**House Price Prediction in Ames, Iowa**

**Executive Summary**

This report presents a predictive model for house prices in Ames, Iowa, to assist a company in purchasing, renovating, and reselling residential properties. Using the Ames dataset (1460 properties, 80 predictors), we conducted exploratory data analysis (EDA), preprocessing, and regression modeling. The LASSO regression model achieved the best performance with a Test RMSE of 0.1286 (log scale) and Adjusted R-squared of 0.8000, indicating accurate price predictions. Key predictors include Above-Grade Living Area (18.70% price increase per standard deviation), Overall Quality, and premium neighborhoods (e.g., Stone Brook, Northridge Heights). We recommend targeting high-quality properties in desirable neighborhoods, prioritizing renovations to kitchen and exterior quality, and using the model for accurate price estimation to maximize profit margins. Multicollinearity and residual diagnostics suggest areas for model refinement.

**Introduction**

The objective of this project is to develop a model to predict house prices in Ames, Iowa, and identify factors influencing prices to support strategic decisions for a real estate company. The Ames dataset contains 1460 properties with 80 predictors, including numerical (e.g., Lot Area, Total Basement Square Feet) and categorical variables (e.g., Neighborhood, House Style). We applied regression techniques from the ISOM5610 course, including multiple linear regression, quadratic regression, and LASSO regression, to model the log-transformed SalePrice, addressing skewness and multicollinearity. This report details the analysis process, model performance, and business recommendations.

**Analysis and Results**

**Exploratory Data Analysis**

EDA revealed significant skewness in SalePrice (mean $180,921, median $163,000) and predictors like GrLivArea and LotArea, justifying log-transformation. Strong correlations with SalePrice were observed for OverallQual (0.79), GrLivArea (0.71), and TotalBsmtSF (0.61). Missing values were prevalent in LotFrontage (17.74%) and categorical variables like FireplaceQu (47.26%), handled via median imputation and "None" replacement, respectively. Outliers (e.g., 61 in SalePrice) were noted, and moderate multicollinearity (VIF ~3.5–3.6 for TotalBsmtSF, 1stFlrSF) was identified, informing preprocessing strategies. See Appendix Figures 1–2 for visualizations.

**Preprocessing**

The dataset was preprocessed by:

* Imputing numerical missing values (e.g., LotFrontage) with medians.
* Replacing categorical missing values (e.g., FireplaceQu, GarageType) with "None."
* Dropping high-missingness columns (PoolQC, MiscFeature, Alley, Fence).
* Encoding ordinal variables (e.g., KitchenQual: Poor=1 to Excellent=5) and nominal variables (e.g., Neighborhood) as dummy variables.
* Log-transforming SalePrice, GrLivArea, LotArea, and TotalBsmtSF to address skewness.
* Creating features: TotalSF (TotalBsmtSF + 1stFlrSF + 2ndFlrSF), Age (YrSold - YearBuilt), and Qual\_GrLivArea (OverallQual × Log\_GrLivArea).
* Standardizing numerical predictors.

The processed dataset has 1460 rows and 227 columns, split into 70% training (1022 rows) and 30% testing (438 rows) sets.

**Modeling**

Three regression models were developed:

* **Multiple Linear Regression**: Included all predictors, capturing linear relationships.
* **Quadratic Regression**: Added squared terms for Log\_GrLivArea and Log\_TotalBsmtSF to model non-linearity.
* **LASSO Regression**: Used cross-validation to select the optimal penalty parameter (alpha = 0.000655), shrinking insignificant coefficients to zero.

Models were evaluated using Test MSE, RMSE, Adjusted R-squared, and 5-fold cross-validation RMSE. Residual diagnostics (Breusch-Pagan, Shapiro-Wilk) and VIF were used to check assumptions.

**Model Evaluation**

The model comparison (Table 1) shows LASSO regression as the best performer, with a Test RMSE of 0.1286 and Adjusted R-squared of 0.8000, balancing accuracy and sparsity. The linear (Test RMSE = 0.1580) and quadratic (Test RMSE = 0.1539) models performed worse, confirming LASSO’s superiority in reducing overfitting via variable selection. Residual diagnostics indicate potential issues: Breusch-Pagan p-value (0.0000) rejects homoskedasticity, and Shapiro-Wilk p-value (0.0000) rejects normality, suggesting residual patterns not fully captured by the model. VIF values (Table 3) show high multicollinearity for OverallQual (441.80) and Qual\_GrLivArea (565.00), likely due to their interaction term, indicating a need for further feature refinement. See Appendix Figure 3 for LASSO residual plots.

**Table 1: Model Comparison**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Train MSE | Test MSE | Train RMSE | Test RMSE | Train Adj. R-squared | Test Adj. R-squared | 5-Fold CV RMSE |
| Linear Regression | 0.0083 | 0.0250 | 0.0909 | 0.1580 | 0.9316 | 0.6981 | 0.1797 |
| Quadratic Regression | 0.0083 | 0.0237 | 0.0909 | 0.1539 | 0.9316 | 0.7108 | 0.1817 |
| LASSO Regression | 0.0127 | 0.0165 | 0.1125 | 0.1286 | 0.8953 | 0.8000 | 0.1601 |

**Key Findings**

LASSO selected 98 predictors (Table 2), including Log\_GrLivArea, OverallQual, TotalSF, and Neighborhood dummies (e.g., Neighborhood\_StoneBr, Neighborhood\_NridgHt). Coefficients indicate that a 1-standard-deviation increase in Log\_GrLivArea increases SalePrice by 18.70%, and Neighborhood\_StoneBr increases it by 11.37%. Premium neighborhoods (e.g., Stone Brook, Northridge Heights) significantly boost prices. The high VIF for OverallQual and Qual\_GrLivArea suggests multicollinearity, which LASSO mitigates but does not fully resolve.

**Table 2: Top LASSO Coefficients**

|  |  |
| --- | --- |
| Feature | Coefficient |
| Log\_GrLivArea | 0.1870 |
| Neighborhood\_StoneBr | 0.1137 |
| Exterior1st\_BrkFace | 0.1017 |
| Neighborhood\_NoRidge | 0.0935 |
| Qual\_GrLivArea | 0.0878 |

**Table 3: Variance Inflation Factors (Linear Regression)**

|  |  |
| --- | --- |
| Predictor | VIF |
| Log\_GrLivArea | 13.5866 |
| Log\_LotArea | 1.3197 |
| Log\_TotalBsmtSF | 1.7690 |
| TotalSF | 7.6462 |
| Age | 1.5771 |
| OverallQual | 441.7955 |
| Qual\_GrLivArea | 564.9997 |

**Conclusion**

The LASSO regression model effectively predicts house prices, with key predictors being Above-Grade Living Area, Overall Quality, Total Square Footage, and premium neighborhoods (e.g., Stone Brook, Northridge Heights). The company should target properties with high overall quality and large living areas in desirable neighborhoods. Renovations should focus on improving kitchen and exterior quality to maximize resale value. The model’s Test RMSE of 0.1286 supports accurate price estimation for profitable purchasing and selling decisions. However, multicollinearity (high VIF for OverallQual, Qual\_GrLivArea) and residual diagnostics (Breusch-Pagan and Shapiro-Wilk p-values = 0.0000) suggest limitations, potentially due to omitted variables (e.g., school quality). Future analyses should explore additional features or alternative models (e.g., ridge regression) to improve robustness.

**References**

* Pochiraju, B., & Seshadri, S. (2019). *Essentials of Business Analytics: An Introduction to the Methodology and its Applications*. Springer.

**Appendix**

**EDA Visualizations**

*Figure 1: Distribution and Boxplot of SalePrice*  
A graph of sales and sales

AI-generated content may be incorrect.

*Figure 2: SalePrice by Neighborhood*  
A graph of a number of houses

AI-generated content may be incorrect.

**Model Outputs**

*Figure 3: LASSO Regression Residual Plots*  
A graph with blue dots

AI-generated content may be incorrect.