Week8_Final_Project_Step1

Akila Selvaraj

2/12/2022

Introduction:

House Price Prediction

House is one of human life's most essential needs, along with other fundamental needs such as food, water, and much more. Demand for houses grew rapidly over the years as people's living standards improved. There are many people who make their house as an investment and property. Purchasing a house is a big decision in a person's life and needs a considerable amount of thought and research. One would like to buy a house at the best rate and minimum risk and would like it to be the best investment for the future. Various online websites, real estate agents and realtors try to guide home buyers by letting them compare different houses available for purchase.

We are trying to predict the house prices using Machine learning algorithms Linear Regression considering factors such as Median income in a county, Crime rate in that county, public schools, hospitals, hospital ratings and unemployment rate in the county. House Price prediction is important to drive Real Estate efficiency. As mentioned earlier, House prices were determined by calculating the acquiring and selling price in a locality. Therefore, the House Price prediction model is very essential in filling the information gap and improve Real Estate efficiency.

Goal

Our goal is to predict the house prices in a county in the next few months. The ultimate goal of the project is to build a prediction engine capable of predicting housing price. This is supervised learning problem as our data set consists of labelled observations and we are going to use the option of multiple regression.

What is the predicted price per square foot of a home in a given zip code/ county in the next few months? Can we predict the price based on crime rates, schools and other information/metrics provided by Zillow for a zip code/neighborhood/county?

House price prediction helps individuals

Before putting the house in the market, people look at comparable properties in the areas and determine the sale price. Machine learning builds an end to end application or solution that is capable of predicting the house prices better than individuals. This helps individuals, realtors to make data driven decisions on house price.

House price Prediction - Data Science Problem

The appraisal of real estate is traditionally conducted by a licensed professional, who would carry out a holistic survey based on several factors such as location, surroundings, areas, and facilities of a real estate. Nevertheless, the manual appraisal would inevitably have the possibility to involve the appraisers' factors and vested interest. This potential risk would likely cause a biased or subjective evaluation of a particular real estate, bringing loss for investors or households. Thus, constructing a feasible algorithm and automated model

which could appraise the real estate impartially and objectively has critical significance for any potential parties participating in these transactions.

Research questions

1. What are the important features that affect the house price? Which factors are related to sale price at the city/county level?
2. How the multiple independent variables, either by themselves or together, influence changes in the dependent variable.
3. To what extent and in what manner do the predictors explain variation in the criterion?
4. How to build a model to predict the house price?
5. How to select the independent factors that decide the house price?
6. Do relationships exist between the datasets chosen for this study? If so, what kind of relationship?
7. How to evaluate our prediction performance?
8. How accurate the model will be ?

Approach

I am planning to use data science framework and going to follow the below steps to solve this problem.

1. Data Analysis/Exploratory Data Analysis
2. Data cleaning/Feature Engineering
3. Feature selection
4. Modeling
5. Model Evaluation

Addressing the problem using this approach.

Exploratoty Data Analysis: In this step, we first approach the data and analyze categorical and numerical variables. This gives the information and will guide me in which predictors I would want in my model. Points covered in this step will be analysing data types, identifying outliers, missing values and distribution of data. In this step, we will also try to identify high-unbalanced variables,minimum and maximum of each variable, and check the variable with outliers. Basically we gather all the information on the main characteristics, the variables and their relationships and finding out the important variables that can be used in our problem.

Data Cleaning: Here, we are going to identify the errors in data, correct and remove inaccurate data for the next step. Data irregularities, structural errors will be corrected in this step. Building a model using bad data will result in bad model.So, cleaning the data will help us to build unbiased model.

Feature Selection: In this process, we are going to evaluate relationship between each individual variable and the target variable which in our case is sales price of the house. This helps to reduce the computational cost of modeling and to improve the performance of the model. We are going to choose the variables which have strong relationship with the target variable sales price. Non-informative , irrelevant variables, redundant predictors which will add uncertainty to the predictions will be be removed from our study and we are not going to consider those variables in our model.

Modeling: Based on the input variables we selected in the previous step, I am planning to use multiple regression model. We will split the data into training and test data and create the model and model coefficients will be calculated.

Model Evaluation: Once the model is created , we will evaluate the model by calculating the accuracy of the model by evaluating how model is performing on test data.

Input Data

To predict house prices using supply-demand features, I am planning to use below datasets.

Income dataset :

Hyperlink: https://www.kaggle.com/goldenoakresearch/us-household-income-stats-geo-locations?select=kaggle_income.csv

This dataset has 32,000 records on US household income statistics and geo locations with granularity on neighborhood scale. This dataset has around 19 fields like the state, zip , county, mean, median.This documentation was provided by the U.S. Census Report retrieved on August 2, 2017.

Zillow economics data (County_time_series and Crosswalk) :

Source: https://www.kaggle.com/zillow/zecon?select=County_time_series.csv

This dataset has smoothed, seasonally adjusted measure of the typical home value and market changes across a given region and housing type. It reflects the typical value for homes in the 35th to 65th percentile range. This dataset has currently 10 variables with County Name, state name, state FIPS, County_FIPS,Metroname_zillow. Dataset county time series has median amount of listing price with county information. Variables in this dataset are date, region name, median listing price of 1 bedroom, 2 bedroom , 3 bedroom houses.

Crime Rates dataset (Crime rates by county) :

Source: https://www.kaggle.com/mikejohnsonjr/united-states-crime-rates-by-county

This dataset which has 24 columns has crime rates by county . arrest, murder, theft, burglary information.

School Dataset- USA Public schools:

Source: https://www.kaggle.com/carlosaguayo/usa-public-schools

This Public Schools feature dataset is composed of all Public elementary and secondary education facilities in the United States. This dataset has 33 attributes which contains the information of name of school, address, city, state, zip, district id and county.zip4 is not available for all the rows which can be ignored for our analysis.

Packages

The requried packages to start with are as follows.When we perform the study, we can add more packages when required.

- readr
- ggplot2
- corrplot
- plotly
- dplyr

Plots and table Needs

Plots

- Correlation plots are a great way of exploring data and seeing if there are any interaction terms.
- Density Plot using ggplot2 to visualize the distribution of the target variable.
- Histogram plot to view the residuals.
- Residual vs Fitted plot

Tables

- Summary table of the model
- Correlation of data

Questions for future steps

1. what are the transformations that can be applied on the dataset to get the desired data for our study?
2. How to combine the datasets chosen for study and what is the grain of each dataset. Is there any duplicates which needs to be deleted before combining the data.
3. What are the attributes which need to be considered in each file and the description and influence of each fields.
4. Does the data give any other information that need to be analysed further.