Ames Housing Market

Dev Mudeppa

GA-DSB 02/16/2024

Outline

01 Objectives

02 Data cleaning

03 Preprocessing

04 Linear Regression

05 Poly, Ridge & Lasso

06 Summary

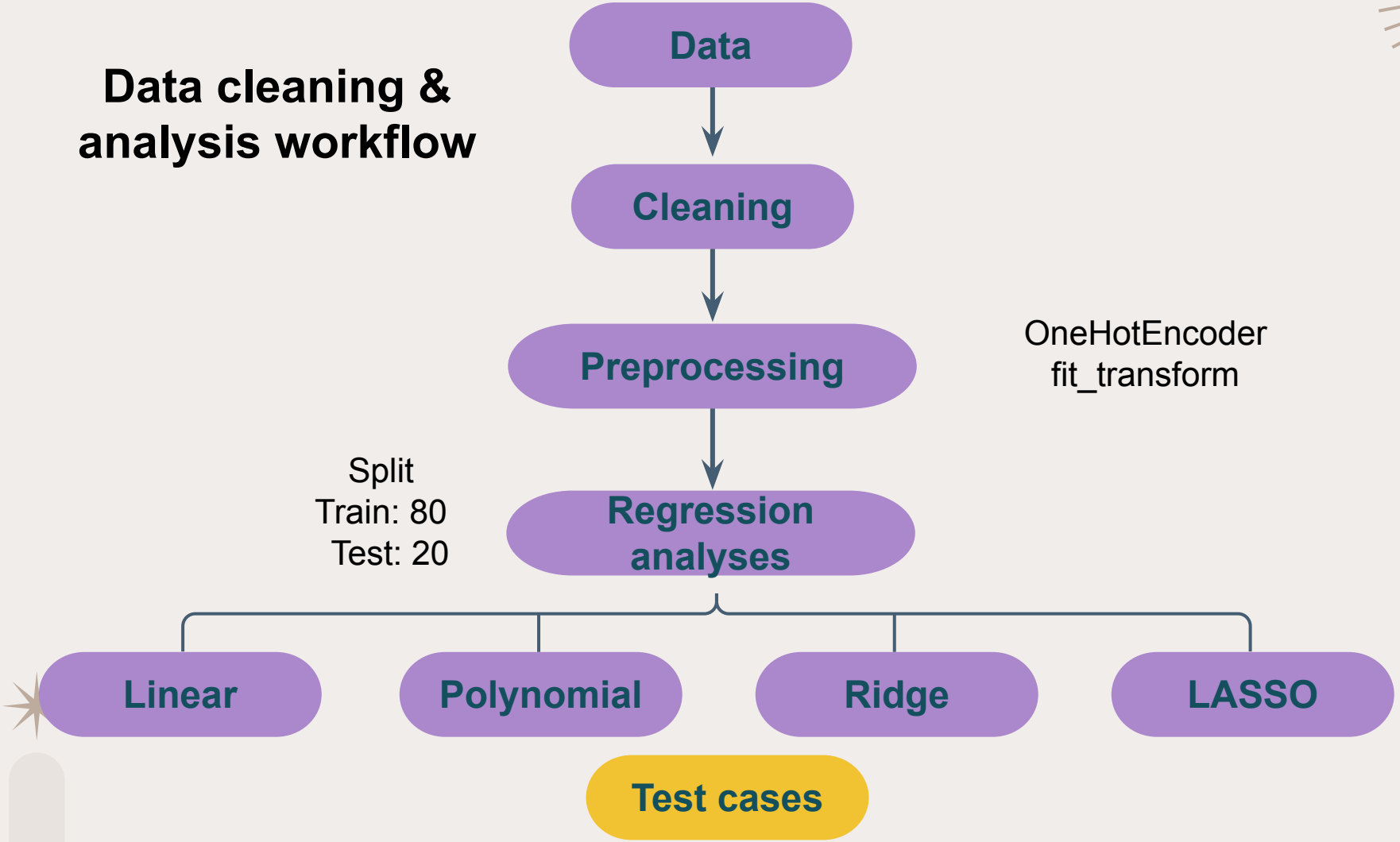
Objective

Predict the housing prices for test set based on the training data set

Build models of linear regressions from training data

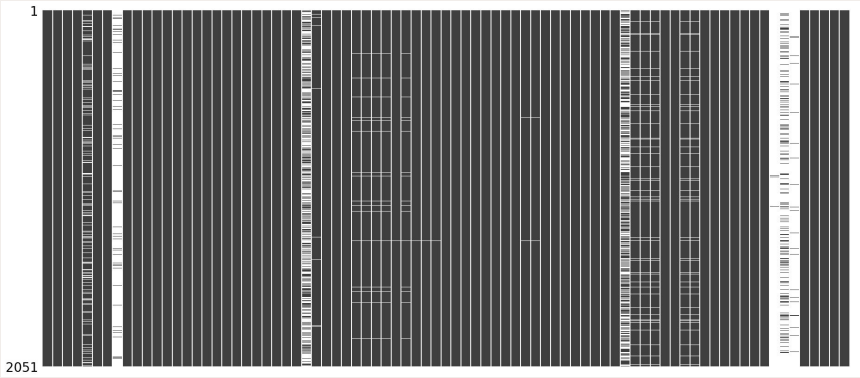
Predict the housing prices in Ames using built models

Data cleaning & analysis workflow



Data cleaning

shape : (2051, 81)



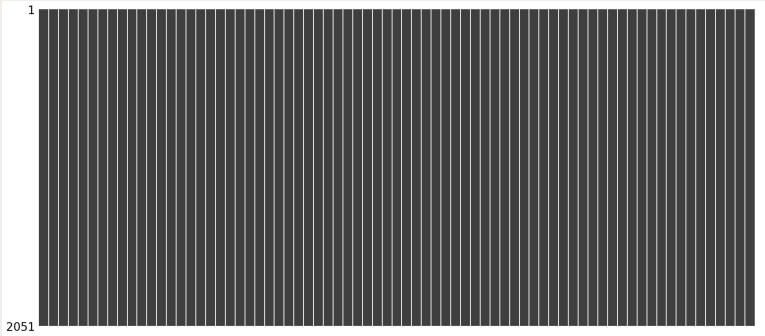
Dropped columns

- Less than 20% data
- Duplicate columns
- Not part of test data

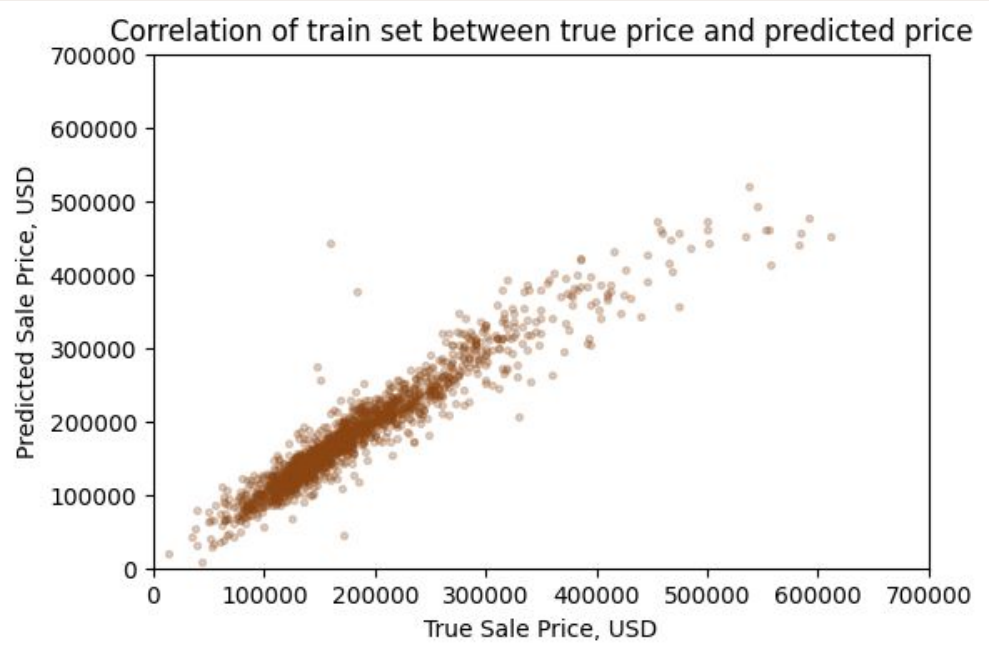
Additional cleaning

- Remove unwanted “ ”, “, ”, etc.
- Fill numerical columns with mean value
- Fill categorical columns with mode

shape : (2051, 73)

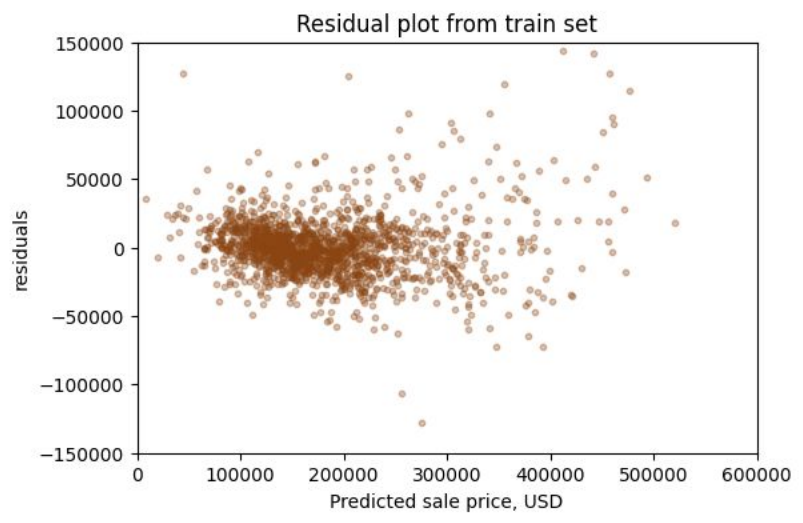


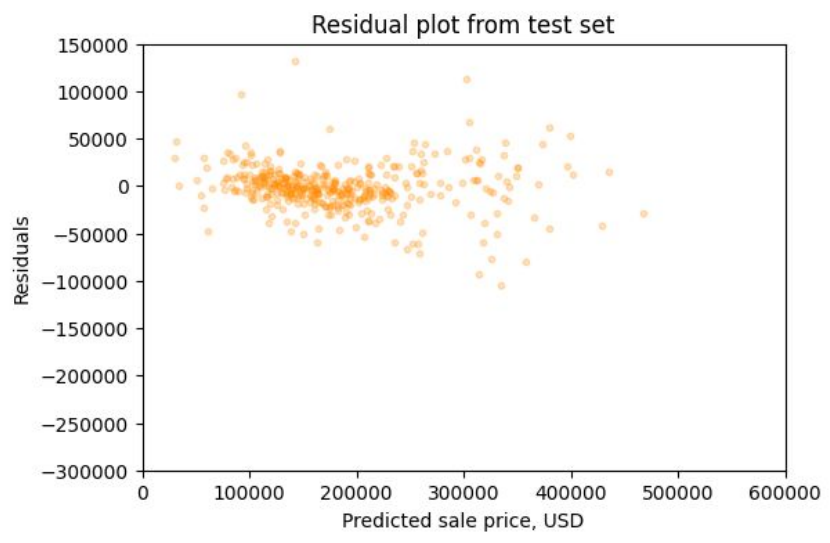
Train set: linear regression analysis



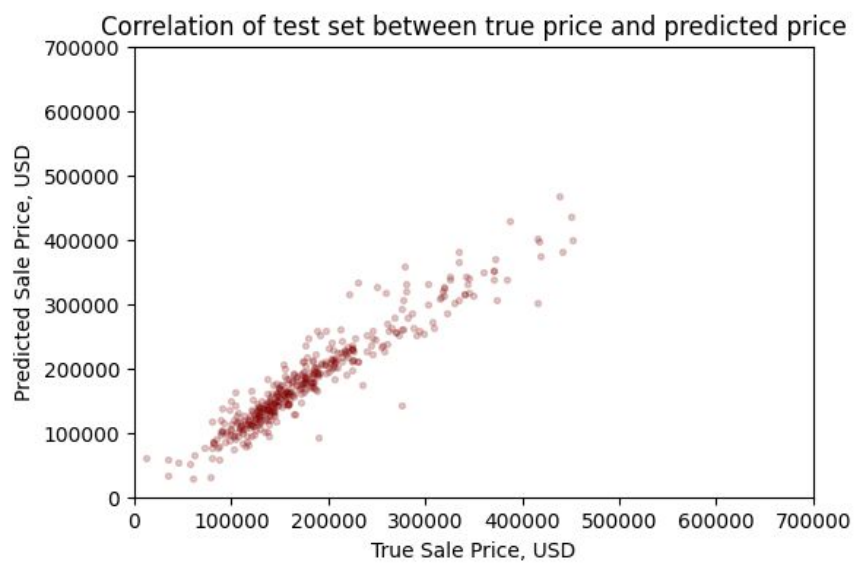
Metrics	Scores	
	Train set	Test set
LR score	0.90	0.90
MSE	580759067	599326238
MAD	15530	16658
Mean CV	0.83	0.84
KFold CV	0.80	0.81

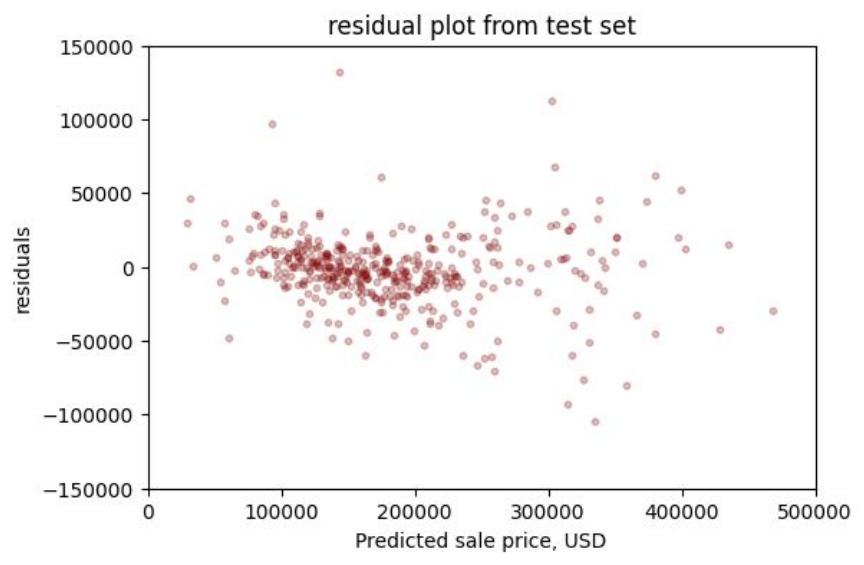
Linear model is a better fit for the data





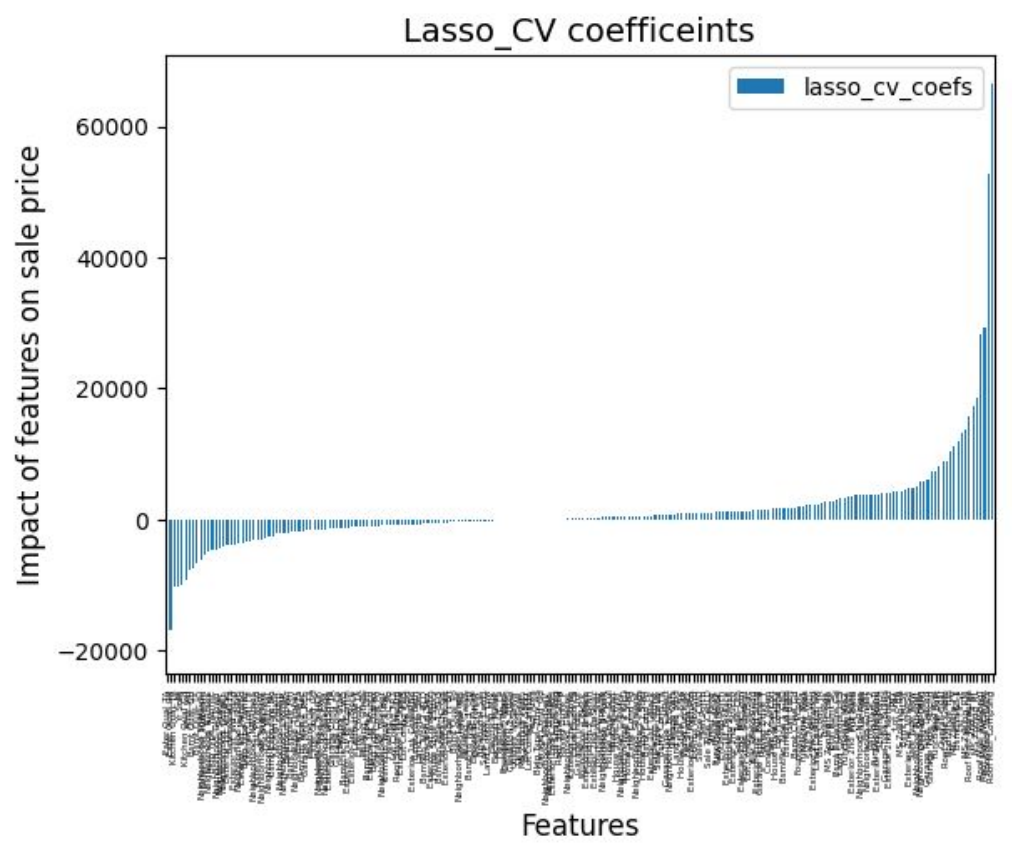
Test set: Linear regression analysis





Metrics

Significant room reduce the independent variables



Lasso_cv score for train set: 0.927
Lasso_cv score for test set : 0.912
Lasso_cv alpha_: 0.001

Summary

- Ames housing market data is exhaustively detailed with over 80 independent variables
- Models were trained and tested using data and regression methods
- With lr score of 0.9, 90% variability in the housing prices can be explained by variables
- Developed models employed to predict the sale price of houses in the test data set
- Lasso analysis suggests significant room in reducing the independent variables

Thanks!!!

Questions and comments please!