zerocode learning

Learning Data Analytics Made Easy

USER GUIDE

ELASTIC NET REGRESSION ANALYSIS

TABLE OF CONTENTS

INDEX

1. MODEL- ELASTIC NET REGRESSION

2. ALL ABOUT LEFT PANEL

3. DATA INPUT AND OVERVIEW TAB

4. DATA SUMMARY TAB

5. ELASTIC NET TAB

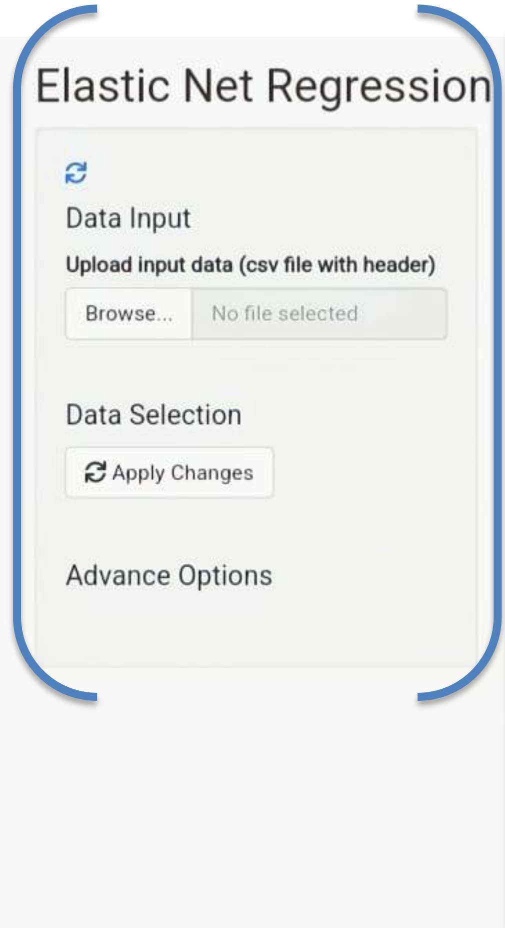
6. RESIDUAL PLOT TAB

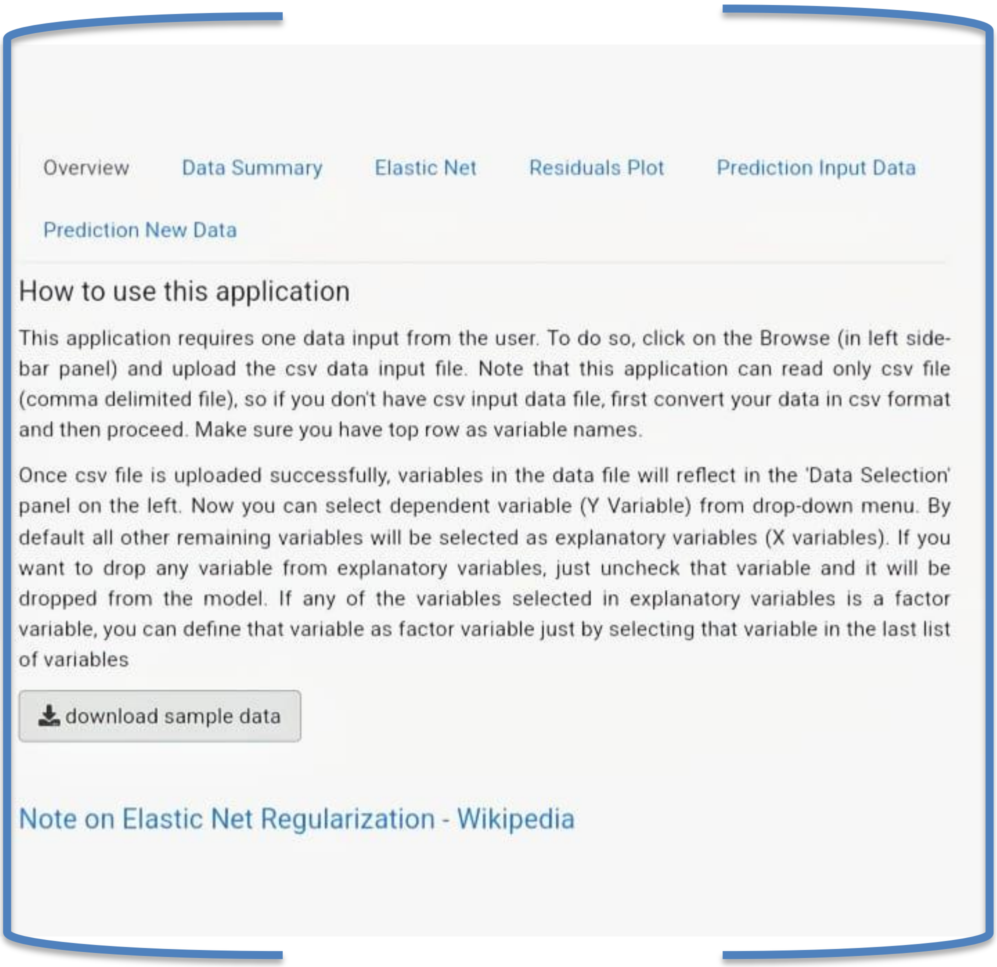
ELASTIC NET REGRESSION ANALYSIS

Elastic net Regression, Elastic net linear regression uses the penalties from both the lasso and ridge techniques to regularize regression models. The technique combines both the lasso and ridge regression methods by learning from their shortcomings to improve the regularization of statistical models. In statistics and, in particular, in the fitting of linear or logistic regression models, the elastic net is a regularized regression method that linearly combines the L_1 and L_2 penalties of the lasso and ridge methods.

LEFT PANEL (INPUT AREA)

OPERATIONAL ANALYSIS TAB (MAIN PANEL)





How to use this application

This application requires one data input from the user. To do so, click on the Browse (in left side-bar panel) and upload the csv data input file. Note that this application can read only csv file (comma delimited file), so if you don't have csv input data file, first convert your data in csv format and then proceed. Make sure you have top row as variable names.

Once csv file is uploaded successfully, variables in the data file will reflect in the 'Data Selection' panel on the left. Now you can select dependent variable (Y Variable) from drop-down menu. By default all other remaining variables will be selected as explanatory variables (X variables). If you want to drop any variable from explanatory variables, just uncheck that variable and it will be dropped from the model. If any of the variables selected in explanatory variables is a factor variable, you can define that variable as factor variable just by selecting that variable in the last list of variables

Note on Elastic Net Regularization - Wikipedia

LEFT PANEL (INP)

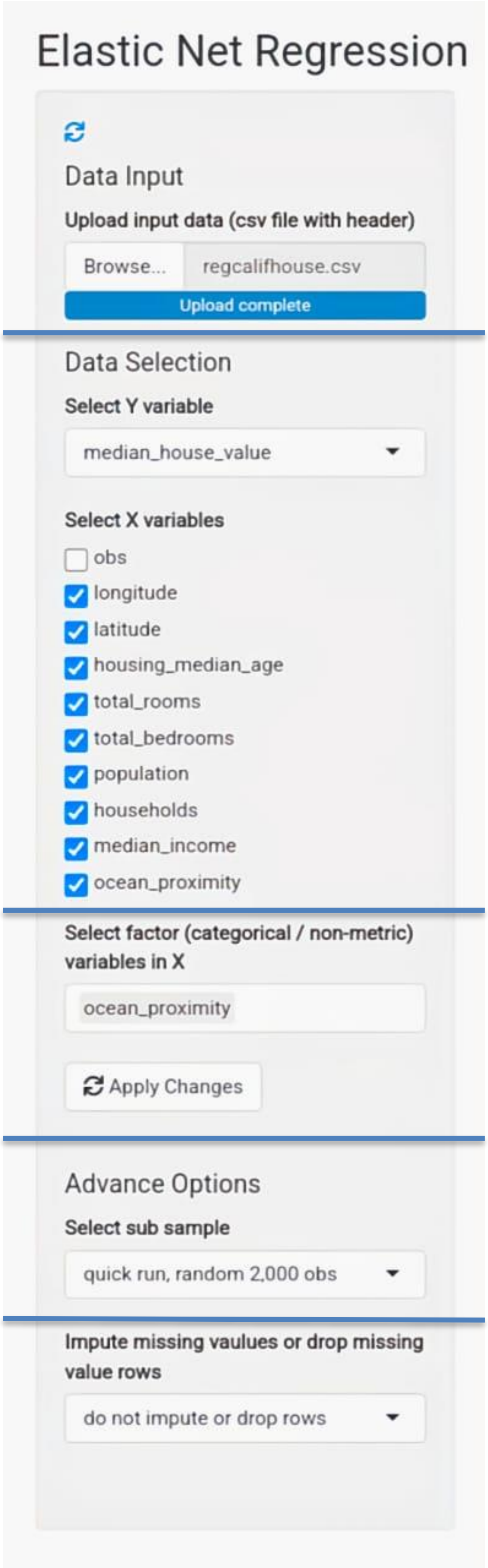
Upload your dataset here

Select your favorable variables required to base the analysis

Apply any changes if you want to do.

Select the subsamples or the whole data for testing.

Deal with missing values either drop or immute it.



Elastic Net Regression

Data Input

Upload input data (csv file with header)

Browse... regcalifhouse.csv

Upload complete

Data Selection

Select Y variable

median_house_value

Select X variables

☐ obs
☑ longitude
☑ latitude
☑ housing_median_age
☑ total_rooms
☑ total_bedrooms
☑ population
☑ households
☑ median_income
☑ ocean_proximity

Select factor (categorical / non-metric) variables in X

ocean_proximity

Apply Changes

Advance Options

Select sub sample

quick run, random 2,000 obs

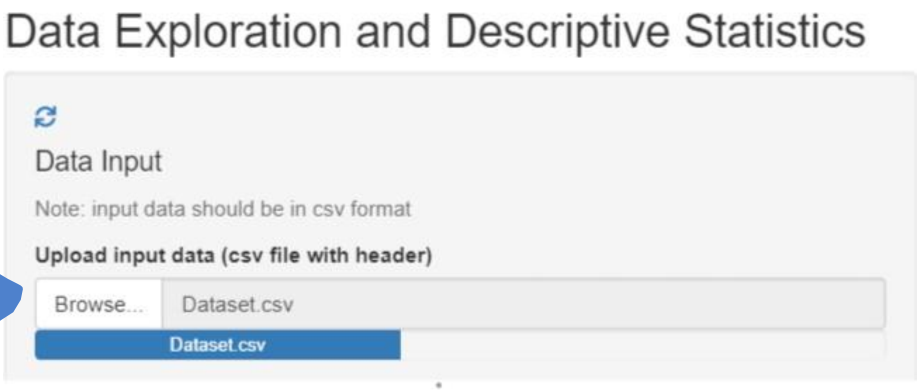
Impute missing vaulues or drop missing value rows

do not impute or drop rows

03

DATA INPUT (UPLOADING DATASET)

- Click on browse
- Select the datafile that is in the form of csv format.(Ex program.csv)
- Browse the file and select the data to train your model for prediction.
- Top rows of the dataset should be of ‘variable names’.



Data Exploration and Descriptive Statistics

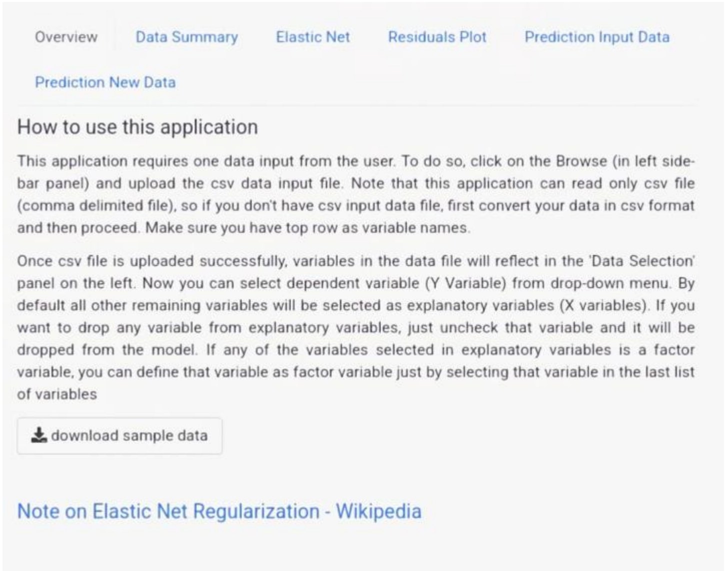
Data Input

Note: input data should be in csv format

Upload input data (csv file with header)

OVERVIEW TAB

This tab provides you with relevant study resources, tutorials, sample datasets and a short overview to start with, which helps you understand and comprehend your data correctly. This tab also provides you the basic idea about text analysis , gives sample data and provides the description about text analysis.



How to use this application

This application requires one data input from the user. To do so, click on the Browse (in left side-bar panel) and upload the csv data input file. Note that this application can read only csv file (comma delimited file), so if you don't have csv input data file, first convert your data in csv format and then proceed. Make sure you have top row as variable names.

Once csv file is uploaded successfully, variables in the data file will reflect in the 'Data Selection' panel on the left. Now you can select dependent variable (Y Variable) from drop-down menu. By default all other remaining variables will be selected as explanatory variables (X variables). If you want to drop any variable from explanatory variables, just uncheck that variable and it will be dropped from the model. If any of the variables selected in explanatory variables is a factor variable, you can define that variable as factor variable just by selecting that variable in the last list of variables

Note on Elastic Net Regularization - Wikipedia

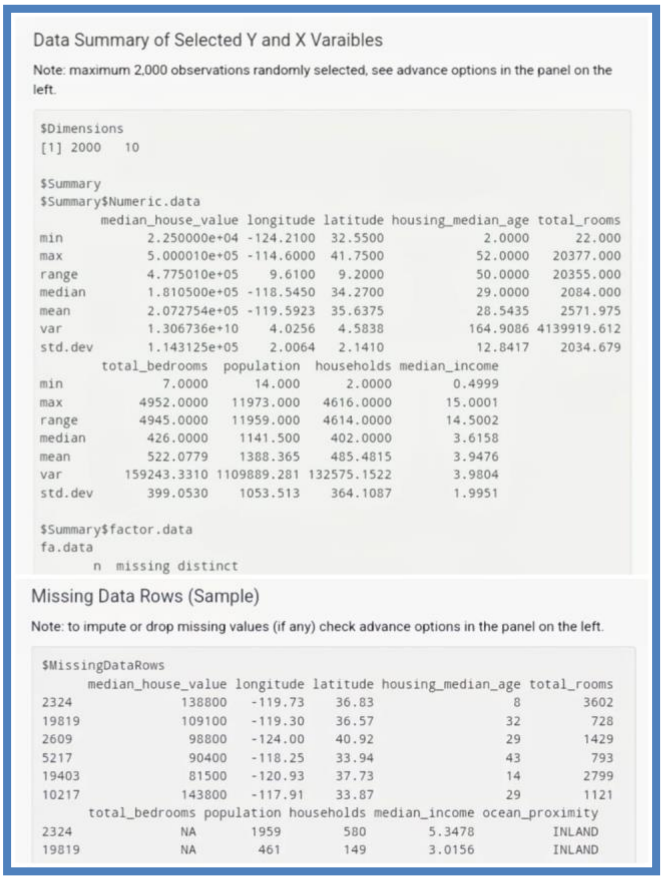
DATA SUMMARY TAB

04

It is very important to understand our data completely to infer meaningful insights and to get an overview of all the data points as a whole, but it is quite impossible to analyze thousand data points manually.

The **'Data Summary'** option enables you to get a comprehensive evaluation through statistical measures that help us form the basis of our analysis.

It will display all the 'descriptive analytics' measures including mean, median, standard deviation, variance etc. for all the data variables present in the dataset. we can review the uploaded data and the contents of it, A brief summary of the data can be seen it includes range of data values, minimum and maximum value missing and null values etc. It also segregates dataset variables into respective data types, such as integer, whole numbers, character etc.



Data Summary of Selected Y and X Varaibles

Note: maximum 2,000 observations randomly selected, see advance options in the panel on the left.

```
$Dimensions
[1] 2000   10

$Summary
$Summary$Numeric.data
       median_house_value longitude latitude housing_median_age total_rooms
min          2.250000e+04 -124.2100  32.5500             2.0000      22.000
max          5.000010e+05 -114.6000  41.7500            52.0000   20377.000
range        4.775010e+05    9.6100   9.2000            50.0000   20355.000
median       1.810500e+05 -118.5450  34.2700            29.0000    2084.000
mean         2.072754e+05 -119.5923  35.6375            28.5435    2571.975
var          1.306736e+10    4.0256   4.5838           164.9086 4139919.612
std.dev      1.143125e+05    2.0064   2.1410            12.8417    2034.679
       total_bedrooms  population  households median_income
min            7.0000      14.000      2.0000        0.4999
max         4952.0000   11973.000   4616.0000       15.0001
range       4945.0000   11959.000   4614.0000       14.5002
median       426.0000    1141.500    402.0000        3.6158
mean         522.0779    1388.365    485.4815        3.9476
var       159243.3310 1109889.281 132575.1522        3.9804
std.dev      399.0530    1053.513    364.1087        1.9951

$Summary$factor.data
fa.data
       n  missing distinct
```

Missing Data Rows (Sample)

Note: to impute or drop missing values (if any) check advance options in the panel on the left.

```
$MissingDataRows
      median_house_value longitude latitude housing_median_age total_rooms
2324              138800   -119.73    36.83                  8        3602
19819             109100   -119.30    36.57                 32         728
2609               98800   -124.00    40.92                 29        1429
5217               90400   -118.25    33.94                 43         793
19403              81500   -120.93    37.73                 14        2799
10217             143800   -117.91    33.87                 29        1121
      total_bedrooms population households median_income ocean_proximity
2324              NA       1959        580        5.3478          INLAND
19819             NA        461        149        3.0156          INLAND
```

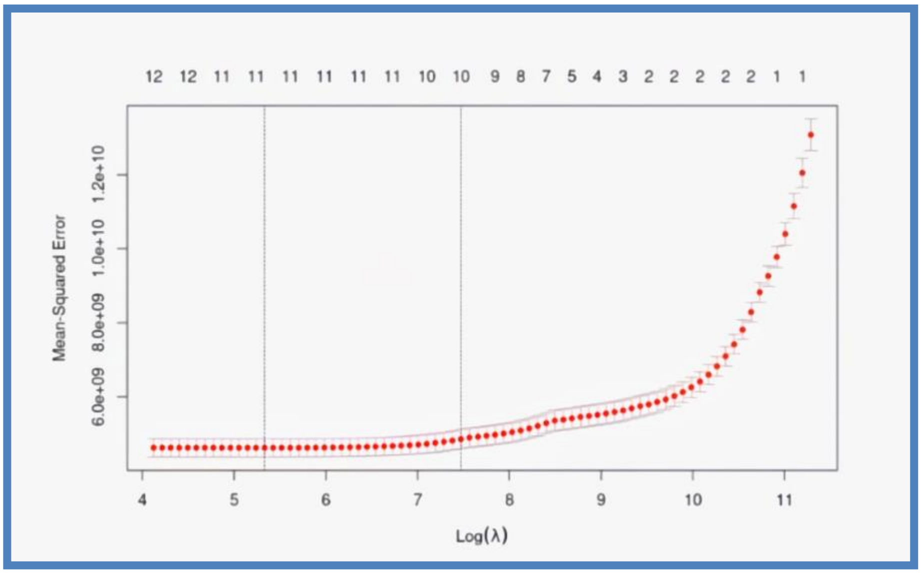
This includes the minimum value maximum value , range between data values ,mean ,median ,mode with standard deviation that is the terms of statistics

Info about missing

Use the left panel to transform selected variables as per the requirement of analysis ,correspondingly the data summary will also change.

ELASTIC NET TAB

05



Elastic net linear regression uses the penalties from both the lasso and ridge techniques to regularize regression models. The technique combines both the lasso and ridge regression methods by learning from their shortcomings to improve the regularization of statistical models. The elastic net simultaneously does automatic variable selection and continuous shrinkage, and it can select groups of correlated variables.

RIDGE REGRESSION : Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values being far away from the actual values

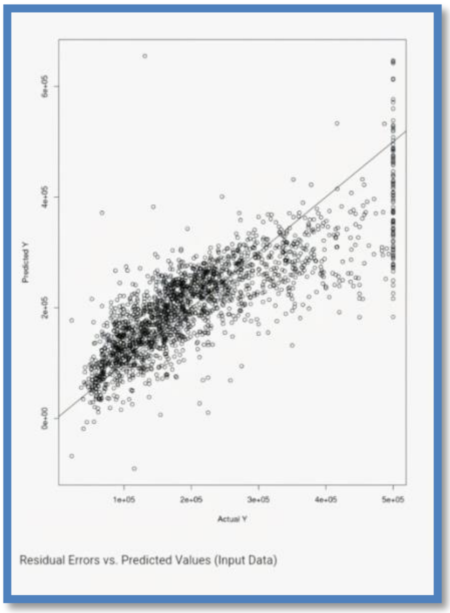
LASSO REGRESSION : Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters).

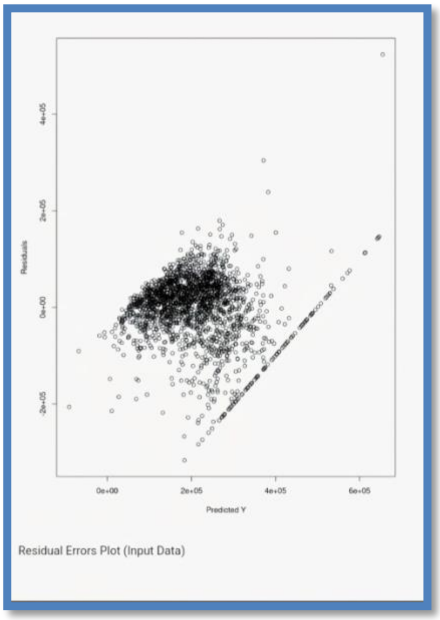
Use the left panel to modify/deal with the outliers identified here.

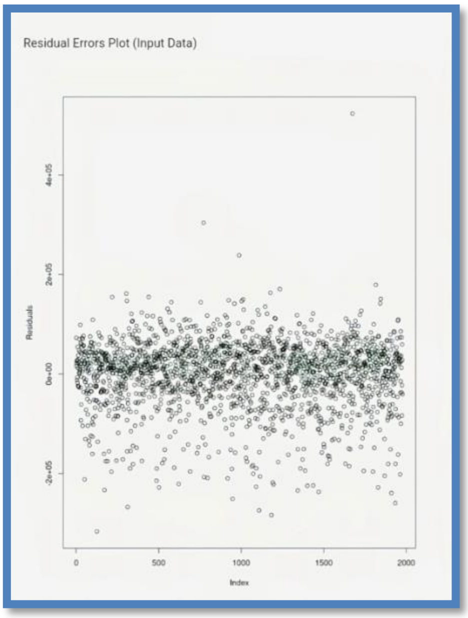
RESIDUAL PLOT

A residual is a measure of how far away a point is vertically from the regression line. Simply, it is the error between a predicted value and the observed actual value. A typical residual plot has the residual values on the Y- axis and the independent variable on the x-axis. As residuals are the difference between any data point and the regression line, they are sometimes called“errors.” Error in this context doesn’t mean that there’s something wrong with the analysis; it just means that there is some unexplained difference. In other words, the residual is the error that isn’t explained by the regression line.

EXAMPLE GRAPH







The equations of calculation of percentage prediction error (percentage prediction error = measured value - predicted value measured value× 100

or

percentage prediction error = predicted value - measured value measured value ×100) and similar equations have been widely used.

Use the left panel to impute or drop the missing values identified here