FLIP ROBO

HOUSING: PRICE PREDICTION

Submitted by:

PARV SHARDA

ACKNOWLEDGMENT

I would like acknowledge the classes and projects shared with us by my mentor Ms. Deepika Sharma from Datatrained institute for our understanding to crack this housing price prediction. Also, I took help through internet surfing via Youtube & Google.

Special thanks to my mentor Mr. Shubham Yadav of Fliprobo Technologies for sharing the feedback of my last assignment which helped me to make my current project more perfect.

INTRODUCTION

- Business Problem Framing

  We 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.

  In real world, we need predictive modelling and recommendation systems for achieving the business goals for housing companies.

- Conceptual Background of the Domain Problem

  We have to find which variables are important to predict the sale price and how various variable values helps in deciding the price of a house.

- Review of Literature

The research done while predicting the best house prices we had 81 columns and 1168 rows and the target variable is SalePrice. The data has null values were treated by either dropping the columns or replacing with mode. Also, through LabelEncoder the huge no. of categorical data was converted to integer or float data. After this data visualization was done by plotting scatter plot, histogram, Catplot and countplot. Maximum relationships were coming positive towards target variable. Skewness was handled through power transform method. Later outliers were detected through boxplot and removed through Z-score method.

After this test data was imported and treated same like training set data replacing all categorical data to integer data. In model building, test and training data were split later model were treated with Linearregression, RandomForestRegressor and DecisionTreeRegressor.

- Motivation for the Problem Undertaken

The motivation behind creating this model was to detect the sale price of houses depending upon various observations such as Lot size in square feet, Type of road access to property, Type of alley access to property, General shape of property, Type of utilities available, Physical locations within Ames city limits, Slope of property, Type of dwelling, Style of dwelling, Rates the overall material and finish of the house, Rates the overall condition of the house, Type of roof, Roof material, Type of foundation, the height of the basement, Type of heating, Central air conditioning, Electrical system, Fireplace quality, Garage condition.

So all these factors will help us to know the approx. pricing of the house. We have to just find the relation in accordance with all.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  Mathematically-: Checking the missing values in the columns and replacing/dropping them, changing the categorical data into numeric data and getting an accurate dataframe.

  Analytics-: Plotting type of graphs such as Category plot graphs, histograms, scatter plots this helped us in knowing the correct available factors in a house and their relationships with the sale price of a house to be sold.

  Modeling-: We used various model creating methods such as LinearRegression, RandomForestRegressor & DecisionTreeRegressor. By comparing the best R2 score value of training and test data we evaluated the best form of data.

- Data Pre-processing Done

  For cleaning of data we first checking the missing or null values the columns with missing values in high quantity were dropped as they were of no such required to interpret data and columns with of less missing values these places were filled with mode values.

  Later, the columns which are categorical were converted into integer or float values through Label encoder (i.e. 0,1,2,3,4,5)

  Skewness was removed through power transform method and outliers were detected by boxplot and removed through Z-score.

Model/s Development and Evaluation

- Visualizations

  Mention all the plots made along with their pictures and what were the inferences and observations obtained from those. Describe them in detail.

  If different platforms were used, mention that as well.

  dfcor=df.corr() -: finding correlation b/w eachother

  scatter plot-: Relationship b/w two factors and how much data is scattered.

  sns.countplot-: It gives factor like yes or no, available or not. Type of floor, neighbourhood etc.

  sns.catplot-: Used as swarm plot comparing factors with target variable.

  Boxplot-: Used to detect outliers.

- Interpretation of the Results

  After doing model building through 3 methods-: LinearRegression, RandomForestRegressor & DecisionTreeRegressor and comparing the R2 score value and RMSE value. The perfect model came as DecisionTreeRegressor.