Machine learning project

House prices - Advanced regression techniques

Sattaru Harshavardhan Reddy 19-11-EC-029
Kondra Dharma sena 19-11-EC-

ABSTRACT Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this playground competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.

With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition challenges you to predict the final price of each home.

INTRODUCTION this project is all about predicting house prices using linear regression models this project is all based on a competion on kaggle-https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview all the datasets are from kaggle there are 80 variables of which house prioce is the dependent variable we hav to predict the rest are attributes to houses which affects the price we removed all the null values by replacing them with mode. and used regression model for further details a out data handling please refer code.

LIBRARIES USED

- PANDAS for data handling
- NUMPY for data handling
- SEABORN for data visualization
- SCIKIT LEARN(sklearn) for Machine Learning Algorithms

OBJECTIVE

- Create an analytical framework to understand key factors impacting house price.
- Develop a modeling framework to estimate the price of a house that is up for sale.
- Making a model based on advanced regression techniques which can predict the price of the houses for sale.

DATASETS AND VARIABLES

- 80 variables present in the dataset
- SalePrice - dependent variable it represents final price.
- Remaining variables represent different features of the houses.

DATA PROCESSING

- Data cleansing -

➔ Treating missing values - continuous variables and character variables.

HANDLING AND TREATING NULL VALUES

The real-world data often has a lot of missing values. The cause of missing values can be data corruption or failure to record data. The handling of missing data is very important during the preprocessing of the dataset as many machine learning algorithms do not support missing values.

1. Deleting Rows with missing values
2. Impute missing values for continuous variable
3. Impute missing values for categorical variable
4. Other Imputation Methods
5. Using Algorithms that support missing values
6. Prediction of missing values
7. Imputation using Deep Learning Library — Datawig

Here we used the second method

Impute missing values with Mean/Median:

Columns in the dataset which are having numeric continuous values can be replaced with the mean, median, or mode of remaining values in the column. This method can prevent the loss of data compared to the earlier method. Replacing the above two approximations (mean, median) is a statistical approach to handle the missing values.

Pros:

- Prevent data loss which results in deletion of rows or columns
- Works well with a small dataset and is easy to implement.

Cons:

- Works only with numerical continuous variables.
- Can cause data leakage
- Do not factor the covariance between features.

Contnious variable - missing variables are treated by replacing with mode

Why mode not mean?

Mean is vulnerable to outliers

When the data is skewed, it is good to consider using mode values for replacing the missing values.

Imputation method for categorical columns:

When missing values is from categorical columns (string or numerical) then the missing values can be replaced with the most frequent category. If the number of missing values is very large then it can be replaced with a new category.

Pros:

- Prevent data loss which results in deletion of rows or columns
- Works well with a small dataset and is easy to implement.
- Negates the loss of data by adding a unique category

Cons:

- Works only with categorical variables.
- Addition of new features to the model while encoding, which may result in poor performance

METHOD

We use regression technique and train our model to work on data in predicting the last variable i.e., price using the other 79 variables as features.

LINEAR REGRESSION

It is one of the most-used regression algorithms in Machine Learning. A significant variable from the data set is chosen to predict the output variables (future values). Linear regression algorithm is used if the labels are continuous, like the number of flights daily from an airport, etc. The representation of linear regression is y = b*x + c.

In the above representation, ‘y’ is the independent variable, whereas ‘x’ is the dependent variable. When you plot the linear regression, then the slope of the line that provides us the output variables is termed ‘b’, and ‘c’ is its intercept. The linear regression algorithms assume that there is a linear relationship between the input and the output. If the dependent and independent variables are not plotted on the same line in linear regression, then there will be a loss in output. The loss in output in linear regression can be calculated as:

Loss function: (Predicted output – actual output)2.

And we use linear regression techniques for predicting prices.