

PFA-HOUSING-PROJECT

SUBMITTED BY: - TEJENDRA SONI

ACKNOWLEDGMENT

I would like to express my thanks of gratitude to SME SWATI MAHASETH as well as Flip Robo who gave me the golden opportunity to do this wonderful project on the topic PFA HOUSING PROJECT, which also helped me in doing a lot of research and I came to know about so many new things. I am very thankful to them.

INTRODUCTION

Problem Statement:

Houses are one of the necessary needs of each person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors in the world’s economy. It is a very large market and there are various companies working in the domain. Data science comes as a very important tool to solve problems in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in house sales and purchases. Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for housing companies. Our problem is related to one such housing company. A US-based housing company named 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 at a higher price. For the same purpose, the company has collected a data set from the sale of houses in Australia. The data is provided in the CSV file below. The company is looking at prospective properties to buy houses to enter the market. You are required to build a model using Machine Learning in order to predict the actual value of the prospective properties and decide whether to invest in them or not

Objective:

You are required to model the price of houses with the available independent variables. This model will then be used by the management to understand how exactly the prices vary with the variables. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns. Further, the model will be a good way for the management to understand the pricing dynamics of a new market.

Firstly, we will start by importing required libraries and databases.

A picture containing graphical user interface

Description automatically generated

Table

Description automatically generated

Our dataset has 1168 rows and 81 columns.

We will first divide the data for train & test we will do pre-processing for both the data

After the pre-processing of train data, we will build the model & will use that model in making predictions for test data

Let’s check the datatype of all columns:

Text

Description automatically generated

* Printing all the data types of the columns that are present in the train dataset
* Our dataset is large it is not printing all the columns so we will use value counts functions so that we can determine that how many datatypes are present in our train dataset
* Here we can see that 43 columns are of object datatype
* 35 columns that are of integer datatype
* & 3 columns are of float datatype

**VISUALIZATIONS:**

Text

Description automatically generated with medium confidence

* There are some null values present in our dataset showing it with the help of seaborn library & with the heatmap
* All the white spots that you can see in the plot are null values shown here

Graphical user interface, text, application, email

Description automatically generated

* Treating the null values according to the columns
* Filling the null values accordingly
* After removing the null values from the selected columns
* After removing the null values, we are checking the columns that are still pending
* We can see that there are 4 columns remaining

Graphical user interface, application, table

Description automatically generated

Dropping the columns as they don’t contribute towards the analysis part

Chart

Description automatically generated with medium confidence

Plotted the count plot for each column So that we can determine that in each Column which variable is mostly used among All the present once's

Here we have set hue=company means the plot will work with respect to company

Chart, bar chart

Description automatically generated

Plotting the bar plot of each column with respect to our target column i.e., sale price

Graphical user interface

Description automatically generated with medium confidence

The LM plot stats that SALEPRICE & YEARBUILT are in positive relationship we can here see as the SALEPRICE increases the YEARBUILT of the HOUSES are also increases

We have set hue to fireplace Qu so the plot will work with respect of that column

Chart

Description automatically generated

Here we have set hue=MSZoing means the plot will work with respect to MSZoing

Chart, box and whisker chart

Description automatically generated

Plotting our basement type columns with our target column

Diagram

Description automatically generated

Plotted the line plot of each column with respect to sale price

To determine the positive & negative relationships among the

Columns

* OVERALL-QUALITY, OVERALL-CONDITION, FIREPLACES & SALES PRICE ITSELF ALSO

ARE SHOWING HIGHLY POSITIVE RELATIONSHIP WITH THE SALE PRICE

* FROM THE ABOVE GRAPH WE CAN SEE THAT THERE IS NOT A SINGLE GRAPH THAT

IS STATING THE NEGATIVE RELATIONSHIP WITH OUR TARGET VARIABLE

* THERE ARE SOME COLUMNS WHICH ARE GOING UP & DOWN SO MIGHT BE THEY

ARE CHANGING DUE TO SOME EXTERNALS REASONS BUT THEY ARE NEITHER FULLY POSITIVELY

RELATED NOR NEGATIVELY THEY WILL DIFFER OCCORDING TO THE EXTERNALS REASONS

* 4. GARAGE-AREA, WOODDECKSF, OPENPORCHSF, ENCLOSED-PORCH & MANY

MORE GRAPH ARE MEDIUM IN THE RANGE

Graphical user interface

Description automatically generated

Chart, histogram

Description automatically generated

OUTLIERS & SKEWNESS REMOVAL:

Text

Description automatically generated

Graphical user interface, text

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Scaling the data with standard scaler

Graphical user interface

Description automatically generated with low confidence

Using variance inflation factor we are removing the

multicollinearity

PRINCIPLE COMPONENT ANALYSIS

A picture containing graphical user interface

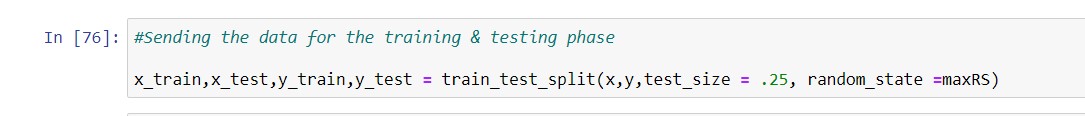
Description automatically generated

Creating the instance of PCA & plotting the curve for that

70 components explain around 95% variance in Data

Graphical user interface, text, application, email

Description automatically generated



Text, letter

Description automatically generated

Text

Description automatically generated

Here in the table, we can see that Ridge model is giving

Highest R2 score among all the different present models

So, we select the Ridge best for our model

But before proceeding further we will cross validate

Each of the models so that we can finally select which model

Will be the best for our dataset

Text

Description automatically generatedAfter determining the cross-validation scores of the models we will keep the Ridge as our model

Because the Ridge model shows the least difference in all the models

Graphical user interface, text, email

Description automatically generated

SAVING THE MODEL

A picture containing graphical user interface

Description automatically generated

PREPROCESSING FOR TEST DATA

Table

Description automatically generated

Its stats that there 292 rows & 80 columns in our test dataset

Text

Description automatically generated with medium confidencePrinting all the datatypes of the columns &

Also, the values counts so that we can

Determine that which column belongs to

The datatypes

Graphical user interface

Description automatically generated

Table

Description automatically generated

Filling the null values accordingly to the

columns

After filling out the values checking how

many columns still have null values

As we can see there are 5 columns that

have null values

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, application

Description automatically generated

Dropping the 5 columns as they don’t

contribute to the analysis part

Printing the data frame

STORING X & Y AGAIN SENDING THE DATA FOR TRAINING & TESTING

Graphical user interface, text, application, email

Description automatically generated

Table

Description automatically generated

Storing the model back & making prediction from it for test dataset

Graphical user interface, text, application, email

Description automatically generatedAlso, making a data frame of predicted answers & measured answers & printing it

Graphical user interface, text, application, email

Description automatically generated

Printing the names of the columns that are positively helped us for prediction

Graphical user interface, text

Description automatically generated

The listed columns helped us for making the prediction