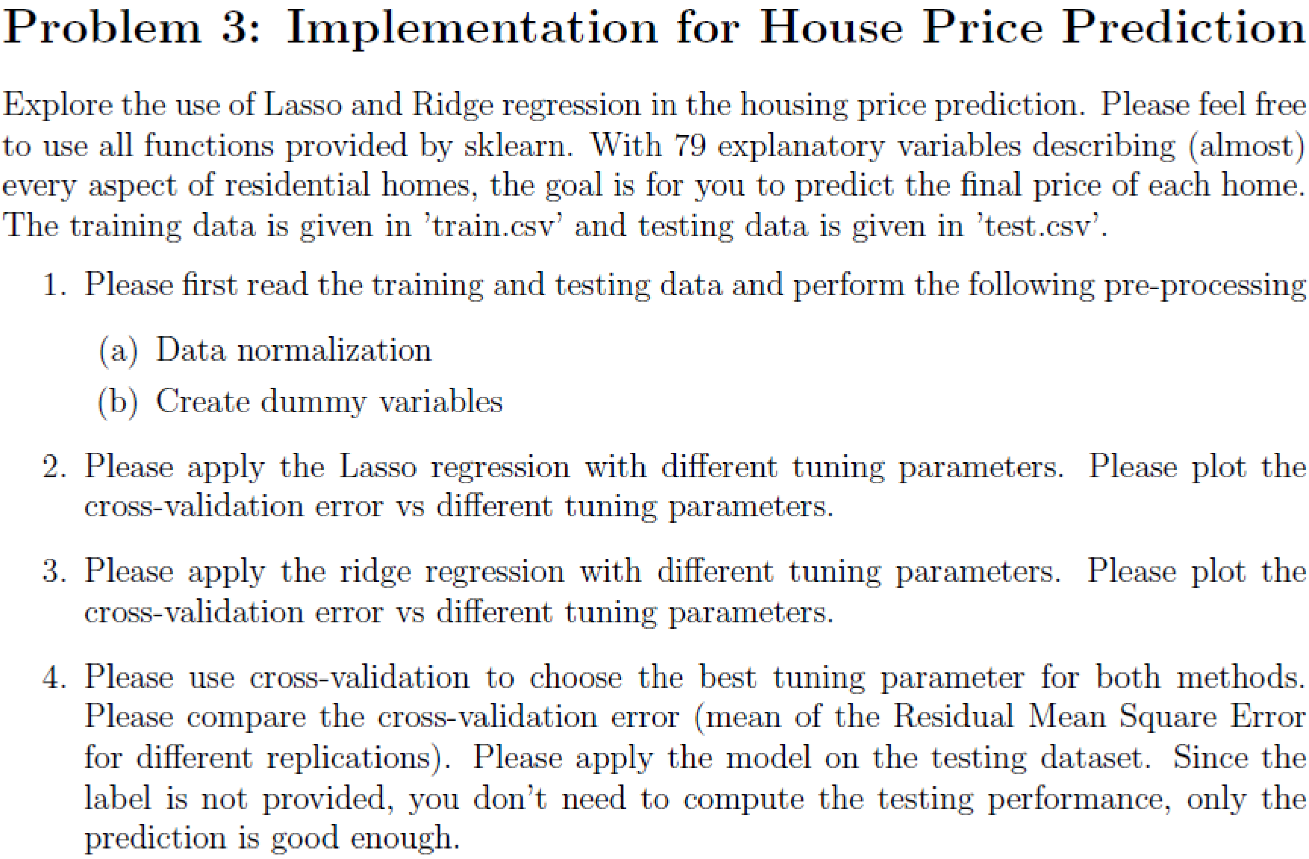


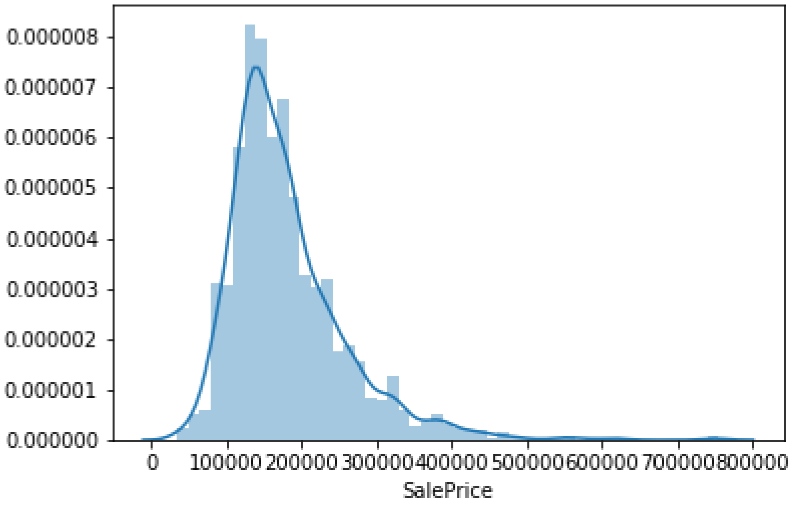
Problem 3: Implementation for House Price Prediction

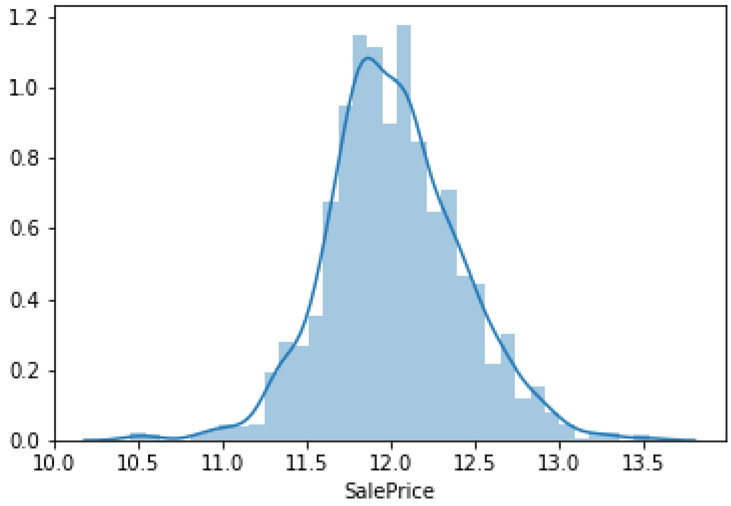
Explore the use of Lasso and Ridge regression in the housing price prediction. Please feel free to use all functions provided by sklearn. With 79 explanatory variables describing (almost) every aspect of residential homes, the goal is for you to predict the final price of each home. The training data is given in 'train.csv' and testing data is given in 'test.csv'.

1. Please first read the training and testing data and perform the following pre-processing
   (a) Data normalization
   (b) Create dummy variables
2. Please apply the Lasso regression with different tuning parameters. Please plot the cross-validation error vs different tuning parameters.
3. Please apply the ridge regression with different tuning parameters. Please plot the cross-validation error vs different tuning parameters.
4. Please use cross-validation to choose the best tuning parameter for both methods. Please compare the cross-validation error (mean of the Residual Mean Square Error for different replications). Please apply the model on the testing dataset. Since the label is not provided, you don't need to compute the testing performance, only the prediction is good enough.

1. Transformation on the respones variable SalePrice, Preprocessing: Deal with the missing data (NA) and create dummy variable for categorical variables

Transformation:





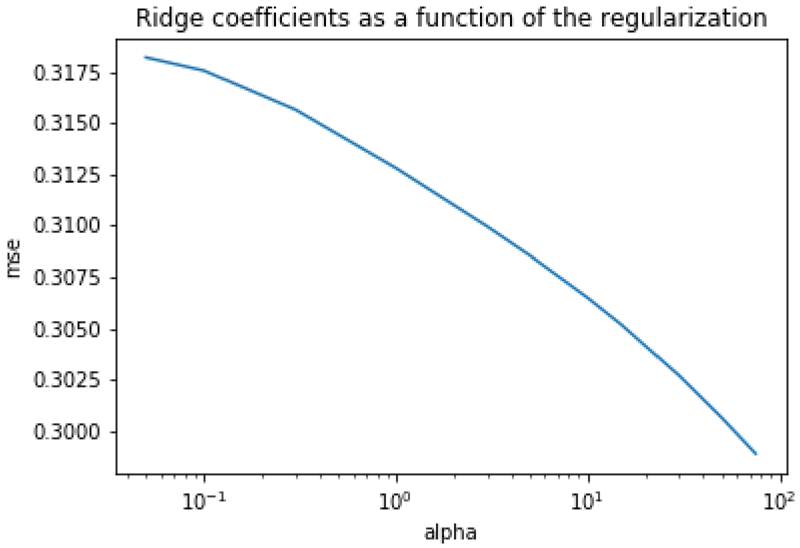
The first plot shows that the SalesPrice his left skewed and hence The response variable Sales Price is applied with log transformation to get a normally distributed dataset which is shown in the plot 2.

Data Cleaning:

I have filled all the Nan values of numerical variables with the mean value of columns. Then i created dummies for all the categorical variables, in this process i remove all the Nan values. I further have matched the number of observations in the Training and Testing datasets. I have also made sure there are no 0 Columns. The missing columns of Test datset was filled with the missing columns functions and Master datasets are created for Train and Test.

2. Model 1: Use Ridge regression and select the best tuning parameter

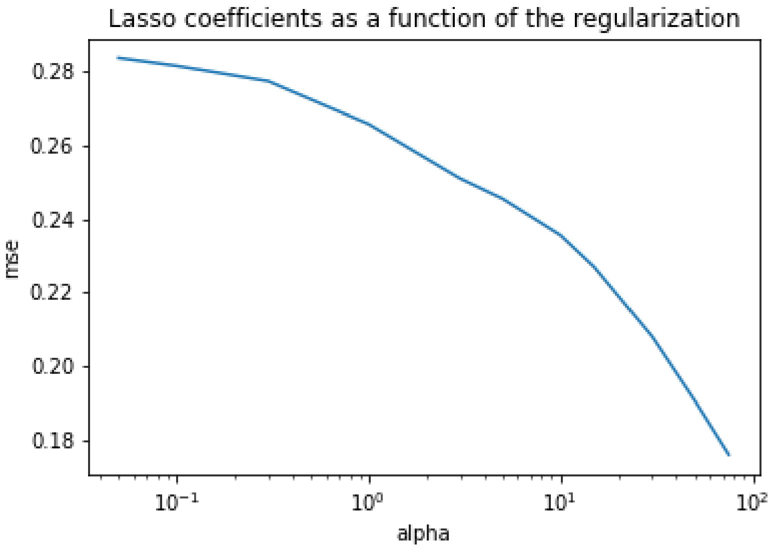
- Please plot the cross-validation error for different tuning parameters
- Choose the best tuning paramter using cross-validation



The plot shows that the cv error decreases as the alpha or the tuning parameter increases. Minimum most value of mse is 0.298906994932 for alpha value 75 Hence the Best tuning parameter using Cross validation is 75.

3. Model 2: Use Lasso regression and select the best tuning parameter

- Please plot the cross-validation error for different tuning parameters
- Choose the best tuning paramter using cross-validation



The plot shows that the cv error decreases as the alpha or the tuning parameter increases. Minimum most value of mse is 0.175893380704 for alpha value 75 Hence the Best tuning parameter using Cross validation is 75.

4. Selecting Best Tuning Parameter, Comparing CV Error, Prediction

Please use cross-validation to choose the best tuning parameter for both methods. Please compare the cross-validation error (mean of the Residual Mean Square Error for different replications). Please apply the model on the testing dataset. Since the label is not provided, you don't need to compute the testing performance, only the prediction is good enough.

Output for Ridge model: 10.0 0.306509346484

Output for Lasso model: 0.05 0.283742358775

Inference:

The Best tuning parameter for Ridge regression model using Cross validation on the Test Dataset is 10.0. The CV Error for Ridge regression model using Cross validation on the Test Dataset is 0.306509346484.

The Best tuning parameter for Lasso regression model using Cross validation on the Test Dataset is 0.05 The CV Error for Lasso regression model using Cross validation on the Test Dataset is 0.283742358775.

Prediction: array([11.84461032, 11.8848263 , 12.16662403, ..., 12.09224462, 11.72875786, 12.36048755])