CSCI 550: Adv. Data Mining
Fall 2024

Project 2: Feature Selection and Regression
Due date: Nov 3, 2024

This project is designed for team collaboration, allowing groups of 2 to 4 members. Individual submissions are not permitted.

Dataset: Residential Sales Data for Cook County, IL

In this project, your goal is to specify and fit various regression models to predict house prices in Cook County. The dataset consists of 204,792 records from Cook County, Illinois, which includes Chicago. It contains 61 features, with the 62nd variable, Sale Price, being the target variable for prediction. Detailed explanations of each variable are available in the attached codebook.txt file.

Learning Objectives:

- **Data Cleaning and Exploratory Data Analysis (EDA):** Develop skills in understanding the structure of the data and performing necessary cleaning.
- **Feature Selection**: Gain proficiency in applying different feature selection techniques.
- **Regression Models:** Learn to implement both linear and non-linear regression methods.
- **Regularization Methods:** Understand and apply l1 (Lasso) and l2 (Ridge) regularization techniques.
- **Cross-Validation:** Implement and appreciate the importance of cross-validation for model tuning and selection.

Deliverables:

You may use either Python or R for this project. Your final submission should include:

- A Well-Formatted PDF Document: This document should reflect the contents of your Jupyter Notebook or Markdown file.
- Code File (Notebook or Markdown): This file should contain your inline code, neatly structured and free from unnecessary or error-prone code. Code that generates warnings is acceptable but should be clearly noted.

Both the PDF and code file will be assessed together for consistency. Ensure that all results in the PDF are supported by corresponding code.

Report Guidelines:

1. **Executive Summary:**

Audience: Assume your reader is a large-scale real estate investor in the Chicago area with little to no background in data mining or advanced statistics.

Content: Summarize the problem, highlight key findings, and discuss any limitations of the models. Provide practical, data-driven recommendations for future real estate investments.

Tone: Avoid technical jargon in this part and focus on clear, actionable insights. Keep the summary brief and accessible, emphasizing investment recommendations.

2. **Data Preprocessing and Exploratory Data Analysis:**

Data Cleaning: Document the steps taken to handle missing data, outliers, categorical variables, and irrelevant predictors.

Visualizations: Include well-labeled visualizations and provide interpretations that highlight significant trends or patterns.

Hypothesis Development: Based on your EDA, discuss any hypotheses regarding the importance of specific predictors.

3. **Model Development and Performance Evaluation:**

You will need to train and evaluate the following models:

- **Simple Regression:** Start with a simple regression model for baseline comparison.
- **Multiple Regression (Full Model):** Train a multiple regression model using all available features.
- **Subset Selection:** Apply a feature selection technique to choose a subset of variables that performs best on the test set.
- **Regularization:** Implement l_1 (Lasso) and l_2 (Ridge) regularization to reduce model complexity.
- **PCA:** Use Principal Component Analysis (PCA) to aid in model selection
- **Non-Linear Models:** Select a quantitative variable based on above models, and develop polynomial and smooth spline models for that variable.
- **Model Comparison:** Compare prediction errors across the different models and provide a thoughtful analysis of model performance.
- **Cross-Validation:** Use 5-fold cross-validation to select tuning parameters for each model and ensure robust performance evaluation.