Week11_Final_Project

Akila Selvaraj

2/27/2022

House Prediction

Introduction

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. I am 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. I am trying to find the what is the impact of such predictors on the house price in this study. House Price prediction, is important to drive Real Estate efficiency. As 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. ion.

Problem statement

Goal of this study is to predict the house prices in a county in the next few months based on multiple factors. 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 I am going to use the option of multiple regression.

What is the predicted price of a house in a given zip code/ county in the next few months? Can we predict the price based on crime rates, schools and other metrics provided by Zillow for a county? Is this the correct time to buy a house? Where can we buy a house based on the budget we have?

This study tries to predict these answers with the help of various sources and by considering different factors which can affect the house prices.

Addressing problem statement

Analysis

To address the problem statement, I am planning to use data science framework and going to follow the below steps.

1. *Data Analysis/Exploratory 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, I will also try to identify high-unbalanced variables,minimum and maximum of each variable, and check the variable with outliers. Basically I will 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.

2. *Data cleaning/Feature Engineering* Here, I am 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.
3. *Feature selection* In this process, I am going to evaluate relationship between each individual variable and the target variable which in this case is sales price of the house. This helps to reduce the computational cost of modeling and to improve the performance of the model. I am planning 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 those variables will not be considered in the model.
4. *Modeling* Based on the input variables I 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.
5. *Model Evaluation* Once the model is created , I will evaluate the model by calculating the accuracy of the model by evaluating how model is performing on test data.

Data Collection

```
Initial step is to collect the data from various sources.
1. Zillow Economics Dataset - https://www.kaggle.com/zillow/zecon?select=County_time_series.csv
2. Crosswalk dataset - https://www.kaggle.com/zillow/zecon?select=CountyCrossWalk_Zillow.csv
3. Umemployment rate by county - https://www.kaggle.com/carlosaguayo/2018-unemployment-rate-by-county
4. Federal Reserve Interest rates - https://www.kaggle.com/federalreserve/interest-rates
5. Crime rate - https://www.kaggle.com/mikejohnsonjr/united-states-crime-rates-by-county
6. Public schools data - https://www.kaggle.com/carlosaguayo/usa-public-schools
7. Income dataset - https://www.kaggle.com/goldenoakresearch/us-household-income-stats-geo-locations?se
8. Zip_County_FIPS - https://data.world/niccolley/us-zipcode-to-county-state
```

Data Preparation

After collecting all the datasets, I analyzed the data to get an idea of what kind of information each and every dataset provides. The datasets I have collected has lot of attributes but not all are necessary for my study. So, I just selected the required ones.

I have chosen datasets such as income, crime rate, house pricing for each county, school to predict the house price. All the datasets have different level of information on different grain. If the data is not clean, the end result will not be accurate and as expected. The first step in data preparation is looking at the data and understand what information each file has. Also, each file has lot of fields which are not needed for my study. Understanding what each field tells and find out the fields relevant for my topic is the initial step performed. Once I figured out the columns required, then I filtered only those.

County_time_series I would say that county_time_series is the parent dataset which has House prices for each county and it has 82 columns. Most of the columns gives the information on median listing price based on bedroom level. The attributes which are useful for my topic is county related information and the house price. So, I have loaded only those attributes to my dataframe. It also has data starting from 1996. To keep only necessary volume, I am considering only the rows where sale year is greater than 2000.

Crime rate In crime rate dataset, the detail I am more interested on, is crime rate per county as I want to find how crime rate in that region impacts house price. Attributes which are not relevant here is detailed information about type of crime and the count of each and every crime, as I am not going in detail on which crime has impact over the price. I am filtering out only the county information and the crime rate for each county, other attributes like count of murder, robbery, burglary will be dropped. In this crime dataset, FIPS_ST and FIPS_CTY are two separate fields whereas in other datasets, these two fields are combined as a single column. As this is a crucial field which I would say as key field, I want this field to be in

correct format in sync with other datasets I am using. Before combining these two fields, FIPS_CTY will be converted to string and leading zeros will be concatenated. Then, both the fields will be concatenated so that FIPS_CD will match with other datasets.

School Dataset School dataset contains list of schools in each county and their location and the details about the school. As I am going to use school dataset to find if presence of school in the county influence house rates, what I need from this school dataset is number of schools in that county. Other details like school name, website information of the school are less relevant and I will drop those columns. To get the number of schools in each county, I will use summarize function and aggregate the data based on FIPS code to get the count of schools in each county.

*Income Dataset*Income Levels dataset has ZIP codes and no FIPS code, so I had to find a solution to convert ZIP codes to FIP codes. For this I had to find the dataset which links ZIPS and FIPS, and merge that with income dataset.

Unemployment dataset Unemployment dataset has unemployment rate for each region for each year. But the year information is column wise one column per year. To merge this dataset with my other datasets and to get the unemployment rate for each year, I had to pivot the data into row-wise so that it will have one row for each year. By this way, I can combine the sale year with the year and get the unemployment rate.

After preparing each dataset separately, I merged all these datasets and prepared one final dataset.

Data cleansing

Once all the datasets are combined, I had to clean the data by looking into missing values, outliers and bad values. Each missing value had to handled differently. For some of the missing values, if a value for particular county is missing, I calculated the mean on state level and replaced the missing values. For some of the missing values, I calculated the median and replaced missing values. Pricing dataset has records even before 2000. For my study, I considered the rows only after 2000. So, I sliced the data and reduced the volume and considered only the rows after 2000.

This is how the data looks like after performing slicing and dicing, and cleansing.

```
##   X FIPS       Date HousePrice StateName CountyName Median_Income
## 1 1 1001 2000-04-30      96300   Alabama    Autauga       48172.5
## 2 2 1001 2000-06-30      96200   Alabama    Autauga       48172.5
## 3 3 1001 2000-08-31      94500   Alabama    Autauga       48172.5
## 4 4 1001 2000-03-31      94400   Alabama    Autauga       48172.5
## 5 5 1001 2000-10-31      94200   Alabama    Autauga       48172.5
## 6 6 1001 2000-02-29      93100   Alabama    Autauga       48172.5
##   count_of_schools crime_rate_per_100000 Real_GDP_Percent_Change Inflation_Rate
## 1               15              251.6019                     7.80            2.3
## 2               15              251.6019                     2.95            2.5
## 3               15              251.6019                     2.95            2.6
## 4               15              251.6019                     2.95            2.4
## 5               15              251.6019                     2.30            2.5
## 6               15              251.6019                     2.95            2.2
##   unemp_rate
## 1          4
## 2          4
## 3          4
## 4          4
## 5          4
## 6          4
```

Feature selection

In machine learning, it is a general practice to rely on a correlation matrix to decide what features to be incorporated into our models. Table 3 presents the correlations between LRPi and each feature, and shows that housing floor area has the largest correlation with prices, followed by property age, travelling time to Central District and floor level. In comparison, correlations between each orientation and prices are very close to zero and so they are excluded from our estimation.

Correlation Matrix

```
library("corrplot")
```

```
## corrplot 0.92 loaded
```

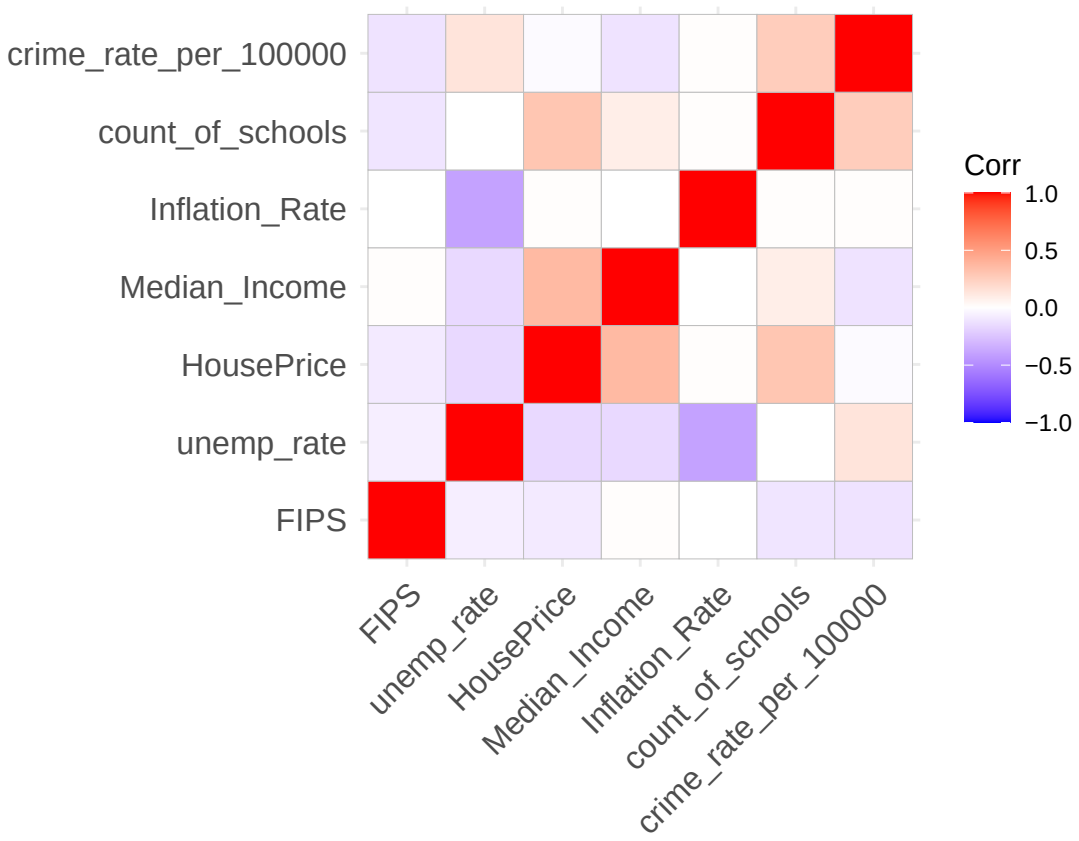
```
cor(House_price_df[, c('FIPS','HousePrice','Median_Income','count_of_schools','crime_rate_per_100000','
```

```
##                                FIPS   HousePrice Median_Income
## FIPS                    1.0000000000 -0.092653701  9.199414e-03
## HousePrice             -0.0926537010  1.000000000  3.663168e-01
## Median_Income           0.0091994136  0.366316807  1.000000e+00
## count_of_schools       -0.1064315443  0.298684081  8.985717e-02
## crime_rate_per_100000  -0.1249954448 -0.019390585 -1.178404e-01
## unemp_rate             -0.0756160292 -0.151127165 -1.649863e-01
## Inflation_Rate          0.0001851431  0.008710466  3.171831e-03
## Real_GDP_Percent_Change 0.0011038637 -0.028270359  9.956758e-05
##                        count_of_schools crime_rate_per_100000   unemp_rate
## FIPS                      -0.1064315443         -0.1249954448 -0.075616029
## HousePrice                 0.2986840809         -0.0193905852 -0.151127165
## Median_Income              0.0898571664         -0.1178403849 -0.164986299
## count_of_schools           1.0000000000          0.2618926924 -0.002570946
## crime_rate_per_100000      0.2618926924          1.0000000000  0.136564642
## unemp_rate                -0.0025709456          0.1365646416  1.000000000
## Inflation_Rate             0.0127920230          0.0069060351 -0.407834008
## Real_GDP_Percent_Change   -0.0002557675         -0.0006271534 -0.072630015
##                        Inflation_Rate Real_GDP_Percent_Change
## FIPS                      0.0001851431            1.103864e-03
## HousePrice                0.0087104657           -2.827036e-02
## Median_Income             0.0031718310            9.956758e-05
## count_of_schools          0.0127920230           -2.557675e-04
## crime_rate_per_100000     0.0069060351           -6.271534e-04
## unemp_rate               -0.4078340079           -7.263001e-02
## Inflation_Rate            1.0000000000           -2.306954e-01
## Real_GDP_Percent_Change  -0.2306954070            1.000000e+00
```

```
#cor(House_price_df)
#cor.test(House_price_df$Median_Income,House_price_df$HousePrice )
```

Plot Correlation Matrix

```
library("corrplot")
library("ggcorrplot")
library("ggplot2")
cor_data <- cor(House_price_df[, c('FIPS','HousePrice','Median_Income','count_of_schools','crime_rate_p
ggcorrplot(cor_data, hc.order = TRUE, method ="square")
```



From the final dataset I prepared, I selected state, county, house price, sale year, median income, crime rate, count of schools, GDP percentage, inflation rate and unemp rate. R2 and adjusted R2 states that the fields GDP rate has not much influence of the house price. So, I decided to take that feature off from my study.

Model Implementation:

I am planning to incorporate multiple linear regression technique to answer the research questions.

Model building

Summary of the model

```
##
## Call:
## lm(formula = HousePrice ~ Median_Income + count_of_schools +
##     crime_rate_per_100000 + unemp_rate + Inflation_Rate, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -513677  -42178  -15500   22134 1334197
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            1.302e+05  1.183e+03  110.08   <2e-16 ***
## Median_Income          9.906e-01  5.827e-03  170.00   <2e-16 ***
## count_of_schools       2.513e+02  1.662e+00  151.25   <2e-16 ***
## crime_rate_per_100000 -1.844e+01  8.891e-01  -20.75   <2e-16 ***
```

```
## unemp_rate            -4.366e+03  7.428e+01  -58.78   <2e-16 ***
## Inflation_Rate        -8.593e+03  4.083e+02  -21.05   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 83920 on 244023 degrees of freedom
## Multiple R-squared:  0.2169, Adjusted R-squared:  0.2169
## F-statistic: 1.352e+04 on 5 and 244023 DF,  p-value: < 2.2e-16
```

Implications

Getting an overview of the correlation helped to understand the difference between the two datasets. It showed the value of the variables and their effect on the prediction.

Data processing and feature engineering are crucial in machine learning to build a prediction model. Furthermore, a model cannot be made without some data processing. For instance, as shown in the study, the model could not be trained before handling the missing values and converting the text in the dataset into numerical values. From the study, we saw that pre-processing the data does improve the prediction accuracy and matches the result.

The independent factors that have been discussed and the features of houses in the local data are combined in order to study the correlation between them and the sale price of house. Our results shows that crime rate and unemployment rate has a negative correlation with sale price indicating that when these factors increase the house price decrease. In addition,Median Income, count of schools, inflation have a positive correlation with the sale price. It means when these factors increase the sale price increase.

Limitations

The calculation of house prices are done without the necessary prediction about future market trends and price increase. The factors that have been studied in this study has a slightly weak correlation with the sale price. Hence, more factors can be added to the dataset that affect the house price, such as population,lending, deposit rates, major occupation for better finding of my research questions.

From the results, it can be concluded that the proposed multiple linear regression model can effectively analyze and predict the housing price to some extent. Admittedly, the prediction accuracy is still limited at specific points, and the model still needs to be improved in further research.

This paper considered only the current year's information of the houses. The time effect of the house price, which could potentially impact the estimated results was not included (the same house should have different price in different years, assuming that age factor is constant). Finally, the house price could be affected by some other economic factors such as exchange rate are not included in the estimation.

Conclusion

In this study, I built multiple regression model to predict the price of house based on some of the house features.I also evaluated the model to determine the performance of the model. I also followed the data science process starting with data collection, then cleaning and preprocessing of the data, followed by exploring the data and building models, then evaluating the results and communicating with visualizations. I have also mentioned the step by step procedure to analyze the dataset and finding the correlation between the parameters. Thus I can select the parameters which are not correlated to each other and are independent in nature. These feature set were then given as an input to multiple regression model and predicted the house prices.