MONTGOMERY COUNTY HOUSE SALES PREDICTION

Tamrat Workineh
Department of Data Science, University of Maryland Baltimore County
DATA 602- Introduction to Machine Learning

Professor Murat Guner (Ph.D.)

December 8, 2020

Table of contents

Acknowledgment:

I would like to thank my professor Dr Murat Guner (Ph.D.) for unreserved support and excellent teaching methodology he used for students to grasp the objective of the course. For his so keen to help students who ask his help and encouraging students to participate in the class activities and do homework assignments on time.

1. Introduction

1.1 Context

The "Montgomery County House Sales Price Prediction " is a project prepared after attending a three-month course organized for grad students in partial fulfillment of the course DATA602- Introduction to Machine Language for Data Science program at the University of Maryland Baltimore County.

1.2 Motivations

House price varies in the US from place to place and from time to time. I also noticed a couple of my friends struggling to find their dream houses. As a Data Science student attending the Machine Learning course, I gained the skills to apply in real problems. I chose the above topic to manipulate regression analysis tools.

1.3 Goal

I am aware of the availability of several dataset from open resources. As a first experience, I preferred to engage in the house sale regression analysis algorithm and predict to focus on some features in the dataset.

1.4 Source

The data was obtained from the state of Maryland open data source. The dataset contains 1500 rows and 13 Columns. The dataset also has the following features: ID, Date, House Price, Bedrooms, Bathrooms, Living room in sq. ft, Lot in sq., Floors of the house in sq. feet, condition of the house (Excellent, Very Good, Good , or Require Refurbishment) grade, Basements, year built , Zip Code and Latitude, Longitude and year built .

2. The Project

2.1 Data

The central element in data analysis is understanding of the data and setting a specific and achievable goal. For this project after finding the relevant data and setting goals , the following machine learning steps were taken to get, clean and transform the data.

2.1.1 Getting the data

The first problem was where to find the data relevant for the intended project and build a large enough dataset to implement the science I want to be able to predict the price for a given house . Thank God , as it is mentioned in the previous section , after obtaining the dataset , it was imported on Jupyter notebook. All the necessary machine learning tools were also imported from Panda , scikit-learn and python libraries

2.2 Data Exploration

Data exploration has been made to uncover initial patterns, characteristics, and points of interest. The following graph and heatmap created using seaborn and Matplotlib shows the dataset collected for analysis.

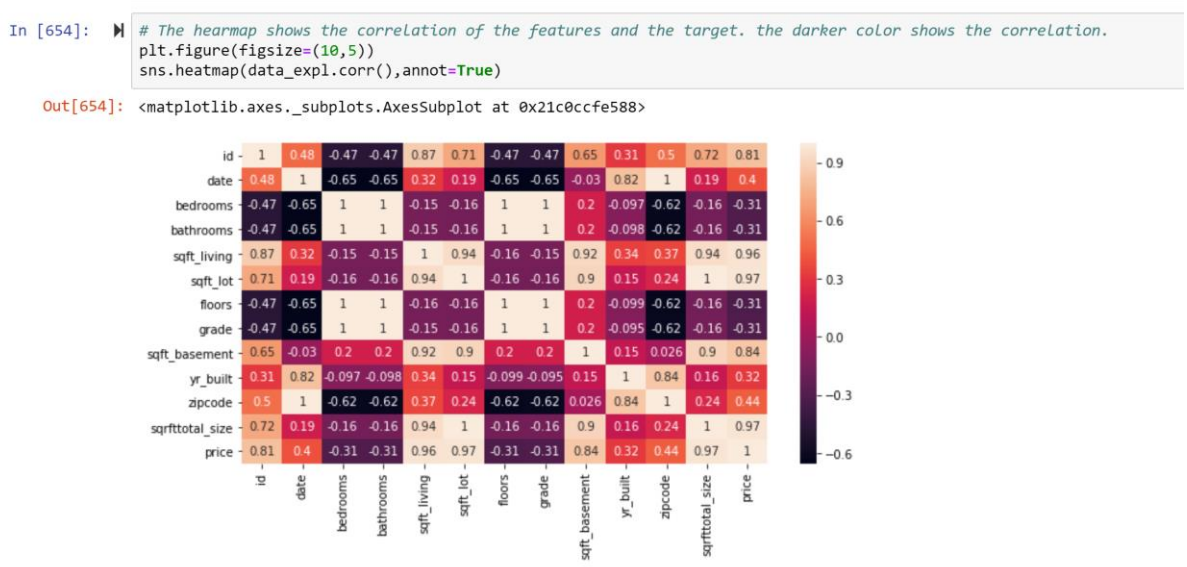

Figure 1 above depicts the heatmap showing the original dataset. From the graph the white diagonal refers to the ID. Which should be cleared later. The graph has two triangles the lighter color indicates the features are highly correlated.

2.2.1 Data Correlation

Data correlation is a static that evaluate how closely two variables move together or in oppostite directions or even if they are related at all. So it ranges in value from negative 1 which means perfectly negatively correlated in that when one decrease the other also decreases or 1 perfectly positively correlated. if one increases the other also increase. if the correlation is zero ,then the relationship will be weak or nor relationship at all.

```
In [658]: data_expl.corr()['price'].sort_values (ascending =False)

Out[658]: price              1.000000
          sqrfttotal_size    0.973793
          sqft_lot           0.972918
          sqft_living        0.962604
          sqft_basement      0.844460
          id                 0.813699
          zipcode            0.440844
          date               0.399209
          yr_built           0.315419
          bedrooms          -0.308582
          grade             -0.309220
          bathrooms         -0.309603
          floors            -0.312281
          Name: price, dtype: float64
```

Table 1: The graph above show positive correlation between price and total size of the house.

2.3 Data Preparation (Preprocessing)

Data preprocessing was conducted to ensure the accuracy, efficiency, or meaningful analysis. The dataset which was originally above 20 thousand rows and and 12 columns were reduced to manageable size as mentined in the previous section. Besides, irrelevant dates were removed to get rid of the noise, and checked that all the values are not empty, otherwise the item is dropped.

2.4 Visualizing the Data

visualization is used to explore and communicate the findings and is the next phase of the data analytics phase

```
In [681]: plt.figure(figsize=(10,5))
          sns.heatmap(tr.corr(),annot=True)

Out[681]: <matplotlib.axes._subplots.AxesSubplot at 0x21c8ebfe908>
```

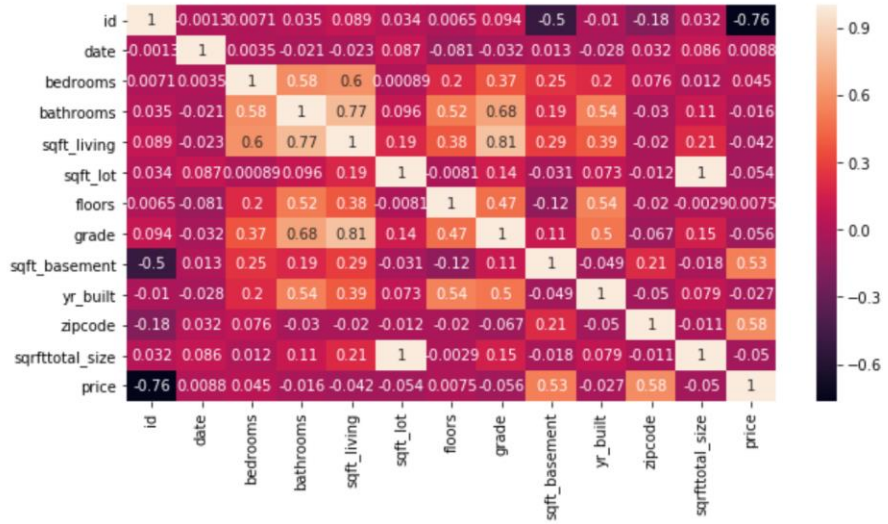

Figure 2 above depicts the heatmap showing the after preprocessing the dataset. From the graph the white diagonal refers to the ID. Which should be cleared later. The graph has two triangles the lighter color indicates the features are highly correlated.

2.5 Splitting Data into Training and evaluation sets

The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the mode.

Splitting data into training and evaluation sets

The final step is to split your data into two sets; one for training your algorithm, and another for evaluation purposes

```
In [671]: shuffled_data = mydata.sample(frac=1., random_state=42)
          shuffled_data.head()
```

Out[671]:

	id	date	bedrooms	bathrooms	sqft_living	sqft_lot	floors	grade	sqft_basement	yr_built	zipcode	sqrfttotal_size	price
1116	349400100	20160908	3	1.75	1480	7830	1.0	7	0	1980	20814	9310	1280000.0
1368	624100010	20161208	3	2.50	2930	19900	1.5	9	0	1983	20816	22830	1070000.0
422	476000017	20161003	2	2.00	1400	1512	2.0	8	460	2006	20814	3372	1180000.0
413	7600065	20160605	3	2.25	1530	1245	2.0	9	480	2016	20815	3255	3710000.0
451	567000755	20161205	2	3.00	1790	1709	2.0	7	390	2001	20816	3889	1100000.0

Table 2 In this project, the dataset was splitted 80% train and 20% test to discover how to evaluate machine learning models using the train-test split.

2.6 Building a Model

In this section, linear regression, ridge regression, and lasso regression will be compared and see how well each model makes a prediction.

2.6.1 Linear Regression

Linear Regression Analysis: regression analysis used to essentially crunch the numbers to help to make better decisions for the prediction. The regression analysis used to predict the future sales.

```
In [678]: plt.scatter(tr['sqrfttotal_size'] , tr['price'])
```

```
Out[678]: <matplotlib.collections.PathCollection at 0x21c0ce4dc50>
```

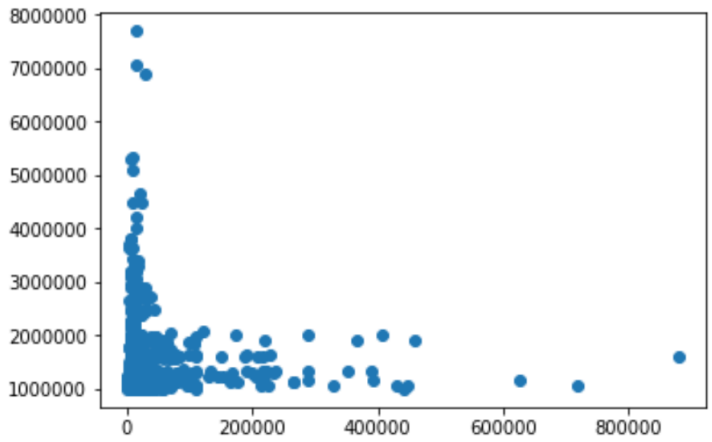

Table 3 The regression model show that there is outlier which greatly affect the result of the algorithm.

- **Outliers Detection**

According to different measures and charts of my data, there was no doubt about the presence of outliers. The outliers, in my case, are tuple taking abnormal value such as very large or very small, even 0, in one or many of variables. These outliers can affect greatly the results of my learning algorithm. They are several types of outliers :

2.6.2 LASSO Regression model

LASSO is a regression model that does variable selection and regularization. The LASSO model uses a parameter that penalizes fitting too many variables. It allows the shrinkage of variable coefficients to 0, which essentially results in those variables having no effect in the model, thereby reducing dimensionality. Since there are quite a few explanatory variables, reducing the number of variables may increase interpretability and prediction accuracy.

2.7 Result

2.7.1 Here are the scores I received from my two models

- **Linear Regression**

```
In [695]: reg=LinearRegression()
          reg.fit(tr[["sqrfttotal_size"]],tr.price)

Out[695]: LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None,
                           normalize=False)

In [696]: reg.score(tr[["sqrfttotal_size"]],tr.price)

Out[696]: 0.002505758812497927
```

- **LASSO Regression**

Lasso Regression

Lasso regression is used for shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean.

```
[703]: from sklearn import linear_model
       lasso_reg  =  linear_model . Lasso ( alpha = 50 , max_iter = 100 , tol = 0.1 )
       lasso_reg.fit(X,y)

Out[703]: Lasso(alpha=50, copy_X=True, fit_intercept=True, max_iter=100,
             normalize=False, positive=False, precompute=False, random_state=None,
             selection='cyclic', tol=0.1, warm_start=False)

[704]: lasso_reg.score(X_train, y_train)

Out[704]: 0.8139867399955776
```

2.1 Evaluation

The training set and the test set contain 80% and 20% of the total sample, respectively. To evaluate the forecasting accuracy of both models, (Regression and Lasso)an out of sample forecasting is operated, subsequently, the later model with 0.81 accuracy was considered to be a relatively superior model.

3 Conclusion

Based on the analysis conducted to predict the price of house sale in Montgomery County for classification problem, the best-performing model was Lasso which shrink the data and reduced the outlier to perform 0.81 accuracy rate the best-performing model to resolve the regression problem.

4 Future Work:

This project is my first experience. I planned to work on same project with additional data models and optimize the performance.

References

Allen C. G., 1977, “Hedonic Prices, Price Indices and Housing Markets”, Center for metropolitan the Johns Hopkins , University, Baltimore, Maryland.

“ The home of the U.S. Government’s open data,” Data.gov

Frew J., and G. D. Jud, 2003, “Estimating the Value of Townhome Buildings”, The Journal of Real Estate Research, 25(1): 77 - 86.