Housing Price Prediction Analysis

Kaleb Jordan

Introduction

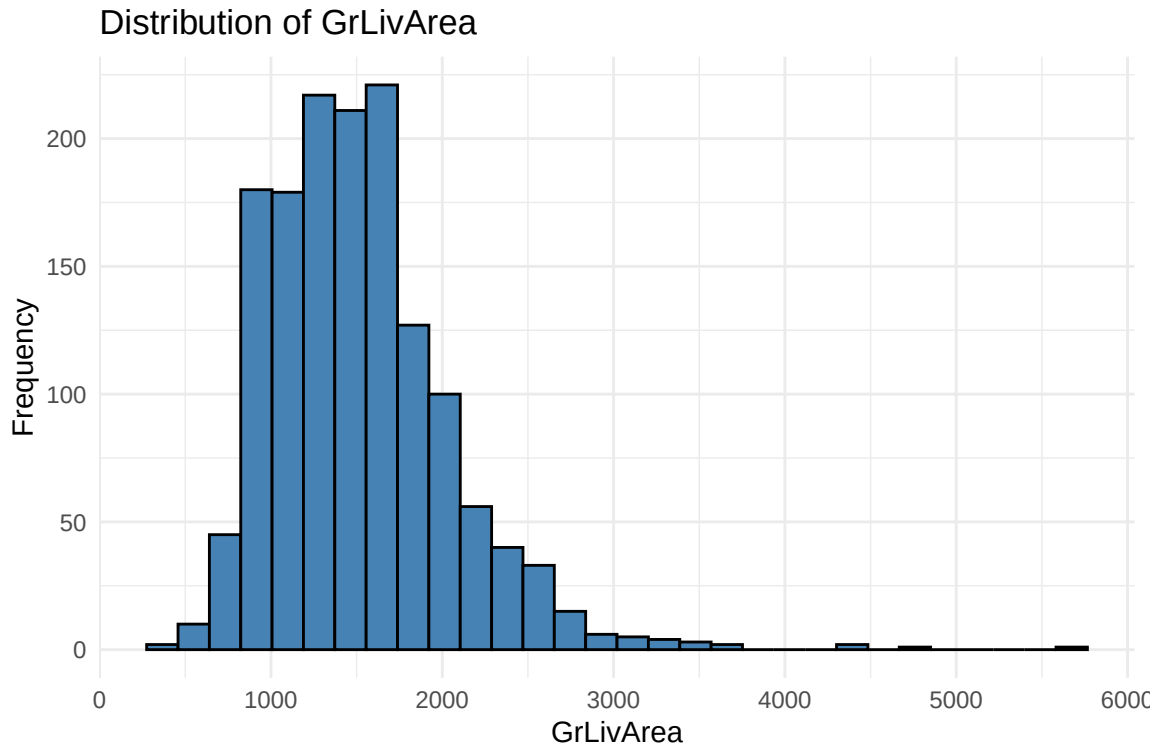
Accurately predicting housing prices is essential for real estate professionals, investors, and homeowners by analyzing housing data from the Ames Housing Data Set. This analysis utilizes a Random Forest model to predict housing prices, addressing data challenges such as skewed distributions and missing values. The approach is robust and demonstrates potential applications beyond housing data.

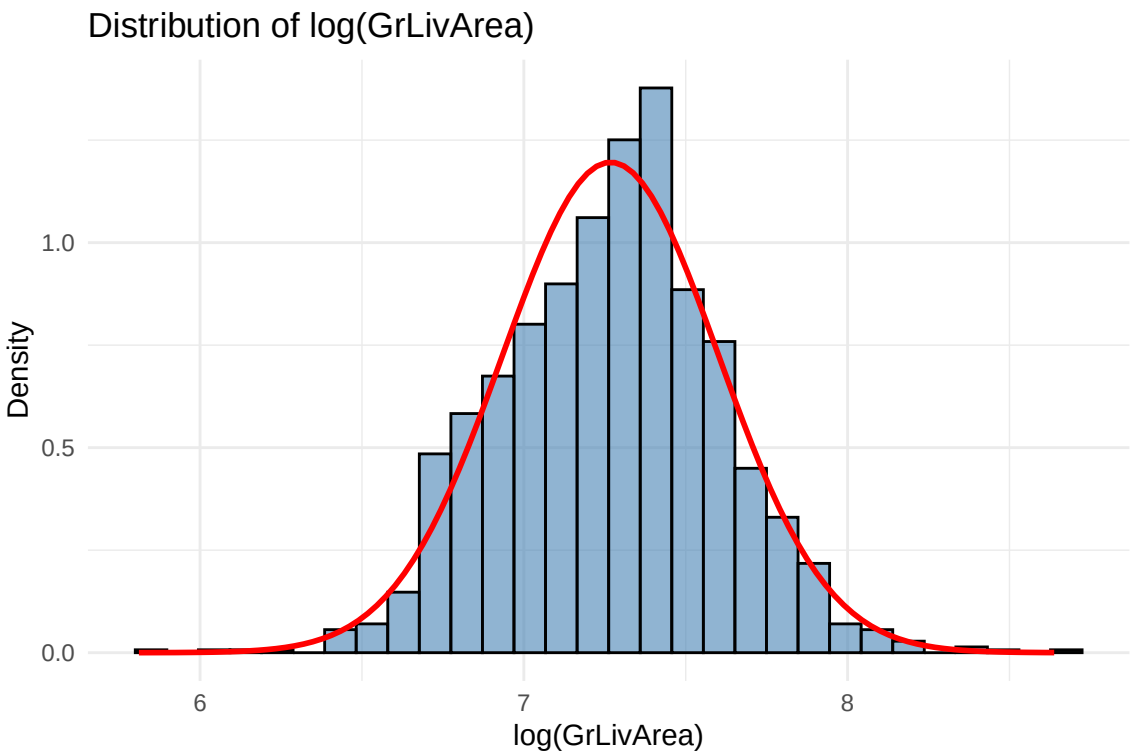
Data Overview

The analysis used two data sets: a training set with known housing prices and a test set for predictions. Key pre-processing steps included:

- Identifying and correcting skewed numerical variables.
- Imputing missing values using the mean for numeric variables and the mode for categorical variables.
- Applying logarithmic transformations to reduce skewness and improve model performance.

For example, the graph below shows before and after the transformation for **GrLivArea**:





Methodology

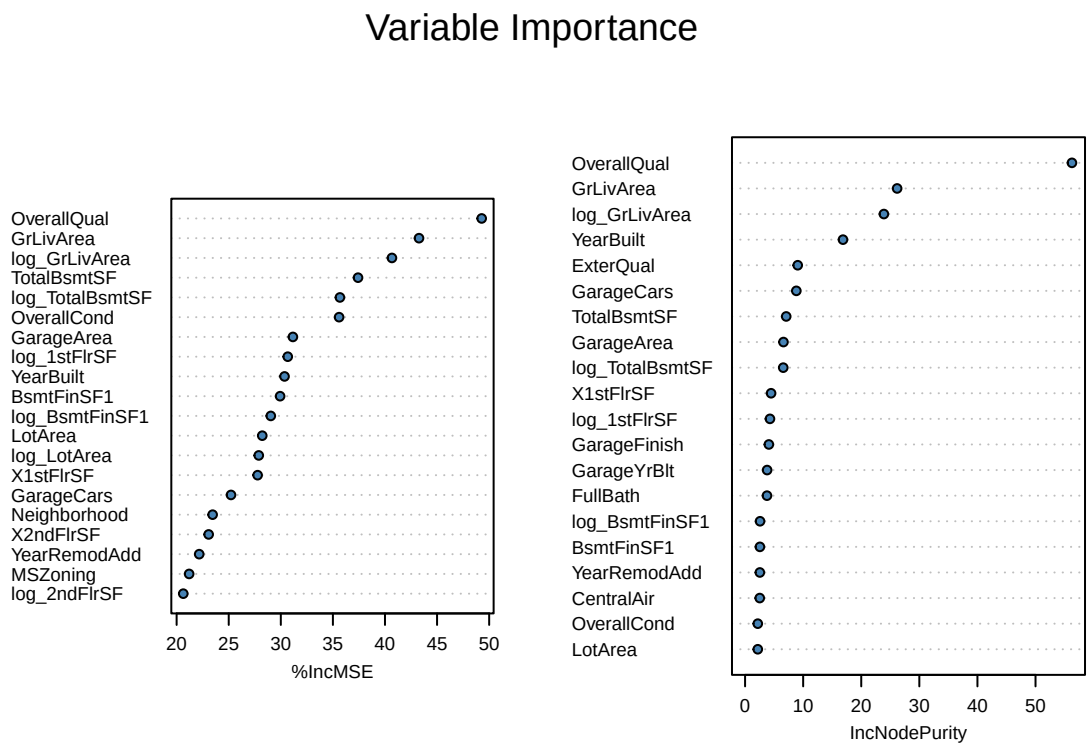
A Random Forest regression model was trained to predict housing prices. This method offers several advantages:

- Handling high-dimensional data sets by selecting the best features.
- Capturing non-linear relationships between features and the target variable.
- Robustness to outliers and missing data.

The model was tuned with:

- **Number of Trees**: 1,700 trees to ensure sufficient decision trees for stable and reliable predictions.
- **Number of Variables at Each Split (mtry)**: 35 variables considered at each split, optimized based on cross-validation results.

Variable Importance



Top Predictors:

- **Overall Quality (OverallQual)**: This variable ranks as the most important predictor, emphasizing the crucial role of overall construction quality in determining house prices.
- **GrLivArea (Above Grade Living Area) and TotalBsmtSF (Total Basement Area)**: Size-related features significantly influence the sale price.
- **YearBuilt (Year Built) and YearRemodAdd (Year of Remodel)**: Modernity indicators show that buyers often prefer newer or updated properties.

Log-Transformed Features:

- Several variables, such as **log__GrLivArea** and **log__TotalBsmtSF**, show high importance. Handling skewed distributions through transformations improves the model's predictive power, making these variables more normally distributed and contributing better to the model's performance.

Garage and Basement Features:

- Variables like **GarageArea** and **BsmtFinSF1** underscore the importance of functional spaces in housing.

Other Variables:

- Location features like **Neighborhood** are critical in reflecting regional price differences.
- Features like **CentralAir** are crucial for capturing home amenities preferred by potential buyers.

Practical Implications

- **Model Enhancements**: By focusing on these high-impact features, future models can prioritize key variables to improve performance and interpretability.
- **Real Estate Analysis**: Realtors and property assessors can use this feature ranking to appraise property values more effectively by focusing on top predictors like **Overall Quality** and **square footage**.
- **Broader Applications**: This ranking can be leveraged for applications such as urban planning, property investments, or predictive analytics in similar housing data sets.

Results

The model achieved a high explanatory power, explaining 88.26% of the variance in housing prices in the training data set. Key predictors identified include:

- Overall quality of the house.
- Total square footage of the property.
- Neighborhood characteristics.

Prediction Performance

Predicted prices for the test set were back-transformed to their original scale, resulting in realistic and actionable estimates. A brief summary of the results (5 of the 1459):

Housing ID	Square Footage (GrLivArea)	Bedrooms (BedroomAbvGr)	Bathrooms (FullBath + HalfBath)	Predicted Sale Price (SalePrice)
1461	896	2	1	$124,722
1462	1329	3	1.5	$153,006
1463	928	3	2.5	$180,575
1464	1629	3	2.5	$182,851
1465	1604	2	2	$193,117

Insights

Lessons Learned

- **Data Pre-processing**: Addressing skewness and missing values before modeling is critical to obtaining accurate predictions. Transforming skewed variables using logs greatly improved the model's ability to handle outliers.
- **Feature Importance**: Understanding the features that contribute most to housing prices allows for targeted interventions in pricing and property assessments.
- **Robust Models**: Random Forest excels in capturing complex non-linear relationships between predictors and target variables.

Future Directions and Suggestions

Advanced Feature Engineering While the current model identified important features like **Overall Quality** and **GrLivArea**, further improvements can be made by:

- **Interaction Terms**: Exploring interaction features, such as **OverallQual * GrLivArea**, to uncover potentially meaningful relationships between variables.
- **Non-Linear Transformations**: Further non-linear transformations like applying square root or exponential transformations could reveal complex patterns not evident in raw features.

Generalization Testing To ensure robustness and model applicability:

- **Cross-Validation**: Utilize k-fold cross-validation to reduce overfitting and better assess model performance across different data subsets.
- **Out-of-Sample Testing**: Test the model on data sets from different regions or real estate markets to assess its effectiveness outside of the training region.

Practical Applications The insights derived from this analysis can be adapted into actionable tools:

- **Real Estate Advisors**: Provide agents and homeowners with an easy-to-use tool that estimates property values based on important predictors, allowing them to make informed decisions.
- **Property Developers**: Use the findings to prioritize property features (such as square footage, basement area, and age) that offer the best returns on investment.

Conclusion

This analysis demonstrates the effectiveness of Random Forest in predicting housing prices. The methodology and findings can be adapted to other fields that require accurate predictions from structured data sets. By focusing on data pre-processing and interpreting the results, similar approaches can be leveraged to inform decision-making in various industries, not just real estate.