Assignment 3 Report - CS 482

Data Preprocessing

Using Machine Learning to Estimate the Cost of Housing

Maddy Connell & Dylan Lozon

Kettering University

11/28/2023

Table of Contents

Ridge and Lasso Regression

Meet the Data

Number of Features: 80
Names of Features:
['Id' 'MSSubClass' 'MSZoning' 'LotFrontage' 'LotArea' 'Street' 'Alley'
'LotShape' 'LandContour' 'Utilities' 'LotConfig' 'LandSlope' 'Condition1'
'Condition2' 'BldgType' 'HouseStyle' 'OverallQual' 'OverallCond'
'YearBuilt' 'YearRemodAdd' 'RoofStyle' 'RoofMatl' 'Exterior1st'
'Exterior2nd' 'MasVnrType' 'MasVnrArea' 'ExterQual' 'ExterCond'
'Foundation' 'BsmtQual' 'BsmtCond' 'BsmtExposure' 'BsmtFinType1'
'BsmtFinSF1' 'BsmtFinType2' 'BsmtFinSF2' 'BsmtUnfSF' 'TotalBsmtSF'
'Heating' 'HeatingQC' 'CentralAir' 'Electrical' '1stFlrSF' '2ndFlrSF'
'LowQualFinSF' 'GrLivArea' 'BsmtFullBath' 'BsmtHalfBath' 'FullBath'
'HalfBath' 'BedroomAbvGr' 'KitchenAbvGr' 'KitchenQual' 'TotRmsAbvGrd'
'Functional' 'Fireplaces' 'FireplaceQu' 'GarageType' 'GarageYrBlt'
'GarageFinish' 'GarageCars' 'GarageArea' 'GarageQual' 'GarageCond'
'PavedDrive' 'WoodDeckSF' 'OpenPorchSF' 'EnclosedPorch' '3SsnPorch'
'ScreenPorch' 'PoolArea' 'PoolQC' 'Fence' 'MiscFeature' 'MiscVal'
'MoSold' 'YrSold' 'SaleType' 'SaleCondition' 'SalePrice']

Name of Target: SalePrice
Number of Samples: 1461

First 5 Rows of Data:

- 1,60,RL,65,8450,Pave,NA,Reg,Lvl,AllPub,Inside,Gtl,Norm,Norm,1Fam,2Story,7,5,2003,2003,Gable,CompShg,VinylSd,VinylSd,BrkFace,196,Gd,TA,PConc,Gd,TA,No,GLQ,706,Unf,0,150,856,GasA,Ex,Y,SBrkr,856,854,0,1710,1,0,2,1,3,1,Gd,8,Typ,0,NA,Attchd,2003,RFn,2,548,TA,TA,Y,0,61,0,0,0,0,NA,NA,NA,0,2,2008,WD,Normal,208500
- 2,20,RL,80,9600,Pave,NA,Reg,Lvl,AllPub,FR2,Gtl,Feedr,Norm,1Fam,1Story,6,8,1976,1976,Gable,CompShg,MetalSd,MetalSd,None,0,TA,TA,CBlock,Gd,TA,Gd,ALQ,978,Unf,0,284,1262,GasA,Ex,Y,SBrkr,1262,0,0,1262,0,1,2,0,3,1,TA,6,Typ,1,TA,Attchd,1976,RFn,2,460,TA,TA,Y,298,0,0,0,0,0,NA,NA,NA,0,5,2007,WD,Normal,181500
- 3,60,RL,68,11250,Pave,NA,IR1,Lvl,AllPub,Inside,Gtl,Norm,Norm,1Fam,2Story,7,5,2001,2002,Gable,CompShg,VinylSd,VinylSd,BrkFace,162,Gd,TA,PConc,Gd,TA,Mn,GLQ,486,Unf,0,434,920,GasA,Ex,Y,SBrkr,920,866,0,1786,1,0,2,1,3,1,Gd,6,Typ,1,TA,Attchd,2001,RFn,2,608,TA,TA,Y,0,42,0,0,0,0,NA,NA,NA,0,9,2008,WD,Normal,223500

- 4,70,RL,60,9550,Pave,NA,IR1,Lvl,AllPub,Corner,Gtl,Norm,Norm,1Fam,2Story,7,5,1915,1970,Gable,CompShg,Wd Sdng,Wd Shng,None,0,TA,TA,BrkTil,TA,Gd,No,ALQ,216,Unf,0,540,756,GasA,Gd,Y,SBrkr,961,756,0,1717,1,0,1,0,3,1,Gd,7,Typ,1,Gd,Detchd,1998,Unf,3,642,TA,TA,Y,0,35,272,0,0,0,NA,NA,NA,0,2,2006,WD,Abnorml,140000
- 5,60,RL,84,14260,Pave,NA,IR1,Lvl,AllPub,FR2,Gtl,Norm,Norm,1Fam,2Story,8,5,2000,2000,Gable,CompShg,VinylSd,VinylSd,BrkFace,350,Gd,TA,PConc,Gd,TA,Av,GLQ,655,Unf,0,490,1145,GasA,Ex,Y,SBrkr,1145,1053,0,2198,1,0,2,1,4,1,Gd,9,Typ,1,TA,Attchd,2000,RFn,3,836,TA,TA,Y,192,84,0,0,0,0,NA,NA,NA,0,12,2008,WD,Normal,250000

Data Preprocessing

Removed Column: ID because it is a unique identifier

Categorical to Numeric Data Conversion:

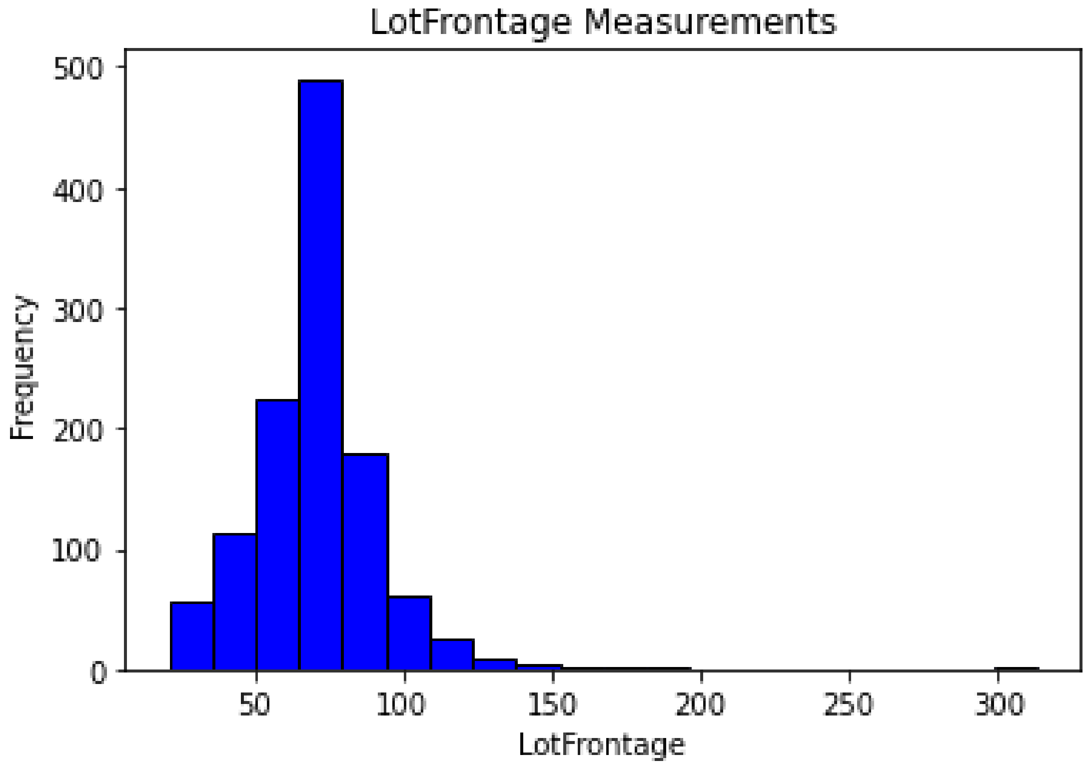
Categorical Columns	Encoding Used	Reason
MSSubClass	One-Hot Encoding	Converting this to a binary matrix was the most straightforward and consistent way to make it numeric.
MSZoning	One-Hot Encoding	Converting this to a binary matrix was the most straightforward and consistent way to make it numeric.
Street	One-Hot Encoding	Converting this to a binary matrix was the most straightforward and consistent way to make it numeric.
Alley	One-Hot Encoding	Converting this to a binary matrix was the most straightforward and consistent way to make it numeric.
LotShape	One-Hot Encoding	Converting this to a binary matrix was the most straightforward and consistent way to make it numeric.
LandContour	One-Hot Encoding	Converting this to a binary matrix was the most straightforward and consistent way to make it numeric.
Utilities	One-Hot Encoding	Converting this to a binary matrix was the most straightforward and consistent way to make it numeric.
LotConfig	One-Hot Encoding	Converting this to a binary matrix was the most straightforward and consistent way to make it numeric.
LandSlope	One-Hot Encoding	Converting this to a binary matrix was the most straightforward and consistent way to make it numeric.
Condition1	One-Hot Encoding	Converting this to a binary matrix was the most straightforward and consistent way to make it numeric.
Condition2	One-Hot Encoding	Converting this to a binary matrix was the most straightforward and consistent way to make it numeric.
BldgType	One-Hot Encoding	Converting this to a binary matrix was the most straightforward and consistent way to make it numeric.
HouseStyle	One-Hot Encoding	Converting this to a binary matrix was the most straightforward and consistent way to make it numeric.
RoofStyle	One-Hot Encoding	Converting this to a binary matrix was the most straightforward and consistent way to make it numeric.
RoofMatl	One-Hot Encoding	Converting this to a binary matrix was the most straightforward and consistent way to make it numeric.
Exterior1st	One-Hot Encoding	Converting this to a binary matrix was the most straightforward and consistent way to make it numeric.

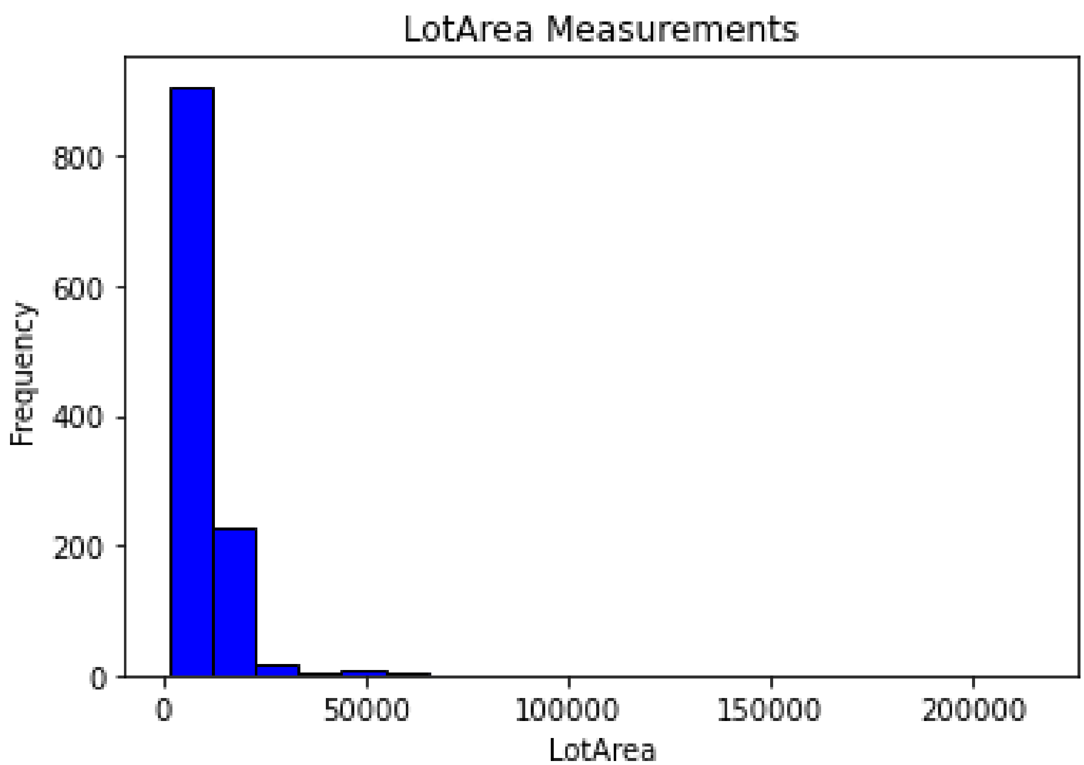
Exterior2nd	One-Hot Encoding	Converting this to a binary matrix was the most straightforward and consistent way to make it numeric.
MasVnrType	One-Hot Encoding	Converting this to a binary matrix was the most straightforward and consistent way to make it numeric.
Foundation	One-Hot Encoding	Converting this to a binary matrix was the most straightforward and consistent way to make it numeric.
BsmtExposure	One-Hot Encoding	Converting this to a binary matrix was the most straightforward and consistent way to make it numeric.
BsmtFinType1	One-Hot Encoding	Converting this to a binary matrix was the most straightforward and consistent way to make it numeric.
BsmtFinType2	One-Hot Encoding	Converting this to a binary matrix was the most straightforward and consistent way to make it numeric.
Heating	One-Hot Encoding	Converting this to a binary matrix was the most straightforward and consistent way to make it numeric.
CentralAir	One-Hot Encoding	Converting this to a binary matrix was the most straightforward and consistent way to make it numeric.
Electrical	One-Hot Encoding	Converting this to a binary matrix was the most straightforward and consistent way to make it numeric.
Functional	One-Hot Encoding	Converting this to a binary matrix was the most straightforward and consistent way to make it numeric.
GarageType	One-Hot Encoding	Converting this to a binary matrix was the most straightforward and consistent way to make it numeric.
GarageFinish	One-Hot Encoding	Converting this to a binary matrix was the most straightforward and consistent way to make it numeric.
PavedDrive	One-Hot Encoding	Converting this to a binary matrix was the most straightforward and consistent way to make it numeric.
Fence	One-Hot Encoding	Converting this to a binary matrix was the most straightforward and consistent way to make it numeric.
MiscFeature	One-Hot Encoding	Converting this to a binary matrix was the most straightforward and consistent way to make it numeric.
SaleType	One-Hot Encoding	Converting this to a binary matrix was the most straightforward and consistent way to make it numeric.
SaleCondition	One-Hot Encoding	Converting this to a binary matrix was the most straightforward and consistent way to make it numeric.
ExterQual	Assigned each category a numeric value	The ratings poor, fair, good, etc can be easily converted to a numeric scale that makes sense for our data.
ExterCond	Assigned each category a numeric value	The ratings poor, fair, good, etc can be easily converted to a numeric scale that makes sense for our data.
BsmtQual	Assigned each category a numeric value	The ratings poor, fair, good, etc can be easily converted to a numeric scale that makes sense for our data.

BsmtCond	Assigned each category a numeric value	The ratings poor, fair, good, etc can be easily converted to a numeric scale that makes sense for our data.
HeatingQC	Assigned each category a numeric value	The ratings poor, fair, good, etc can be easily converted to a numeric scale that makes sense for our data.
KitchenQual	Assigned each category a numeric value	The ratings poor, fair, good, etc can be easily converted to a numeric scale that makes sense for our data.
FireplaceQual	Assigned each category a numeric value	The ratings poor, fair, good, etc can be easily converted to a numeric scale that makes sense for our data.
GarageQual	Assigned each category a numeric value	The ratings poor, fair, good, etc can be easily converted to a numeric scale that makes sense for our data.
GarageCond	Assigned each category a numeric value	The ratings poor, fair, good, etc can be easily converted to a numeric scale that makes sense for our data.
PoolQC	Assigned each category a numeric value	The ratings poor, fair, good, etc can be easily converted to a numeric scale that makes sense for our data.

Number of features after preprocessing: 243

Learning From the Data

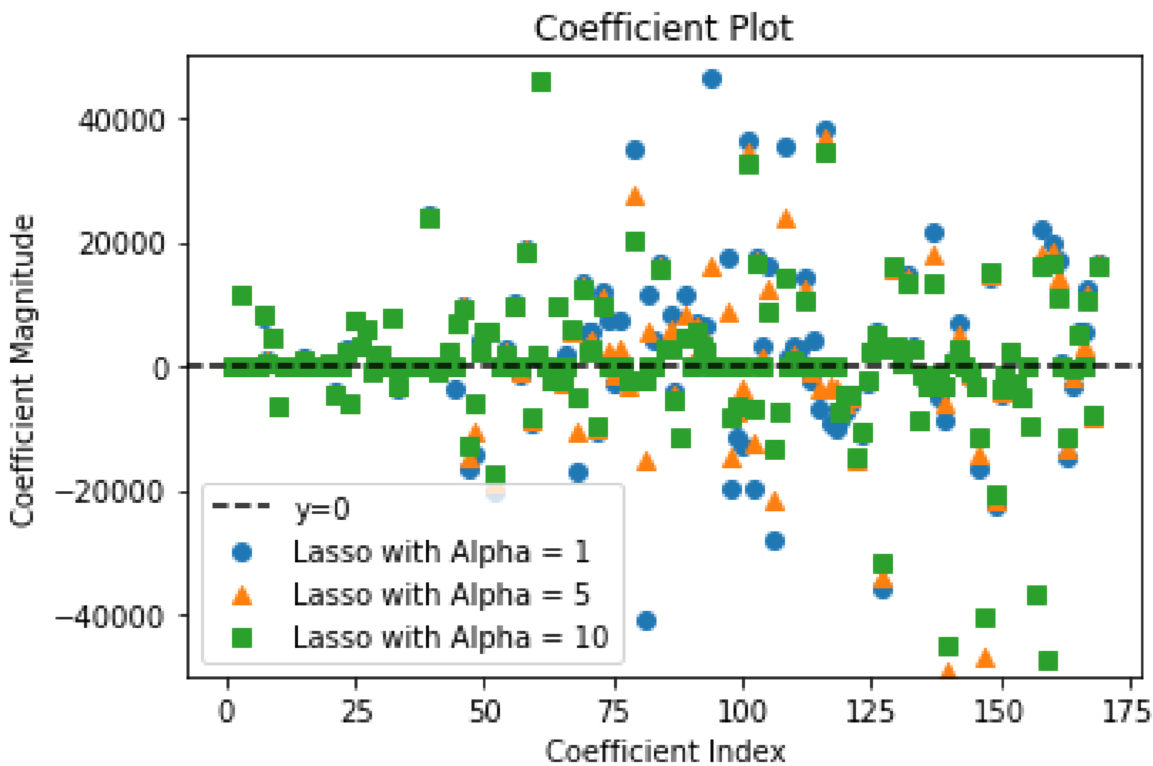




Feature Extraction

Threshold for removing least correlated features: Correlation of -.04866 or below
Features removed: 73
Features remaining: 170

Threshold for removing highest betas: Beta of 2412.3918 or above
Features removed: 51
Features remaining: 119



Best SVM Parameters:
Cost: 100
Gamma: 1