Ames Housing Price Analysis

Project Overview

This project analyzes the Ames Housing dataset to predict house sale prices using various regression techniques. The analysis includes exploratory data analysis, feature engineering, and comparison of multiple machine learning models.

Dataset

Ames Housing Dataset - Contains detailed information about residential properties in Ames, Iowa, including various features that influence house prices.

Project Objectives

- Explore and visualize housing data to identify key patterns
- Clean and preprocess data for analysis
- Build predictive models to estimate house sale prices
- Compare performance of different regression models
- Identify the most important features affecting house prices

Technologies Used

- **R** (Version: R 4.x or higher recommended)
- **Libraries:**
  - tidyverse
  - ggplot2
  - caret
  - dplyr
  - corrplot
  - glmnet
  - gridExtra
  - randomForest
  - rpart

Key Features Analyzed

- Ground Living Area (Gr Liv Area)

- Ground Living Area (Gr.Liv.Area)
- Overall Quality (Overall.Qual)
- Garage Cars
- Garage Area
- Total Basement Square Footage (Total.Bsmt.SF)

Analysis Steps

1. Data Loading and Exploration

- Load the Ames Housing dataset
- Examine data structure and summary statistics
- Identify missing values

2. Data Cleaning

- Impute missing values (e.g., Lot Frontage with median)
- Remove outliers (houses > 4000 sq ft living area)
- Handle categorical variables

3. Exploratory Data Analysis (EDA)

- Distribution analysis of sale prices
- Correlation heatmap of numerical variables
- Scatter plots showing relationships between features and sale price

4. Feature Engineering

- Convert categorical variables to factors
- Encode ordinal variables
- Select relevant features for modeling

5. Model Building and Comparison

Three regression models were implemented and compared:

- **Linear Regression**: Simple interpretable model
- **Decision Tree**: Non-linear regression approach
- **Random Forest**: Ensemble method for improved accuracy

6. Model Evaluation

- Models evaluated using Root Mean Squared Error (RMSE)
- Feature importance analysis using Random Forest
- Diagnostic plots for model validation

Key Visualizations

1. **Distribution of Sale Prices** - Histogram showing price distribution
2. **Correlation Heatmap** - Relationships between numerical features
3. **Scatter Plots** - Feature vs. Sale Price relationships
4. **Residual Plots** - Model diagnostic visualizations
5. **Q-Q Plot** - Normality check for residuals
6. **Feature Importance Plot** - Top predictors identified by Random Forest
7. **Model Comparison Chart** - RMSE comparison across models

Results

- Best-fit line equation for simple linear regression
- RMSE values for each model
- Feature importance rankings
- Model performance comparison

How to Run

1. **Install Required Packages:**

```r
install.packages(c("tidyverse", "ggplot2", "caret", "dplyr",
                   "corrplot", "glmnet", "gridExtra",
                   "randomForest", "rpart"))
```

2. **Download the Dataset:**
   - Obtain the AmesHousing.csv file
   - Update the file path in the script to match your local directory
3. **Run the Script:**
   - Open the R script in RStudio or your preferred R environment
   - Execute the entire script or run sections sequentially

File Structure

```
Project-AmesHousing/
├── analysis.R          # Main R script with all analysis code
├── report.pdf          # Detailed project report
├── report.docx         # Editable report document
└── README.md           # This file
```

Future Improvements

- Implement additional models (XGBoost, Ridge/Lasso Regression)
- Perform hyperparameter tuning
- Create interactive visualizations
- Add cross-validation for more robust evaluation

Author

College Project - Data Analysis Course

Date

2024-2025 Academic Year