# # DSC412-project_Tharun Mandadi

# Title: Iowa House Price Prediction Using ML Techniques

## Proposal

### Problem Statement

Accurately predicting house prices is a vital challenge for real estate professionals and prospective homeowners. The Iowa House Price Prediction project aims to address this challenge by leveraging machine learning techniques to develop a robust predictive model. The project focuses on understanding the relationship between property attributes and sale prices while addressing potential issues like multicollinearity and overfitting in the dataset.

### Proposed Solution

The project employs linear regression and two regularization techniques, Ridge and Lasso regression, to build predictive models. These methods help address multicollinearity and improve generalization. By optimizing hyperparameters such as the regularization strength, the models aim to achieve minimal mean squared error (MSE).

What makes this approach novel is its combination of a comprehensive feature engineering process and the comparison of regularization techniques to identify the most effective model. Additionally, this project seeks to evaluate the sensitivity of model predictions to different input features, providing actionable insights into the most influential factors affecting house prices.

### Potential Stakeholders

- Real estate agents and appraisers seeking accurate pricing models.
- Homebuyers and sellers aiming to understand fair market values.
- Financial institutions evaluating mortgage and loan risks.
- Researchers and data scientists exploring regression techniques in real-world applications.

### Potential Obstacles

- **Data Quality**: Handling missing values and ensuring consistency in the dataset.
- **Feature Engineering**: Selecting and transforming features effectively to enhance model performance.
- **Hyperparameter Tuning**: Finding the optimal regularization strength for Ridge and Lasso models.
- **Model Evaluation**: Ensuring that the selected model generalizes well to unseen data.

### Novelty

This project goes beyond standard regression analysis by combining regularization techniques with a rigorous feature engineering process. The use of both Ridge and Lasso regression allows for a detailed comparison of their effects on predictive accuracy and variable selection. Additionally, the project incorporates sensitivity analysis to quantify the impact of various property attributes on model predictions.

### Model Accuracy

The accuracy of the models will be evaluated using mean squared error (MSE) on validation and test datasets. Cross-validation techniques will ensure unbiased performance evaluation. Comparative analysis with baseline models (e.g., simple linear regression) will demonstrate the improvement achieved through regularization.

### Benchmarking

Similar models and studies exist in online repositories and research papers. These will be used as baselines to compare the performance and features of this project's models. Improvements will be documented, including enhanced feature selection, optimized hyperparameters, and lower error metrics.

## Plan

### Data Acquisition

- **Source**: Historical house price data from publicly available sources, including Kaggle's Ames Housing dataset.
- **Access Permissions**: Ensure the data is publicly available or properly licensed for use.

### Data Preprocessing

- **Cleaning Data**: Handle missing values using median imputation for numerical columns and mode for categorical columns.
- **Feature Engineering**: Select relevant features, including two categorical features (e.g., BsmtQual and Neighborhood) and 21 numerical features.
- **Scaling**: Standardize numerical features and target variable (SalePrice) to ensure comparability.

### Data Organization

Organize data into training, validation, and test subsets to ensure unbiased evaluation. Store cleaned and preprocessed data in a GitHub repository for reproducibility.

### Data Analysis

- Explore feature distributions and correlations.
- Identify multicollinearity among features using variance inflation factors (VIF).

### Model Selection

1. **Linear Regression**: Fit a baseline model using scikit-learn's `LinearRegression()`.
2. **Ridge Regression**: Implement with varying regularization strengths ($\lambda$) to minimize

MSE.
3. **Lasso Regression**: Use L1 regularization to perform variable selection and identify key predictors.

### Accuracy Metrics

- Evaluate models using MSE on validation and test datasets.
- Use cross-validation to ensure robustness.
- Benchmark results against simple linear regression as a baseline.

### Sensitivity Analysis

- Assess the impact of changes in key features (e.g., property size, location) on predicted prices.
- Quantify the relative importance of selected features using coefficients from Lasso regression.

### Public Documentation

Maintain a comprehensive project repository on GitHub. Include detailed README files, Jupyter notebooks, and results for public access.