Final Project Statistical Methods in Data Science II

Ethan Cohen-Cole, Ph.D.
Professor of Economics
Department of Economics
University of Maryland, College Park

Table of Contents

1. Data Cleaning
2. Data Exploration
   2.1. Correlation Matrix
   2.2. Scatterplots
   2.3. Boxplots
   2.4. Density Plots
   2.5. Pairwise Interactions Scatterplots
   2.6. Pairwise Interactions Boxplots
   2.7. Pairwise Interactions Density Plots
3. Data Preprocessing
   3.1. Dummy Variables
   3.2. Standardization
   3.3. Log Transformation
   3.4. Interaction Terms
   3.5. Polynomial Terms
   3.6. Final Dataset
4. Model Selection
   4.1. Forward Selection
   4.2. Backward Selection
   4.3. Stepwise Selection
   4.4. Lasso Regression
   4.5. Ridge Regression
   4.6. Elastic Net
   4.7. Principal Component Regression
   4.8. Partial Least Squares
   4.9. Regression Tree
   4.10. Random Forest
   4.11. Gradient Boosting
   4.12. Support Vector Machine
   4.13. Neural Network
   4.14. Ensemble Model
5. Model Evaluation
   5.1. Cross Validation
   5.2. Test Set Evaluation
   5.3. Residual Analysis
   5.4. Feature Importance
   5.5. Partial Dependence Plots
   5.6. Prediction Intervals
   5.7. Bootstrap Confidence Intervals
   5.8. Learning Curves
   5.9. Model Stability
   5.10. Final Model Comparison
6. Conclusion
   6.1. Summary of Findings
   6.2. Limitations
   6.3. Recommendations
   6.4. Future Work
7. Appendix: Conjugate Gradient Algorithm

1. Data Cleaning

```
# Load data
df <- read.csv("Housing.csv")

# Check structure
str(df)

# Check for missing values
colSums(is.na(df))

# Check for duplicate rows
sum(duplicated(df))

# Check unique values in categorical columns
cat("mainroad:", unique(df$mainroad), "\n")
cat("guestroom:", unique(df$guestroom), "\n")
cat("basement:", unique(df$basement), "\n")
cat("hotwaterheating:", unique(df$hotwaterheating), "\n")
cat("airconditioning:", unique(df$airconditioning), "\n")
cat("prefarea:", unique(df$prefarea), "\n")
cat("furnishingstatus:", unique(df$furnishingstatus), "\n")
```

```
'data.frame':	545 obs. of  13 variables:
 $ price           : int 13300000 12250000 12250000 12215000 11410000 10850000 10150000 10000000 9870000 9800000 9800000 9800
```