Week 8-9 - Exercise 8.3 Final Project Step - 1

Ganesh Kale

May 16, 2021

Introduction:

Topic: Predict House Prices in Austin TX USA

To tackle the problem statement given above I am going to follow CRISP-DM methodology that is widely used for handling data science projects. The first step in this process is understanding the business.

Business Understanding:

The topic for the final project I have chosen is predicting the housing prices from Austin Texas area. Real estate is always my favorite area to research in and with recent years it has been really risen. Housing market in USA continue to rise in all of the country's major cities. The Austin TX market is also one of the hottest market in 2021. The reason I chose the Austin Housing data because I stayed in TX for long time and Housing market was in boom that time and wanted to check the housing market in that area which is in high demand these days. To buy house in Austin area one should know the market trend and current house prices, also should know the house prices differ based on the area and features provided comprises the house price.

Problem Statement - To predict the house price based on the houses sold in that area.

The Austin TX Housing Prices data obtained from the Kaggle website link provided here Click here for Data Set Link The data was originally obtained from zillow website and cleaned and removed the unwanted columns by author.

The data set has total 47 unique features and it was uploaded with house images for each house listed in the data set and given the reference of each to the data set as an additional feature. For this project purpose, I have not included the house images since the research and analysis would be completely based on the features provided for each house instead of the photos. There are three files have been created and main file has all the information about houses, second file has information about the school area where the house is located and third file provides the house features information, that mean what are the features are there for each listed house.

Research Questions:

In order to analyze the problem first I need to understand the data and make sure got all the required data to handle this problem. To solve this problem and I am going to find answers to below question that would help in the analysis.

1. Do I have all the required data to solve the business problem?
2. Do I have the target variable?
3. What are the key features that drive house price change?
4. Is there any external feature that influence the house price or any other variable?
5. What are the assumptions need to made for Multiple linear regression?
6. What packages and statistical methods required to handle this problem?
7. What are the units of data features used to solve the problem and need any conversion?

Approach:

As mentioned above , this is the prediction problem, that means we need to perform the regression analysis on the data. Since we need to predict the Sale price of house given house features, area or neighborhood details I am going to use Multilinear Regression to solve this problem. To handle the business problem, the approach will be followed as below -

1. Load all the data in to separate data frames - To store data in data frames for Exploratory analysis.
2. Check the data summary and structure of the data - to understand what kind of variables are in the data and its types
3. Clean the data - For null/na values
4. Explore and Visualize the data - Plot different charts to see the data distribution, outliers etc.
5. Transform the data - change the categorical variable to numerical in order to fit to model, remove duplicates, perform aggregation to explore the data etc.
6. Check the data Distribution and Correlation - checking data distribution and correcting it by removing outliers if any and checking the correlation among the features to decide which one to keep and which one to remove.
7. Finalizing the Predictors - By using different techniques to select the predictors required for handling the regression problem. Creating new data frame with only required features.
8. Selecting Regression Model - Since this is multiple regression problem and we need to predict the house sale price, I am going to use Multiple linear Regression model.
9. Splitting the data for training and validation - Data will be splitted to train and test data sets and train will be used to train the model and then tested using test data set.
10. Fit Model - Run multiple regression model using the train data
11. Evaluate the result - The result generated by model will be evaluated
12. Predict the result - the result will be predicted using test data
13. Evaluate the assumptions - based on the test result run different statistics to validate the assumptions made earlier are met or not.
14. Conclusion - Based on the evaluation of result and assumption conclude the result about the model run on sample with produce same result on population and predict the accurate price.

Packages:

Data Understanding:

The second step of CRISP-DM is understanding data, here I am going to use this step to know more about data and tools/techniques required to handle the data.

To solve this problem statement using R programming language we need packages that would help us Load data, clean and Transform the data, visualize the data ,run the statistical models to evaluate the assumptions and test. 1. plyr - For Data wrangling 2. dplyr - For Data wrangling 3. ggplot2 - For data visualization 4. knitr - For creating Rmarkdown reports 5. tidyr - For data cleaning 6. QuantPsyc - For data Screening 7. car - For applied regression

Data:

As mentioned above, the Austin TX housing data was originally obtained from Zillow and consists of three different data sets. There are total 47 unique features in the all 3 data sets. First data set has all the information about house such as -

1. zpid - Zillow Property Id
2. city - City name
3. streetAddress - Address of House/Property
4. zipcode - Zip code of property
5. description - Property Description
6. latitude - Coordinates - location
7. longitude - Coordinates - location
8. propertyTaxRate - Property Tax rate in that area
9. garageSpaces - How many car parking spaces in garage
10. hasAssociation - Is HOA there or not
11. hasCooling - AC units are installed in house or not
12. hasGarage - Property has garage or not
13. hasHeating - Property has heating system or not
14. hasSpa - Property has spa or not
15. hasView - Property has view or not
16. homeType - Property is single family or apartment or townhouse etc
17. parkingSpaces - How many parking spaces
18. yearBuilt - What year property was built
19. latestPrice - WHat is latest house price
20. numPriceChanges - How many times property prices changes since listed
21. latest_saledate - Date of last sold
22. latest_salemonth - Month of last sold
23. latest_saleyear - Year of the last sold
24. latestPriceSource - The party provided the price of the property
25. lotSizeSqFt - Lot size in sq ft
26. livingAreaSqFt - Living area in sq ft
27. numOfBathrooms - Number of bathrooms
28. numOfBedrooms - Number of bedrooms
29. numOfStories - number of stories

Second Data Set has the information about schools in that area -

1. zpid - Zillow Property id
2. numOfPrimarySchools - Number of primary school in that area
3. numOfElementarySchools - Number of elementary school in that area
4. numOfMiddleSchools - Number of middle school in that area
5. numOfHighSchools - Number of high school in that area
6. avgSchoolDistance - The avg school disatnce from the property
7. avgSchoolRating - The school rating near to the house
8. avgSchoolSize - average school size
9. MedianStudentsPerTeacher - median student per teacher ratio

Third Data Set has information about house additional features -

1. zpid - Zillow Property id
2. numOfPhotos - Number of phots of property

3. numOfAccessibilityFeatures - number of accessibility features available
4. numOfAppliances - Number of appliences in the property
5. numOfParkingFeatures - Number of parking features
6. numOfPatioAndPorchFeatures - Number of patio and porch features
7. numOfSecurityFeatures - number of security features
8. numOfWaterfrontFeatures - Number of water front features
9. numOfWindowFeatures - Number of window features
10. numOfCommunityFeatures - Number of community features
11. homeImage - Reference of image file name

Plots and Tables:

To handle any problem we need data and to understand the data and see any pattern or information we need visualization. Plotting different charts to understand the data distribution, patterns etc I am going to use ggplot2 package and inbuilt functions such as plot(), hist() etc. The plots are used to analyze and explore the data -

1. Scatter plot - To see the relation between two variables and possible outliers
2. Histogram - To check data distribution for the different features
3. Q-Q plot - To check standard normal distribution
4. correlation matrix - To check the correlation of all the features from the data set
5. Boxplot - To check outliers and quantiles
6. Density plot - To check the distribution
7. Bar Chart - Data distribution of categorical variables.

Next Step:

To solve this problem, the steps mentioned in approach will be followed. The next step is to start looking into data and make data ready for regression analysis. Based on the knowledge gained so far Multiple regression is the righ choice to solve this problem but if the result produced by model does not meet the assumption of Multiple regression or deviates from it or result is not that accurate as it supposed to be, in such cases next step will be re-transforming the data and changing the predictors and re-run model or find another statistcal approach to solve the problem.