CC7184 DATA MINING AND MACHINE LEARNING

Coursework Report

Module leader- Dr. Qicheng Yu

Project title

To examine the impact of crime rate on house price in London boroughs (using CRISP- DM model)

Student Name- **Jiji Ajitha Kumari Venugopalan Assari**

London Met ID- **20033188**

Email-jia0098@my.londonmet.ac.uk

Table of Contents

1.Introduction

1.1 Business understanding

Empirical studies urge that house pricing is one of the key components of the aggregate expenses that are used in the analysis of the household welfare status. Literature shows the variation of the share of the value of the house in the household aggregated expenditure. Although the share varies across the countries it is worth noting that the house price forms a significant great portion of the family expenditure.

The characteristic of the environment where the house is situated plays a key role in the determination of the price. These features may include poverty, racial composition, and crime rate among others. According to the classical ecological theory, as cited by Olivier et al. (2020), urban migration to the concentrated area can be characterized by a high crime rate. Therefore, empirical studies have associated crime as a key catalyst for the social-economic changes in the community.

The cumulative crime rate effect will directly influence society and the individual security perception. As a result of the deterioration of the security persist people tend to move to places, that seem to be more secure. As a catalyst, the crime rate also affects the family and individual economically. The impact is felt by both victims and non-victims. For instance, the if crime rate suppresses the house price the resident get impacted negatively. For instance, as house ownership increases, the crime rate threatens the appeal and declines the ownership desirability. This, in turn, makes the household remain to rent. This increases demand for the rental houses in the secured places hence increasing their pricing.

As a fact, it is clear that the impact of the crime rate does not end with the usual effects. These effects include bodily injury, police protection costs, and the loss of property. However, the impact of the crime rate especially on the slums and the congested areas can be seen in the house pricing. As people avoid the crime risk places for the secure places, the house demand raise. This in turn causes the house prices in so-called secure places to increase.

1.2 Data understanding

In this coursework, the CRISP-DM (Cross Industry Standard Process for Data Mining) Model, were developed using python programming language to model the average house price in the UK. The features used as the factors in the modeling include the total number of houses sold, the number of crimes committed, and if or not the house is in a borough city.

The coursework seeks to determine the impact of the number of crimes on the house price. \the target variable, price of the house, is continuous rather than categorical. The impact of the number of houses sold on the average house price. And the impact that borough city has on the house price. I created a meaningful new dataset from the different sources of data through cleaning and transforming.

The analysis data was retrieved online via the following link.

https://www.kaggle.com/code/fixfon/real-estate-price-changes-in-london/data?select=housing_in_london_monthly_variables.csv.
https://data.london.gov.uk/dataset/uk-house-price-index
https://data.london.gov.uk/dataset/recorded_crime_summary

2. Literature review

Several empirical studies have explained the impact of the crime rate on the house pricing. One of the main concerns is the demographic changes that are associated with the crime rates. Researcher have shown that crime rate will influence how people perceive security of their cities. According to (Florida, 2021), urban-urban migration sometime is as a result of less threating locations.

As cited by (Chen, 2020), the US cites from 1980 to the peak of 2006 have experience increase in home ownership and low crime rate. For instance, in Los Angeles the median house price rose from approximate $200000 in late 1980s to around $55000 in the year 2006. Researchers such as (Wong, 2019), (Ceccato, 2020), and (Beck, 2018) associates crime rates negatively with the average house price. High crime rate is associated with constraining property value which increases the value in some extent. The study conducted by (Lyndsay, 2018) the impacts of different kinds of the crime was measured. The study shows that high crime rate leads to people selling their properties and move to some safe places. The transaction of the resident makes overcrowding to so called safe places. On the other hand, the house selling due to crime rate increases the poverty concentration. This is because only the low-income individuals will buy houses in the crime prone areas. Thus, this will be a loose to the seller.

As cited by Deaton and Zaidi (2012) there are two approaches to estimating the self-owned house price. The hedonic pricing and the implicit rental value. They continue to add that implicit rental values entail asking the house owner the much they could have paid if they were renting. On the other hand, the hedonic pricing method is a process of modeling the house price as a factor of the house feature. This includes the use of machine learning algorithms such as the ordinary least square methods to model the house price.

3. Data preparation

3.1 Data selection and Preprocessing

The house price data contains a total of 13549 observations. The observations include the average house price in London between the 1st January 1995 to the 1st January 2020. Besides the data contains a total of 7 features. These include; the data which shows the specific date when the observation was taken. The specific area in the United Kingdom. The average house price shows the mean of the house price sold on that day in Euros. The code is the unique identification of the daily observations. The number of houses sold and the number of crimes done. the last feature was the borough flag which shows if the observations were made from the borough city or not.

3.2 Data Measurement setting and Transforming

For the analysis purpose in this article, the date, area, and the code column were omitted for the analysis. Further, the data was found to have missing values, especially on the crime rates and the number of houses sold columns. For the analysis, the missing values were replaced by zeroes (0). Further, the mean average price was found to be 263519.7 euros. This was found to range between 13549 euros and 1463378 euros. The average number of houses sold was found to be 3866.98, the number of the house price was found to range between a minimum of 0 to 132163 houses. The number of crimes was found to range from zero crimes in a day to 7461 crimes in a day. The mean number of crimes was found to be 1185.03 crimes per day. Besides, among the 13549 sampled days, 9936 were taken from the borough city while 3613 were not from the city.

Additionally, the study seeks to determine the distribution of the features and the relationships. For instance, Figure1 shows a positive association between the average house price and the crime rate. However, the average house price was found to be negatively associated with the number of houses sold, See in Figure 2

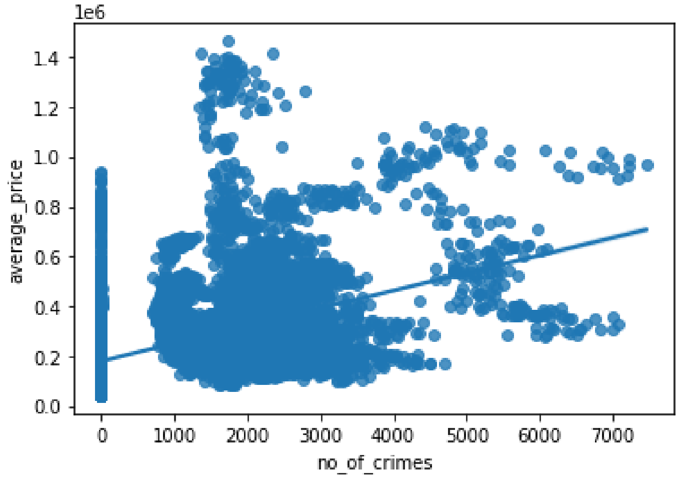

Figure 1. A scatterplot for the average house price and the number of crimes.

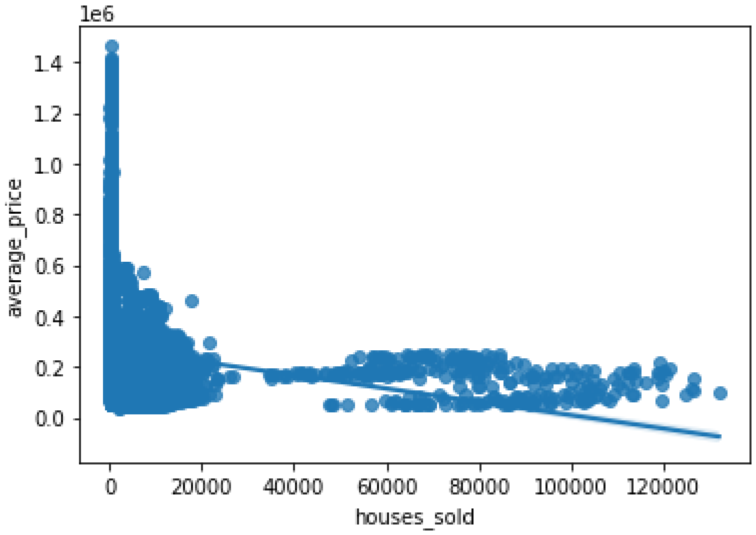

Figure 2. Scatterplot for the average house price and the number of houses sold

Considering the distribution of the average house price. Figure 3 shows that the house price was skewed to the left. Besides the mean average house price was found to be high in the house from the borough city, see in Figure 4.

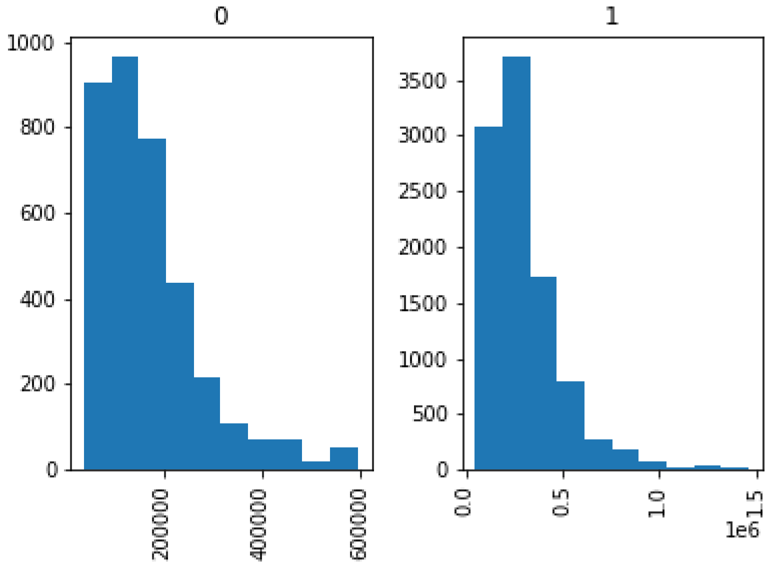

Figure 3. The distribution of the average house price

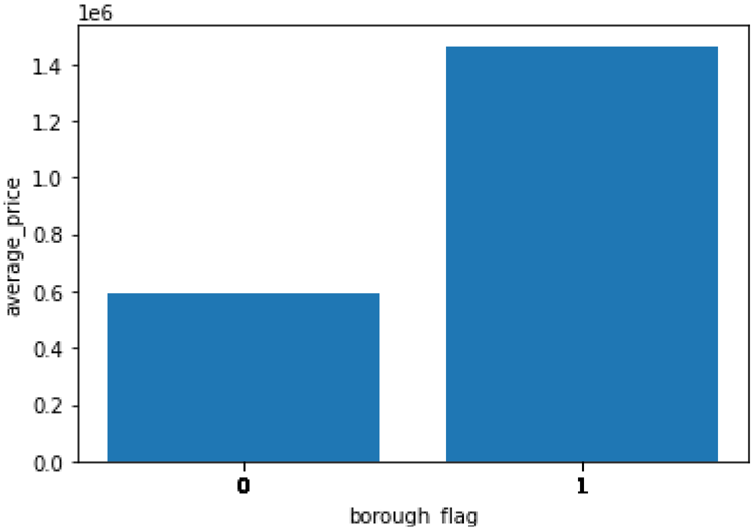

Figure 4. The mean average house price between the borough and others cities.

4. Modelling processing

4.1. Modelling Techniques

4.1.1. Linear Regression Model

Firstly, the ordinally least square method was used to model the average house price. Linear regression is to study the linear relationship between the dependent and independent variables. The condition for Linear Regression is that the dependent variable must be continuous and the independent variable may be continuous, binary or categorical.

4.1.2. Decision Tree

Decision trees are used for both classification and regression. Decision Tree Regression was selected because of the reason that regression tree takes continuous features and the mean of values from a group of datapoints.

4.1.3. Random Forest Regression Model

Random Forest is generally a classification technique which prevents overfitting by creating random subset of the features. In this coursework, I am also going to forecast housing price against crime rate in different London boroughs using Random Forest regression models. This algorithm is stable because of the average value is used. First, we need to import random forest Regressor from sklearn.

4.2. Model executing and test designs

Before determining the best model, we need to consider a training dataset and model the relationship of the variables with the three mentioned Regression techniques. This should be such that it would efficiently predict the new data samples.

The data partitioning was conducted before modeling. This was to set the data into a training and testing set. We consider 80% of the original data as training data and 20% as the testing data. Linear Regression model was considered in two cases where the number of houses sold, the number of crimes, and the borough flag were taken as the explanatory variables in the first case and the borough flag was left out as an explanatory variable in the second case. Then Random Forest Regression Model and Decision Tree Regression Model was executed using the second case.

4.3. Assessing models

The models were used to predict the testing data and the error and variability was checked to get the best predictive model. The residual plot was displayed to check the residual values to assess the model fit. The normality of residuals can be checked using the QQ plot.

5. Evaluation and results

The regression summary for linear regression model is given in table 1. This shows that considering the first model the mean average house price was found to be 177637.9312 this was expected to decrease by 0.5249 with a unit increase in the number of houses sold. Besides, a unit increase in the number of crimes rate was expected to increase the average house price by 69.4688. Concerning the houses sold in other places, the expected average price was expected to be 7445.315 more euros. Conversely considering the second linear model. The expected average house price was found to be 182059.5524 euros. This was expected to decrease by 0.6231 with a unit increase in the number of houses sold. Besides, a unit increase in the number of crimes rate was expected to increase the average house price by 70.6765.

variable	Model 1	Model2
intercept	177637.9312	182059.5524
House Sold	-0.5249	-0.6231
No of crime	69.4688	70.6765
Borough flag	7445.315	

Table 1. The summary of the Linear models

To compare the model performance in the case where the number of houses sold and the num ber of crimes were taken as the explanatory variables, performance metrics such as mean abso lute error, mean squared error and the root mean squared error was computed as below.

	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error	Varianc e Score
Linear model Regressor	114110.47596901383	27613794477.228622	166173.9885698981	0.20
Decision tree Regressor	101679.81962143251	22540803908.762657	150135.9514199136	0.34
Random Forest Regressor	100306.29975953905	22954144478.09397	151506.25227393743	0.33

Table 2: Error Comparison for the three models

The above table shows comparison of Mean absolute error, mean squared error, Root Mean squared error and the explained variance score of three models. When I plot residuals for Random Forest Regressor Model, it shows value of R-squared is 0.876 or basically 87.6% for train data and for test data, it shows 0.331 or basically 33.1%. The mean absolute error of the model is 100306.29975953905 and the variance is 0.33. When it compared to other two models, Random Forest regressor has low absolute error and the variance is 0.337. Lower the Mean absolute error, Mean Squared error and Root mean squared error, better the model. The ideal score of the ‘explained variance score’ should be between 0.60 and 1.0. However, none of our models satisfy this condition. The reason may be that we have considered two response variables and this may not be sufficient to calculate the target variable.

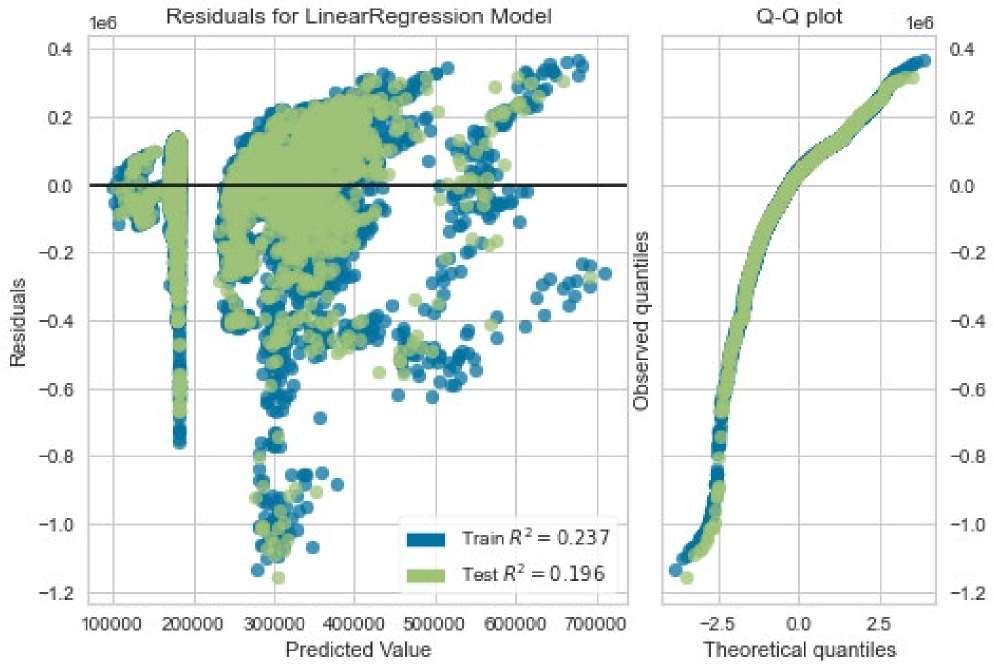

Figure 5: Residual Plot and QQ Plot for Linear Regression

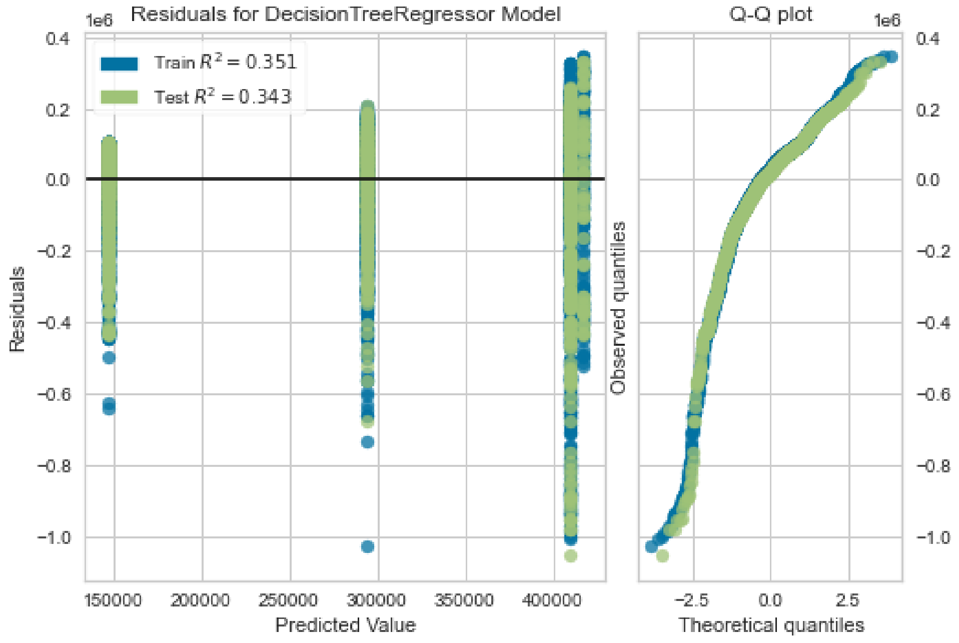

Figure 6: Residual Plot and QQ Plot for Decision Tree Regression

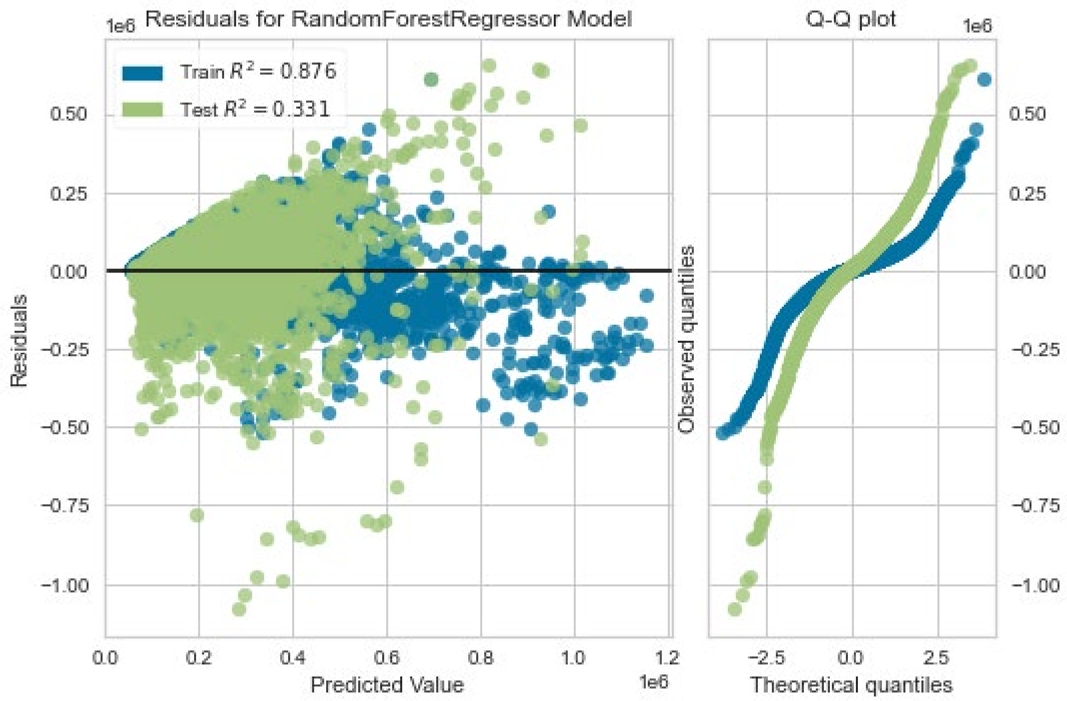

Figure 7: Residual Plot and QQ Plot for Random Forest Regression

When comparing Random Forest and Decision Tree, residuals are more random in Random Forest Regression compared to Decision Tree. Residuals for decision tree is not random so we cannot say it is a model that fits data.

6. Plan for deployment

The impact of crime on society is widely spread. The analysis from the current coursework was suggest that the house price would increase as the number of crimes increase. This suggests that the development of the country is localized. Here we mean it pleases that are at risk of crime people tend not to reside or buy houses in them. This in turn some places are preferred over others. The long-term effect of this is contestation and challenges for the social and amenity.

Considering the effect of the number of houses sold a day. The results show it hurt the house price. They could be associated with the increase in the construction of the standard cheap house. Further, a high number of real estate development can be a major contributor to low prices. Besides, the urban house tends to be of high price than rural. This can be attributed to the reason why the houses in the average house price in borough city.

To harmonize the house price, there is a need for tightening the security by the ruling government. Further, the security can be enhanced by delocalizing and unforming the spread of the development project across the country. This will make the rural places have competitive house prices as those in town. This in turn will increase the number of houses sold.

8. Conclusion

Average house price has been a great concern for many years. This issue has persisted to an extent of some government initiating affordable housing program. As one of the sustainable development goals there is need to have affordable housing for all for quality and development of the nation. However, the house pricing does not leave in isolation. Some of the factors include social demographic characteristics of the location where the house is located. To a greater interest the current study focuses on the impact of the crime and the number of houses sold on the average house price. The study shows that the house average price will increase depending on either it is on the rural areas or in town. The house average house price was found to decrease as the number of crimes increase. Therefore, the number of the crime can be associated with the increase in the house price in so called safe areas. In conclusion, the number of the house sold was found to decrease the average house price.

The study recommends even distribution of the development and the security across the country. This will enhance equality among all parts of the nation. This will contribute to deco gestation of the prior called safe location and the population will be evenly distributed. This increases the number of houses sold while reducing their price. On the other hand, if the supply of the house is above the demand it will create crises for the real estate dealers making them make losses. Therefore, there is a need to harmonize the supply and demand for the houses.

9.References

1. Ian H. Witten, Eibe Frank, Mark A. Hall, Christopher J. Pal (2016) Data Mining: Practical Machine Learning Tools and Techniques, Elsevier Science
2. Mohammed J. Zaki and Wagner Meira, Jr., (2020) Data Mining and Machine Learning: Fundamental Concepts and Algorithms, 2nd ed. Cambridge University Press
3. Berry, M. J. A., and Linoff, G. S. (2004). Data Mining Techniques: for Marketing, Sales and Customer Support. 2nd edition. Wiley. ISBN 0–471-47064-3
4. Cross Industry Standard Process for Data Mining http://www.crisp-dm.org/
5. Mohammed J. Zaki and Wagner Meira, Jr., (2020) Data Mining and Machine Learning: Fundamental Concepts and Algorithms, 2nd ed. Cambridge University Press
6. Deaton, A.&.(2012).Guidelines for constructing consumption aggregates for welfare analysis(Vol.135).World Bank Publications.
7. Dustmann, C.F.(2018). Housing expenditures and income inequality.ZEW-Centre for European Economic Research Discussion Paper,(18-048)
8. Ollivier,M.L.(2020).Characterizing ecological interaction networks to support risk assessment in classical biological control of weeds.Current opinion in Insect Science,.,38,40-47
9. https://www.kaggle.com/code/fixfon/real-estate-price-changes-in-london/data?select=housing_in_london_monthly_variables.csv.