Advanced Linear Regression Case Study

Contributor : Sayak Bhattacharjee

Introduction

Problem Statement

Surprise Housing has decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them on at a higher price

We are required to build a regression model using regularization to predict the actual value of the prospective properties and decide whether to invest in them or not.

The company wants to know:

Which variables are significant in predicting the price of a house, and

How well those variables describe the price of a house.

Business Objective

We are required to model the price of houses with the available independent variables.

The model should provide clarity on the top predictor variables and their influence on the price of the house.

Solution Approach

As it is clearly instructed this this is a linear regression problem with an objective of achieving the predictor variables, we will create linear model by performing the below steps

The solution is divided into the following sections:

Data understanding

Data cleaning

Data Exploration

- Univariate Analysis
- Bivariate Analysis
- Outliers treatment

Data preparation

Feature engineering – encoding of categorical variables

Scaling of the numerical variables

Model building and evaluation

Data Understanding

- Here we have loaded the data into python
- Gone through the data dictionary
- Understanding of the volume and dimensions of the data
- Load required libraries in our notebook

Data Cleaning

- Null Handling
  - Drop those columns where more than 50 records are null (out of total 1460)
  - Evaluate if any rows are with null values

- Data Type change
  - For certain variable we changed the data type from numeric to object as the data is ordinal in nature
  - Similarly changed data type to number for certain variables

- Delete variables
  - if we find the cardinality of the data for a variable is very low

- Imputing Null values
  - If number of null records in a col was less than 50 then we imputed the null values with median (for numeric variable) and mode (for categorical variable)

Data Exploration

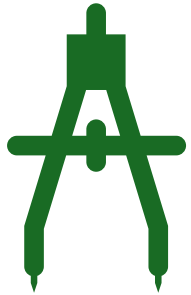
- Univariate Analysis
  - We found the skewness of SalePrice and performed log transformation to adjust the skewness
  - Outliers for numerical variables are replaced with min and max data
  - Drop few categorical variables if 95% or more data in one category
- Bivariate Analysis
  - By plotting co-relation heatmap we found the variables which are 80% or more corelated, dropped them
  - Got an indication about the linearity between SalePrice and other numeric variables by plotting scatter plot

Data Preparation

Feature Engineering

Label Encoding – Performed for the categorical variables which carry ordinal quality of data

One hot encoding – Performed for the nominal categorical variables



Scaling

Standard Scaler – We have performed standard scaling for the numerical variables

Model Building and Evaluation

1. First, we created a linear regression model together with RFE
2. Secondly, we created a linear regression with grid search cross validation method and RFE as estimator
3. VIF – we have not performed for this assignment
4. Thirdly we created a ridge regression model with set of lambda hyper parameters. Through grid search CV we found the best value of the lambda and created the final ridge model with the best lambda value
5. Lastly, we created a lasso regression together with grid search CV.
6. Finally we decided to go for lasso as the best performing model and fetched the top predictors with their co-efficients.