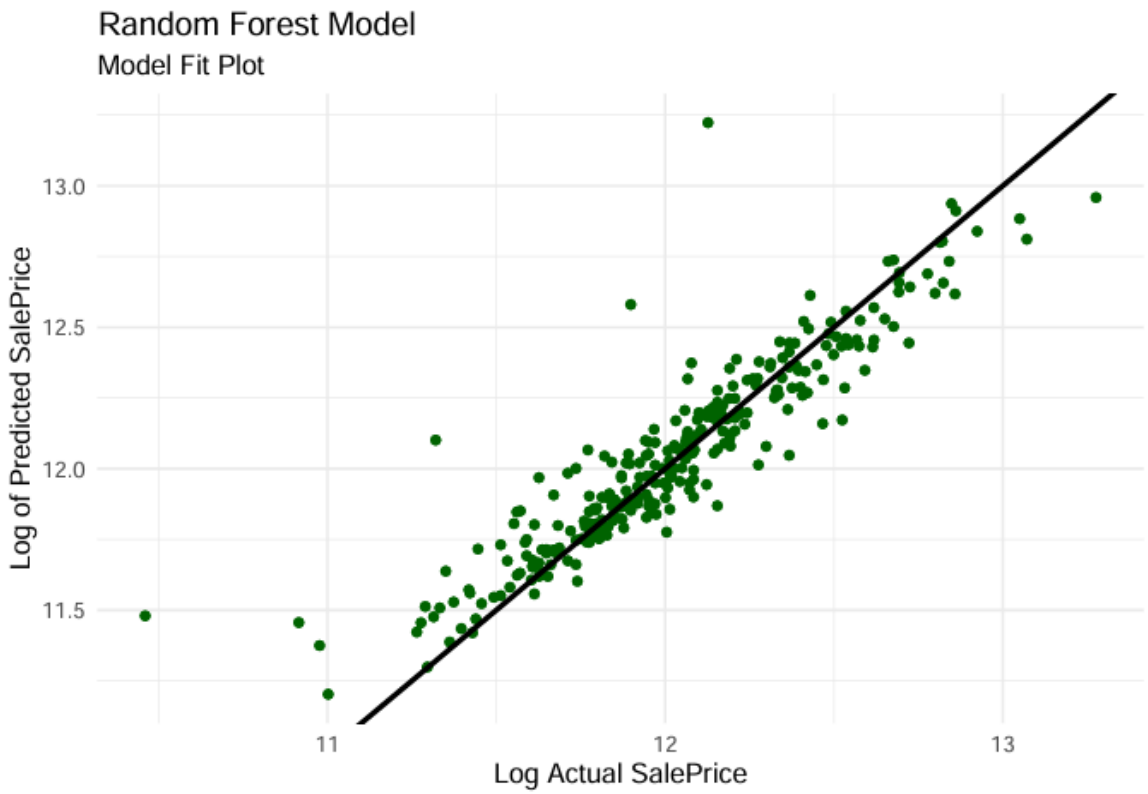


The residual plot for the PCR model using only significant components shows that the residuals are well-centered around zero, with no systematic pattern. The distribution is tight and symmetrical, indicating that the reduced model maintains the assumptions of homoscedasticity and linearity. The presence of a few mild outliers is expected but does not suggest model misspecification.

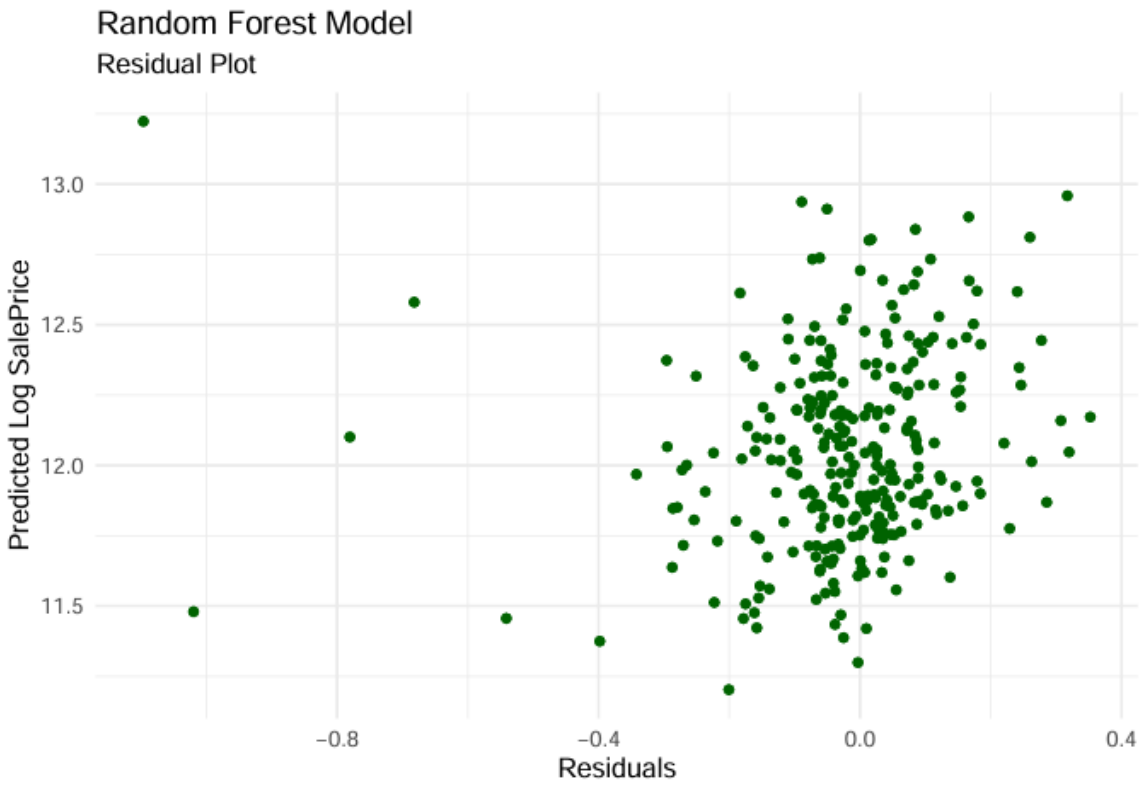
Random Forest

A random forest model was fitted using all numerical variables with 500 trees and optimal mtry. The model achieved an RMSE of 0.1628. Variable importance plots confirmed that the top contributors aligned with the strongest predictors found in the correlation analysis, including *GrLivArea*, *OverallQual*, and *GarageArea.*


Random Forest Variable Importance
Top 20 Variables by Mean Decrease in Accuracy
GrLivArea
OverallQual
X1stFlrSF
YearBuilt
X2ndFlrSF
TotalBsmtSF
GarageArea
GarageCars
YearRemodAdd
Fireplaces
BsmtFinSF1
FullBath
GarageYrBlt
MSSubClass
LotArea
OverallCond
BsmtUnfSF
BedroomAbvGr
KitchenAbvGr
BsmtFullBath
Variable
0
10
20
Importance
Importance
25
20
15
10



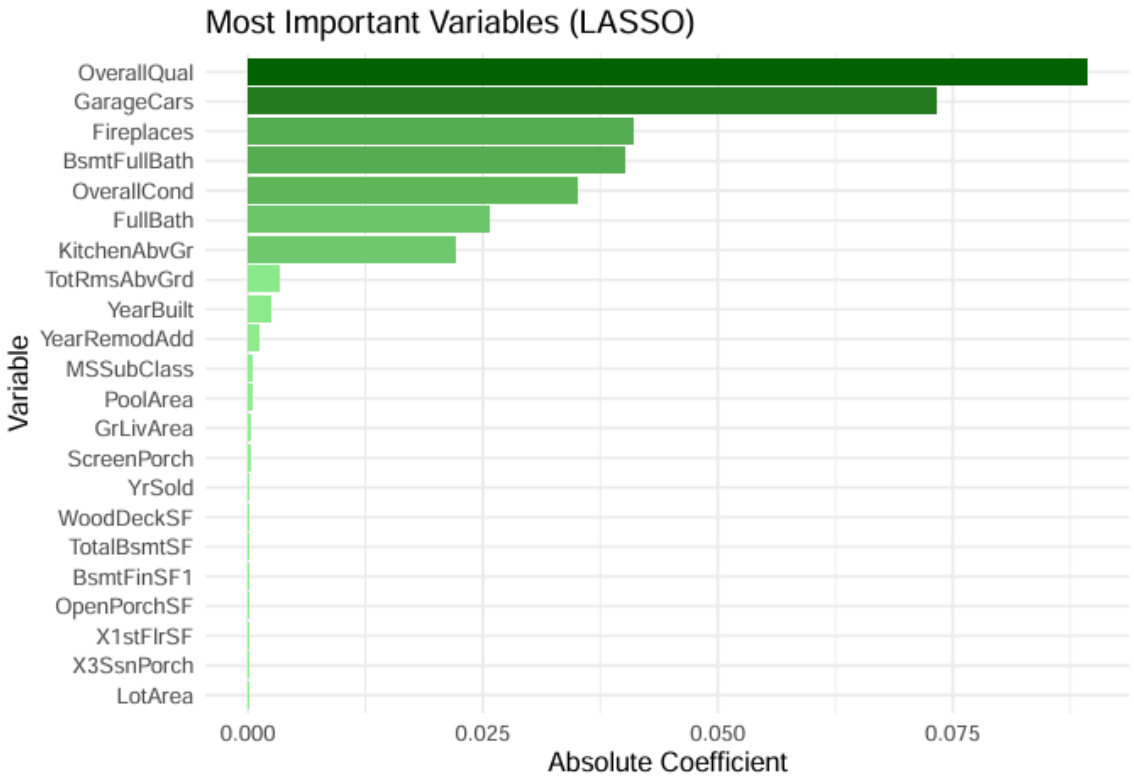
The fit plot for the Random Forest model shows a strong alignment between predicted and actual log-transformed sale prices. Most green points cluster closely around the diagonal reference line, indicating high predictive accuracy. There is slightly more variation at the extremes, but overall, the model performs well, capturing nonlinear relationships that linear models may miss.

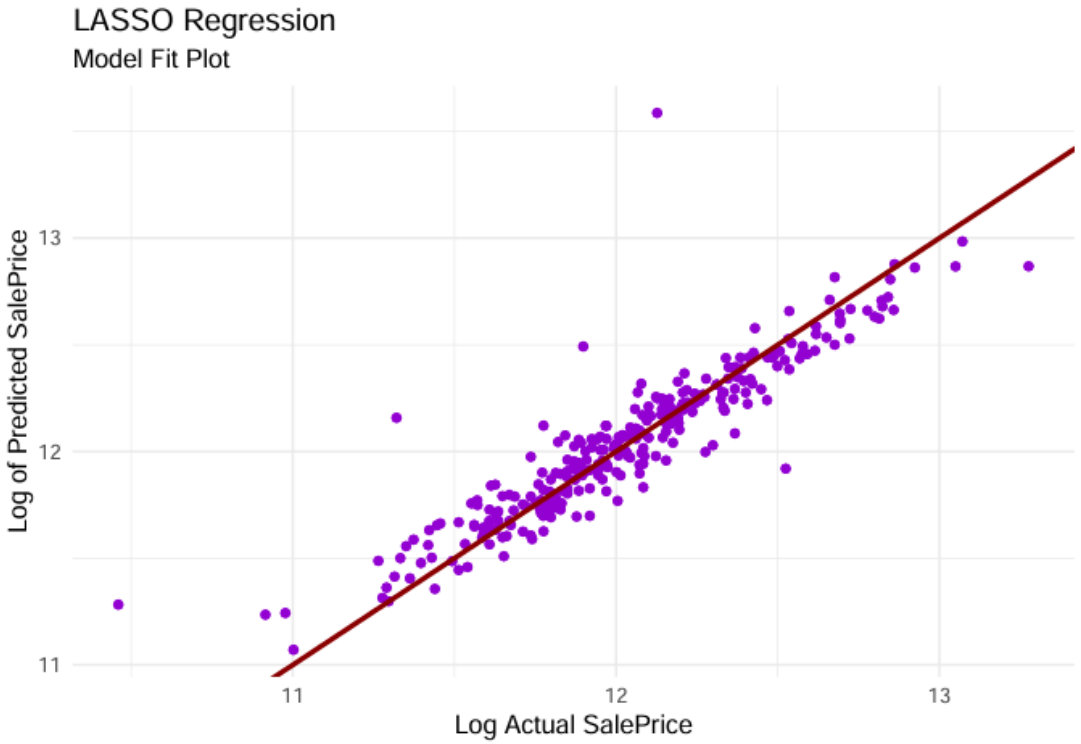


The residual plot for the Random Forest model displays a mostly symmetric and centered distribution of residuals around zero, with no evident pattern. This suggests that the model's predictions are unbiased across the range of predicted log sale prices. A few larger negative residuals indicate underprediction for some homes, but overall, the residual spread is tight, supporting the model's strong performance and good generalization.

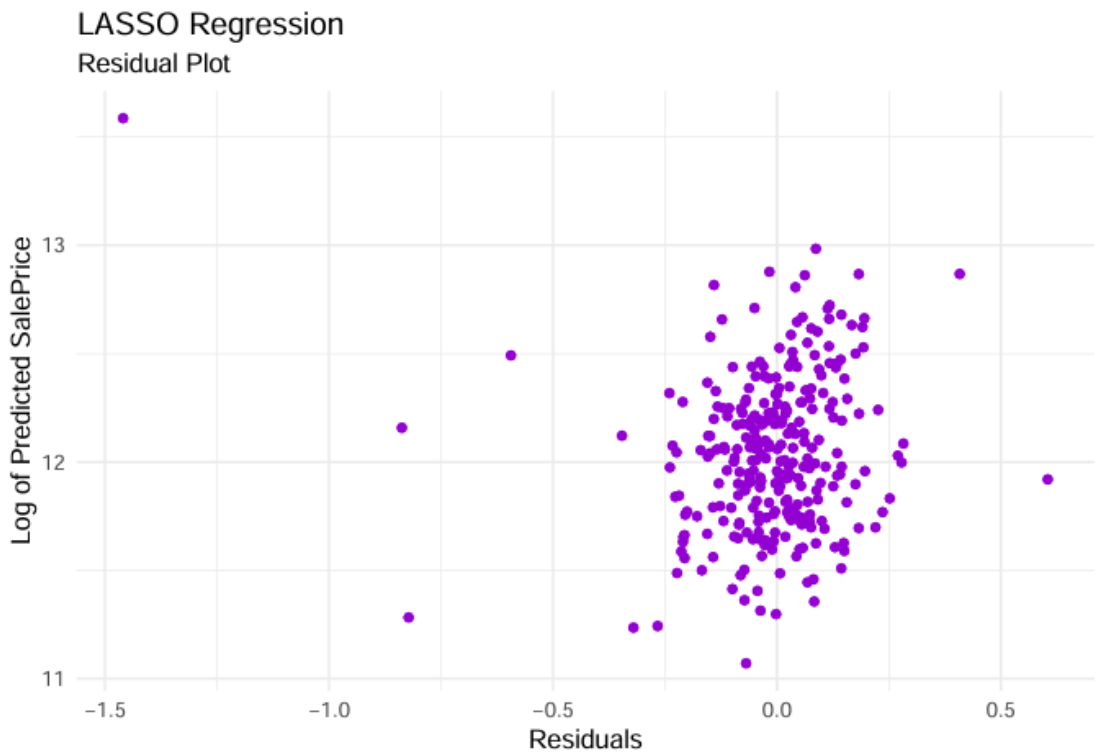
LASSO Regression

Lasso regression was employed using log-transformed sale prices to linearize relationships, stabilize variance, and produce approximately normal residuals—an essential assumption for linear modeling. The method minimizes the residual sum of squares subject to an L1 penalty, effectively shrinking some coefficients to zero and facilitating variable selection. The optimal penalty parameter was identified via cross-validation. Lasso identified *OverallQual*, *GrLivArea*, and *GarageCars* as dominant predictors, while suppressing less informative variables. Notably, both Lasso and Random Forest models indicated that fireplaces had stronger predictive value than full bathrooms above ground, suggesting nuanced, non-intuitive associations in housing data. While Lasso trades a small increase in RMSE for model parsimony, it remains a strong candidate in high-dimensional settings due to its interpretability and robustness to overfitting. The log-transformed sale price was regressed on all numeric predictors, and the final RMSE was 0.1631.


Most Important Variables (LASSO)
OverallQual
GarageCars
Fireplaces
BsmtFullBath
OverallCond
FullBath
KitchenAbvGr
TotRmsAbvGrd
YearBuilt
YearRemodAdd
MSSubClass
PoolArea
GrLivArea
ScreenPorch
YrSold
WoodDeckSF
TotalBsmtSF
BsmtFinSF1
OpenPorchSF
X1stFlrSF
X3SsnPorch
LotArea
Variable
0.000
0.025
0.050
0.075
Absolute Coefficient


LASSO Regression
Model Fit Plot
Log of Predicted SalePrice
13
12
11
11
12
13
Log Actual SalePrice

The model fit plot for the LASSO regression shows a tight alignment between predicted and actual log-transformed sale prices, with most points falling close to the 45-degree line. This suggests that the LASSO model, which performs both variable selection and regularization, captures the key predictors effectively while maintaining good generalization. A few moderate outliers exist, but the overall fit is strong.

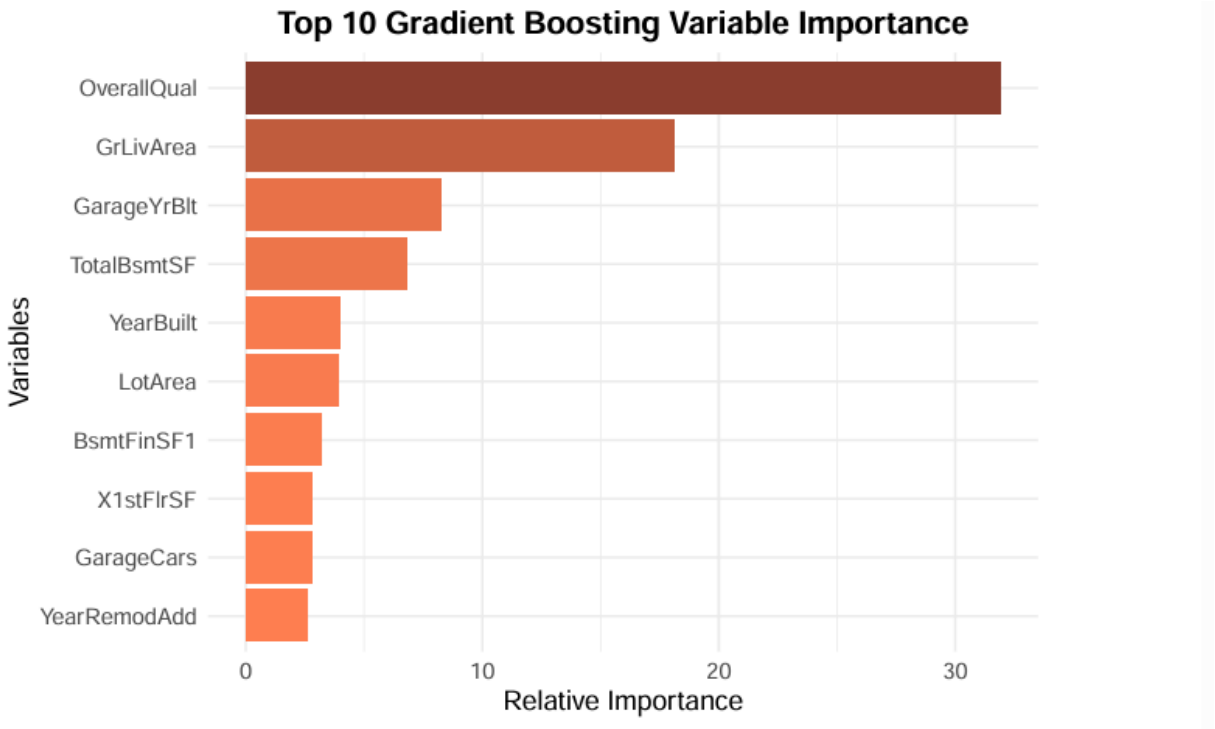


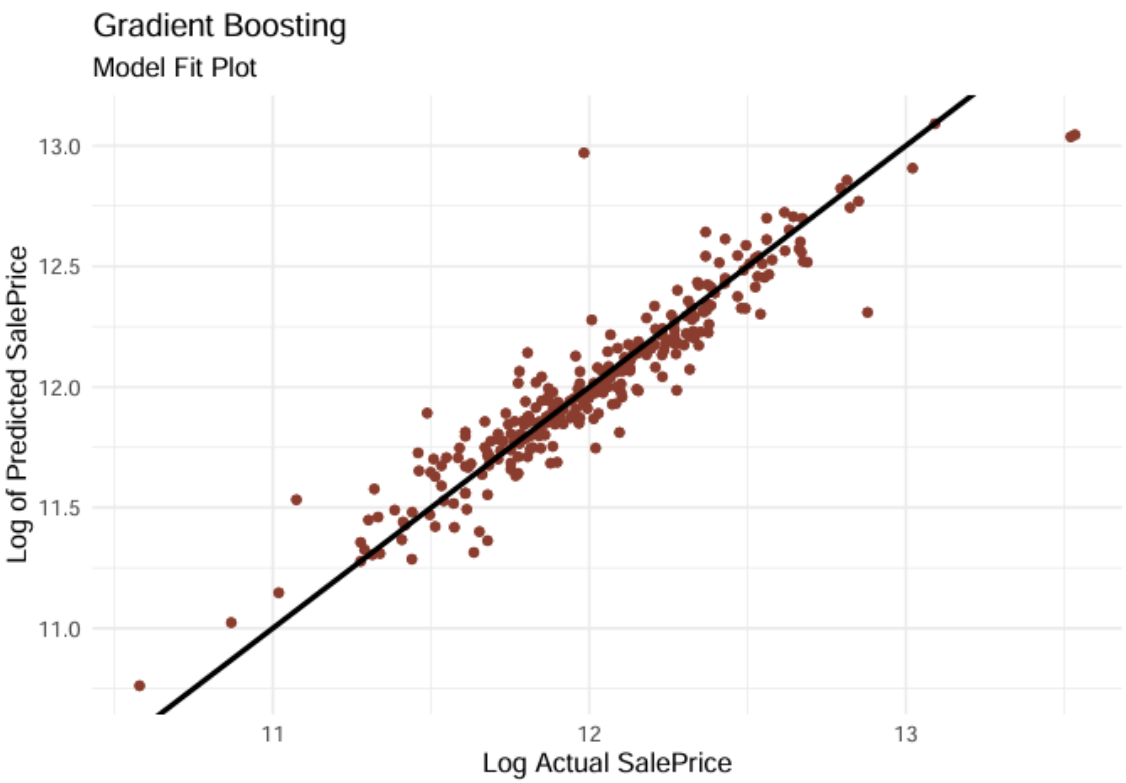
This residual plot for the LASSO regression model reveals a well-centered distribution of residuals around zero, indicating minimal bias in predictions. The residuals show no systematic trend, which supports the assumption of constant variance. Although a few extreme values appear on the left, the overall spread is moderate, suggesting the model performs reliably across most of the prediction range.

Gradient Boosting Trees Model

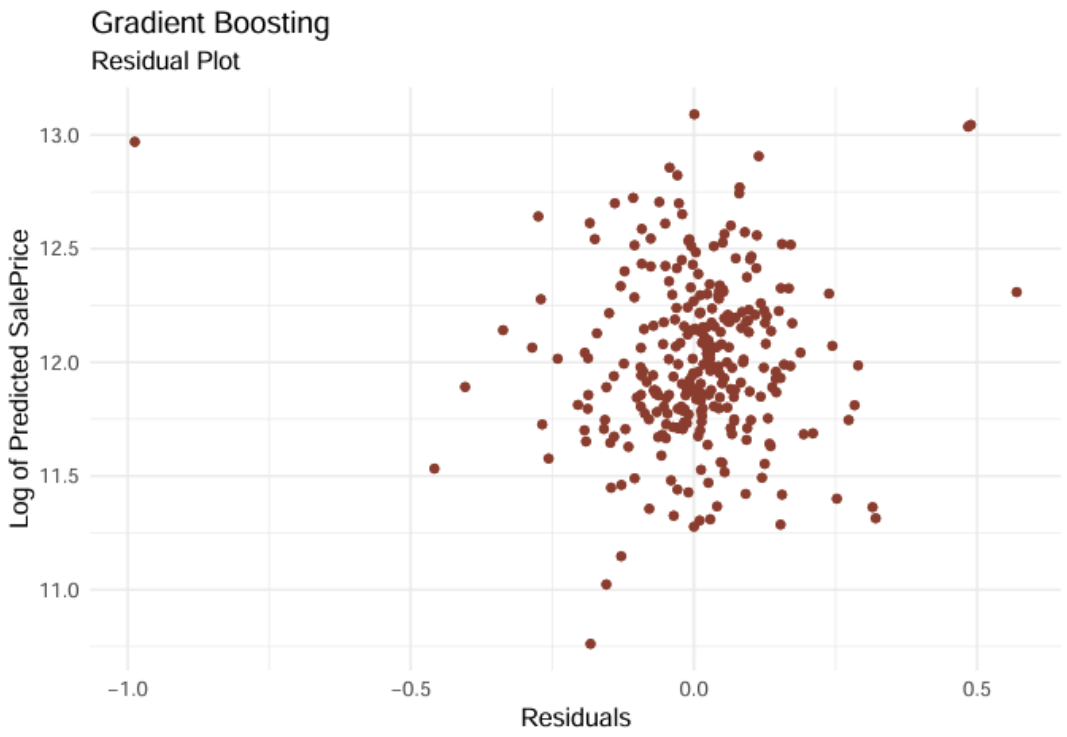
Gradient Boosting Trees demonstrated the strongest predictive performance, achieving the lowest RMS on the validation set. This ensemble method builds additive models in a forward stage-wise fashion by fitting weak learners (typically shallow trees) to the residuals of prior models, iteratively reducing prediction error. Boosting is particularly well-suited to high-dimensional, nonlinear problems like housing price prediction. The model's interpretability analysis revealed *OverallQual* and *GrLivArea* as the most influential predictors, while *LotArea*, *YearBuilt*, and *OverallCond* were less impactful. RMSE consistently decreased across iterations,

confirming that the model generalized well without overfitting. Although complex and sensitive to hyperparameters, the Gradient Boosting model provided the best empirical fit and stability across all methods tested. The validation RMSE for GBM was 0.1358.





The model fit plot showed predicted values closely aligning with the actual log sale prices, indicating good performance. Residuals were centered around zero and showed no major signs of heteroscedasticity or systematic bias. Although GBM tends to be more prone to overfitting than Random Forests, cross-validation and a conservative learning rate helped maintain generalization.



This residual plot for the Gradient Boosting model shows residuals that are tightly clustered around zero, with no clear patterns or trends, indicating that the model does not exhibit systematic bias. The spread is slightly wider than in the Random Forest model, with a few more extreme residuals on both ends, suggesting that while GBM captures complex relationships effectively, it may be more sensitive to outliers. Overall, the residual distribution supports the model's reliability and predictive accuracy.

Conclusion and Model Comparison

This analysis compared multiple predictive modeling approaches—Principal Component Regression (PCR), Lasso Regression, Random Forests, and Gradient Boosting Trees—applied to log-transformed home sale prices to mitigate skewness and heteroskedasticity. Performance was evaluated using Root Mean Square (RMS) on the log scale of sale prices, providing a consistent metric for model comparison.

Table 1: Model Comparison: Validation RMSEs

Model	Validation RMSE
Principal Component Regression (20 PCs)	0.1738
Principal Component Regression (Significant PCs)	0.1762
Random Forest	0.1628
LASSO Regression	0.1631
Gradient Boosted Trees	**0.1358**

- PCR yielded RMSE values of 0.1738 (20 components) and 0.1762 (significant PCs), demonstrating solid but limited flexibility in capturing nonlinear effects.
- Lasso Regression produced a low RMSE (0.1631) while enforcing sparsity, highlighting high-importance features and revealing that fireplaces are more predictive than the count of full bathrooms above ground.
- Random Forests improved upon linear models with an RMSE of 0.1628, benefiting from automatic variable interaction handling.
- Gradient Boosting Trees achieved the lowest RMSE of 0.1358, with the most stable and consistent performance. It captured the complex structures of the data and emphasized that *OverallQual* and *GrLivArea* are the primary predictors of sale price.

Overall, Gradient Boosting offered the most accurate predictions in this high-dimensional context. However, its complexity underscores the importance of model validation and careful tuning. Simpler models like Lasso provide valuable interpretability but may underperform in the presence of nonlinearities and interactions. Finally, limitations such as the inability to robustly detect outliers—due to market forces—and risks associated with ad hoc variable selection in sensitive models must be acknowledged. This reinforces the importance of applying rigorous statistical methodology alongside domain knowledge in predictive modeling.

References

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis* (6th ed.). Pearson Prentice Hall.

Weisberg, S. (2014). *Applied linear regression* (4th ed.). Wiley.

Montoya, A., & DataCanary. (2016). *House prices – advanced regression techniques*. Kaggle. https://kaggle.com/competitions/house-prices-advanced-regression-techniques