Data Mining and Machine Learning 1 (MSCDAD_A) Project Report

Umer Iqbal - x17111854
Course: Master of Science in Data Analytics(A)

Abstract— Is prices of houses is rising at sharper rate in Dublin and London cities? What would be the price of house if we have 4 bedrooms and 2000 square feet area. The answer to these questions is essential for the discussion on buying or selling properties in Dublin and London. Whereas answer to all questions will be answered in this report. Machine learning techniques that have been used for this report are time series, linear regression, multiple linear regression, decision tree, Lasso regression, OLS multiple linear regression.

In our analysis, property prices of London and Dublin datasets have been taken from Kaggle and property price register. Where data mining techniques helped to predict prices of houses of London and Dublin and came up with decision for buying property in Dublin and London. We imported our datasets in RStudio for pre-processing and run machine learning algorithm.

INTRODUCTION

House prices prediction can help property developer to decide selling price of house and can help customers to arrange correct time to purchase house. House price prediction can help those people who are planning to buy houses, so they can have an idea in advance what price should be in coming days or what should be price if they want to buy a house with 3 bedrooms. This report analyses the house price prediction of London and Dublin e.g., What would be the price of house if we have 4 bedrooms and 2000 square feet area. It will also prove visualisations/insights on each dataset. Literature review will be examined in this document in relative to their impact on our approach how they are separate from our implementation/approach, what others machine learning model has been used how accurate they are, and what model we should focused on during our analysis.

During our analysis KDD methodology will be used as part of our data mining and machine learning. It is very helpful for extracting knowledge from large datasets. KDD consists of 5 major methods, Data selection (How our data should be selected in terms of our project), Data pre-processing (clean our datasets, if necessary, check null values etc.), Data transformation (performed exploratory data analysis, creating new data frame from existing datasets, performed principal component analysis to see dimensionality of our London housing dataset), Data mining (apply machine learning methods), and evaluate/interpretate our analysis. At end of report, conclusion and future work will be discussed and further we can use in future. In this part, all question will be answered e.g., what would be price of house in Dublin with 3 bedrooms house etc.

RELATED WORKS

Movements in house price indicate current economic situations, and it also concerns buyers and sellers. A lot of factors can be used that have an impact on house prices e.g., number of bedrooms and number of bathrooms. Whereas house price prediction also depends on its locations. A house which has access to highways, schools, grocery shops, shopping malls, opportunity to employment as compared to houses with no accessibility. "According to Irish times by Eoin Berk-Kennedy on 25th August 2021" House prices inflation is likely to move up to 12% this year, before it fall back roughly in 2022 as supply comes on and affordability constraints kick in, the chief executive of DNG state agent said. Keith Lowe said recent increase in prices was being driven by covid-related factor e.g., increases savings. [1]

It is mostly believed by academia, that accurately predicted the special price for specific real estate is impractical since it involves loads of factors, tries on final cost. By "*Li Ka-Shinga*", a famous property tycoon in Hong Kong, three major factors for determining property prices, first one is location, the second one is location and third one is also location. A particular factor like location plays a vital role in predicting prices of houses and it is worth to explore by adopting a statistical model. [2]

According to economics principal, market price of properties is attained when necessity/demands and supply curves and intersect each other. It is unlikely that market price of property is going to be equal to market values, whereas market for real estate has been too unpredictable and fluctuating to be considered as an ideal market. [3]

Another example of related work was a paper by Park and Bae in 2015, develop a housing price prediction model with machine learning algorithm e.g., Naïve, decision tree, RIPPER etc. and evaluate their performance in terms of classification accuracy. Aims of their research is to help property sellers to make rational decision in property transactions. Their model determine that RIPPER algorithm based on accuracy, constantly outperforms other models in performance of house price predictions. [4]

Satish et al in an article "*Predicting property prices with machine learning algorithm on Feb 2020*" he performed many machines learning algorithm e.g., linear regression, multiple linear regression, gradient boosting algorithm, Lasso regression to predict house prices. Based on their results lasso algorithm outperforms other algorithm based on their accuracy. A records of housing transection for three cities Los Angeles, California, and Huang applied both linear and nonlinear ML algorithm to predict house prices. Unexpectedly, all these methods underestimate Zestimate predictions error. In overall conclusion where Zestimate provided a relative accurate housing price prediction. [5]

Li et al in 2009 has used vector support regression to forecast prices of property in China using quarterly data from 1998 to 2008. During their analysis they selected five feature to predicted property prices as output variable of support vector regression SVR. Based on their conventional evaluation criteria e.g., Mean absolute error MAE, mean absolute percentage error MAPE, and root mean square error RMSE, they come to end that SVR model is best machine learning technique to predict property prices. [6]

Rafiei and Adeli in 2016 have used support vector regression to decide whether property developer should build a new development or stop construction at beginning or project based on prediction of future housing prices. In their study, they trained a model with almost 26 features e.g., ZIP code, area, cost of construction, property prices etc. Their results explain that support vector regression is an excellent technique for prediction of housing prices. Evaluation findings, provide valuable inputs to decision making of property developer. [7]

Wang and Wu min 2018 employ 27, 649 housing assessment price data from Airlington county, Virginia USA and suggest random forest outperforms linear regression in terms of accuracy. According to feature e.g., number of bedrooms, floor level, building age area etc. their study compares multiple machine learning algorithm random forest, Decision tree, Linear regression, during evaluation they decided that random forest is one of best in terms of overall accuracy as evaluated by root mean squared error. [8]

In 2019 Koktashev et al. used housing transection records off 1970 and attempt to estimate housing value in city of Krasnoyarsk. Their feature includes number of rooms, total area, floor, parking, type of repair, type of bathroom, rate of house features. They applied random forest, ridge regression, linear regression, random forest to predict property prices. They conclude that random forest outperforms other two algorithm when they evaluated by mean absolute error MAE. [9]

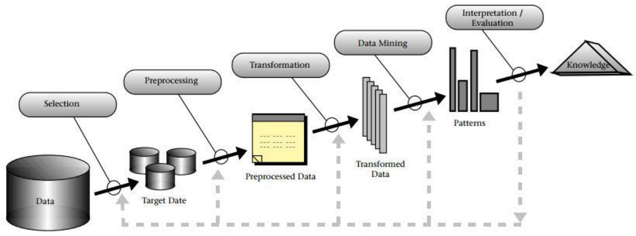
Sifei Lu, Zengxiang Li, Zheng Qin, Xulei Yang, Rick Siow Mong Goh used an advanced house predictions system using linear regression. The aims of their model were to predict a good house price predictions based on other variables. They used liner regression ML techniques because it provides them good accuracy. Their model provides predicted house price to user based on selected locations. [10]

In research paper by "Ayush Varma Abhijit Sharma Sagar Doshi Rohini Nair" recommended that use of neural networks along with linear regression and boosted algorithm can improved prediction accuracy house price. They used neural network mostly to compare all predictions and shows most accurate results. Whereas a neural network with boosted regression was to increase predictions accuracy of their outcomes. [11]

In a research paper by Dr. M. Thamarai "House price prediction modelling using machine learning" Shinde and Gawande predicted house prices using different ML algorithm e.g., Lasso regression, SVM, logistic regression and compare their accuracy. [12]

METHDOLOGY

Diagram: to show steps of the KDD process.



KDD is a framework which has its core, the application for specific data mining method for pattern discovering and method.

Data Selection

Datasets need to be existed to predict house price based on independent variables. We have collected our London house prices data from Kaggle.com and Dublin house prices data from propertypriceregister.ie. All datasets were in comma separated value (.csv) file. Our project idea is to predict house prices of Dublin and London and these datasets contains all relevant information and we can use our independent variables to predict dependent variable.

Characteristics of Datasets.

Summarize	Dublin House price	Average house prices of London	London houses
File Type	Comma separated values	Comma separated values	Comma separated values
File Order	Structured	Structured/Unstructured	Structured
Size of File	1.75MB	658KB	255KB
Number of instances	14310	13550	3479
Number of attributes	10	7	11
Types of attributes	String, character, Int	String, Date, Int, characters	String, int, Characters

Data Pre-processing

Data pre-processing is a process of transforming raw data into understandable format. In this method we can check our quality of data before applying to machine learning algorithm. All our datasets (Average London house prices, Dublin house prices, London house prices) were imported to RStudio where we begin to clean our data with help of R programming language.

First, we checked indexes of each dataset what information is stored in each index, are these columns are suitable for our further analysis if not then we deleted all unnecessary columns from our datasets with help of R programming language in RStudio. Names of columns has been changed, if they were not in proper written form of all our datasets. Converting prices of Dublin from characters to numeric values. To find null values we checked all columns are there any null values in our datasets if there are null values then we filled null values by taking mean of that column. After performing pre-processing, our datasets are ready for machine learning algorithm.

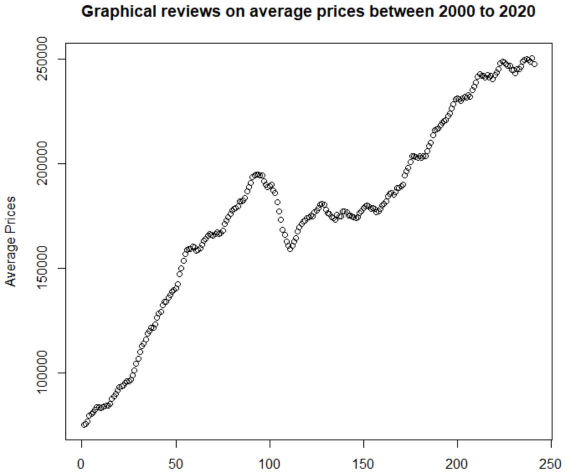
Data Transformation

Data transformation is part pre-processing, it is common approach that can perform to convert data from one type to another type in order to carried out certain type of data mining techniques. Large datasets have been transformed into smaller dataset, taking only those attributes and instances on which, we can perform machine learning algorithm. Exploratory data analysis is used to analyse and summarise main characteristic of datasets, during EDA process descriptive statistics has been performed to check what is mean, median, mode, min, max, 1st quartile, and 3rd quartile of our all datasets.

Descriptive Statistics	Average house price of London	Dublin House Prices	London House prices
Mean	263520	531195	1864173
Median	222919	375000	1220000
Min	40722	6900	180000
Max	1463378	170142820	39750000
1st quartile	132380	285463	2150000
3rd quartile	336843	538500	750000

Transformation of data is a process which mostly focuses on transformation of data from one focus to another so that data mining algorithm can be implemented easily, removing fields, adding new columns from our datasets. Whenever we see few statistical values like mean, median, mode, max, min, and standard deviation, it can help us to understand our data in much easier form.

Figure 1: It represents average prices of London houses from year 2000 to 2020.

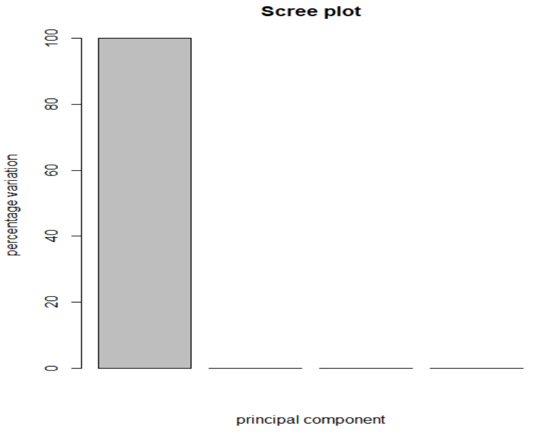


Principal component analysis is used during transformation process. Whereas PCA is a technique for reducing dimensionality of datasets, improving interpretability but at same time also minimizing information loss. We use principal component analysis techniques in RStudio on London housing dataset. Hence summary of PCA:

	PC1	**PC2**	**PC3**	**PC4**
Standard Deviation	2267283	1016	1.329	1.585e-15
Proportion of variance	1	0	0.000	0.000e+00
Cumulative Proportion	1	1	1.000	1.000e+00

From above table we can see we have four types of PCA components PC1, PC2, PC3, PC4, because our input data carried 4 featured and that’s why it calculated four components. What is happening above, in our PC1 cumulative proportion is close to hundred percent variance of data.

Figure 2: Scree plot of principal component analysis on London houses prices datasets.



Data Mining.

Data mining is our second last step in knowledge discovery and data mining. Data mining is used to turn raw data of Dublin and London houses into some useful information. Data mining and machine learning algorithm are performed in RStudio with R programming language. It is a process of analysing a large batch of information to determine trends and patterns. Some data mining algorithms that we used to predict price of houses are, time series analysis, decision tree, multiple linear regression, Lasso regression, OLS regression, simple linear regression.

Based on above ML algorithm we discover that is we want to buy a house in London with 5 bedrooms and 1500 square feet, it will cost us around 972359.3£. Whereas if we want to buy an apartment in Dublin with three bedrooms and it will cost us around 342218.1 euros.

Time series analysis.

Time series analysis is a specific way of analysing a sequence of data points collected over an interval of time. Time series helps us to understand the underlying causes of trends or systematic patterns over time.

Time series analysis is performed on average prices of London housing data, where we checked trends of how many houses has been sold between 1995 to 2020 of different locations of London city e.g., barking and Dagenham, Barnet, Bexley, city of London etc. and performed forecasting how many houses can be sold over the coming periods.

In RStudio we converted our date attribute to time series analysis put frequency is equal to 4 (data points for every quarter of the year) by using tsibble package and we get ts object with one variable and 13, 549 observations. After, using same time series data we performed decomposing analysis, seasonal analysis, corelation analysis, ARIMA forecasting, naïve forecasting, and single exponential smoothing forecasting on city of London houses from 1995 to 2020.

Decision Tree.

A decision tree is known as decision support tool that uses a tree like-model of a decision and their consequence. Decision trees are mostly used in operations research particularly in decision analysis to help identify a strategy mostly to reach a goal. Decision tree algorithm is performed on London housing data where price dependent variable and number of bedrooms were set as independent variable and performed prediction.

During implementation first create a training and test datasets for London housing prices and split it to 70:30. Then we build our model with recursive partitioning function and test performance of our model and checked accuracy which is 0.9923. After that we pruned our tree by using lowest cross validation parameter and evaluate performance of our pruned model.

Multiple linear regression.

Multiple linear regression is a technique that use several explanatory variables to predict outcome of response variable. The aim of this multiple linear regression is to model linear relationship between independent variables and response (dependent) variables. Multiple linear regression is used on London housing data to predict prices of houses based on our independent and dependent variables. During multiple linear regression model, we set prices attribute as our dependent variables and numbers of bedrooms, area per square ft were used as our independent variables. We predict prices of London houses with multiple linear regression model by setting our dependent and independent variable and get accuracy around 0.4634.

Lasso Regression.

Lasso is known regression analysis technique that perform both variable selection and regularisation in case if we want to increase our accuracy. Lasso regression is performed in RStudio on London housing dataset to predict prices of houses. Dependent and independent variables were used for prediction of model, in Lasso regression prices were set to our dependent variable and area per square ft, number of bedrooms, number of baths were used as independent variables and applied glmnet method. During model performance or accuracy, we get our R squared value around 0.4648566.

During the implementation of Lasso regression, we applied another model OLS (ordinary least square) multiple linear regression and compare accuracy of our model 2 with model 1 (Lasso regression). Like Lasso regression, we used price attribute as our dependent/outcome variable and area per square ft, number of bedrooms, number of baths were used as independent variables. Hence during OLS model performance/accuracy we get our R squared value around 0.4730667.

Simple linear regression.

Simple linear regression is used to value relationship between two quantitative variables. It can be used to check a relationship between two variables, how strong they are. In this regression analysis, there will be a dependent variable (prediction value) and other will be independent variable. Simple linear regression model is used to predict Dublin houses prices as our dependent variable and number of bedrooms were set as our independent variable. If we say we want to know price of house with 3 bedrooms then, based on our simple linear regression model (two quantitative variable), the price of house will be 342218.1.

Interpretation/Evaluation

It is our last step of our knowledge discovery and data mining. Now we will analyse all our machine learning models, which model has most accuracy, and based on accuracy should we select this model and why? As mentioned above we have performed 5 different ML models and get different R squared value of each model and based on this R square value we can evaluate and select our best model.

Evaluation

Is price of houses is increasing sharply in London and Dublin? what would be price of house if we want to buy a house in London with 5 bedrooms with 1500 square ft? what is price of a house in Dublin with 3 bedrooms? To answer all these questions, we imported all our datasets into RStudio and performed machine learning algorithm (Time series analysis, decision tree, multiple linear regression, lasso regression, OLS multiple regression, simple linear regression) with R programming language.

During time series analysis we performed some forecasting models (ARIMA forecasting, naïve model, single exponential smoothing) to analyse how many houses can be sold in centre of London in next quarter of the year.

Figure 3: ARIMA forecasting time series analysis. It represents sales of houses sold in centre of London with AIC= 4127.83, AICc=4127.96, BIC=4142.66. Based on ARIMA model we can see that there is no increment in sales of house in coming quarter of the year.

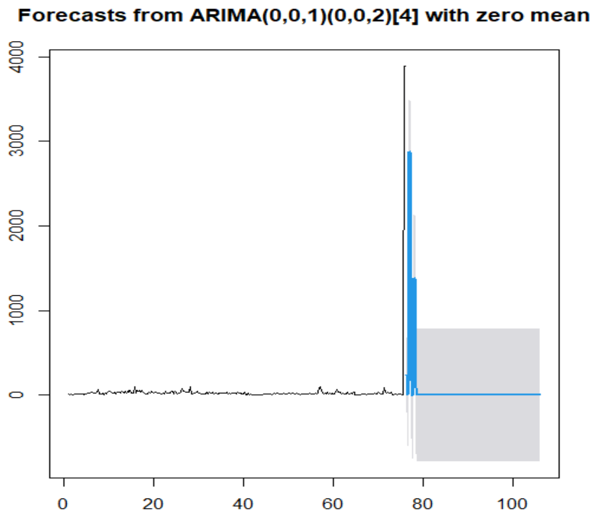


Figure 3: Naïve forecasting time series analysis. It is used to check sales of houses sold in centre of London root mean square error RMSE 225.0973, MAE

value is equal to 22.92331, but if we check our accuracy of our naïve model then our mean absolute error value is 0.5915057 which is around 60 percent of accuracy of this model. Based on our naïve model we can say that there is slightly increment in sales of London houses in coming quarter of year.

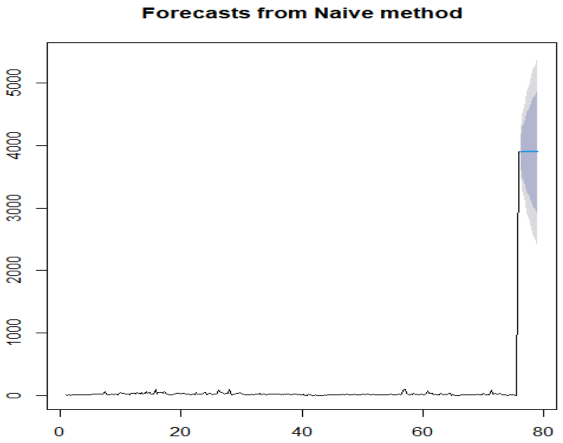
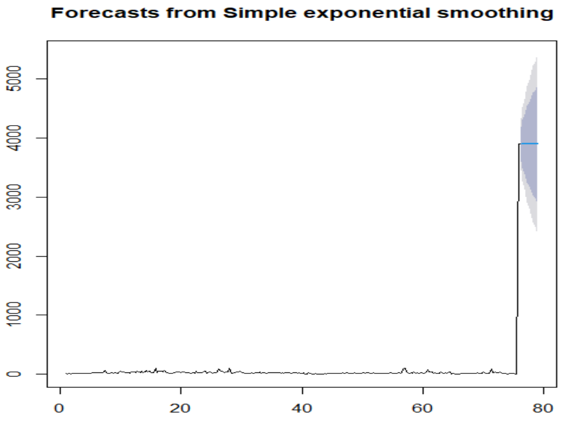


Figure 4: Single exponential smoothing of time series analysis.

SEM model is also used to check sales of houses sold in centre of London where our AIC is 4983.590, AICc is 4983.671, and BIC is 4994.712. Therefore, to check accuracy of single exponential smoothing our mean absolute scaled error is 0.5899239.



Hence based on above two models naïve and single exponential smoothing we can evaluate that our naïve model has highest accuracy than other based on MASE value which is 0.5915057.

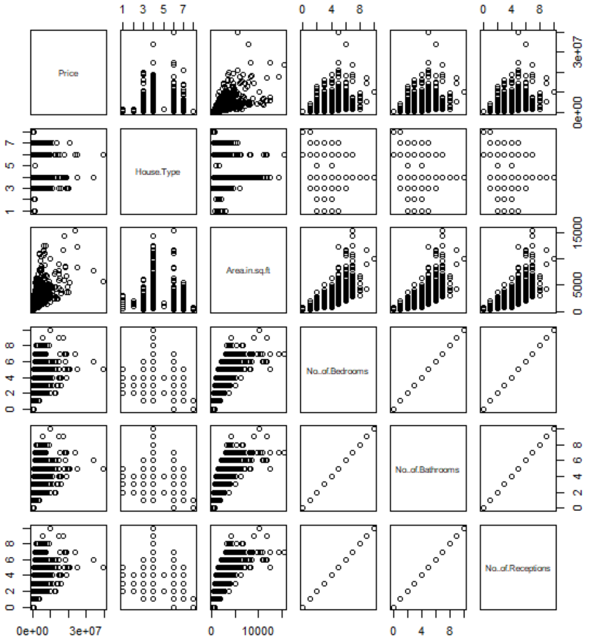
Decision tree analysis is used to predict prices of London houses. During our implementation in RStudio we set our London prices as our dependent variable and Number of bedrooms column as our independent variable. If we check our accuracy of decision model which is around 0.023 which is not high as compared to others.

Multiple linear regression predicts prices of London houses by using number of bedrooms and area per square ft as our independent variable. During evaluation of this model, we have our R squared value around 0.4634.

Corelation Coefficient of multiple linear regression.

	Area sq. Ft	No of bedrooms	No of Bathroom	No of Receptions
Area sq. Ft	1.00	0.78	0.78	0.78
No of bedrooms	0.78	1.00	1.00	1.00
No of Bathroom	0.78	1.00	1.00	1.00
No of Receptions	0.78	1.00	1.00	1.00

Figure 5: Pair plot in multiple linear regression for dependent and independent variables.



Lasso regression is used to predict prices of London houses, by setting our dependent and independent variables. During evaluation of our Lasso regression, we have our mean R squared value around 0.4729806. During implementation of this model, we performed OLS multiple regression and compared R squared (0.4730667) value with Lasso regression. Here we can evaluate that Lasso regression has slightly high accuracy than OLS regression.

Conclusions & Future Work.

Therefore, based on above analysis prices of houses in London and Dublin will increase significantly from time to time it will move up to 12 percent in 2022. From above machine learning algorithm, we have highest accuracy of Lasso regression for prediction of house than all others machine learning models. If we want to buy a house in London with 5 bedrooms and 1500 area square feet our predictions price will be 972359.3. further, if we want to see prediction price of Dublin house with three bedrooms house then our prediction price will be 342218.1. The advantage of this projects is it can help people who wanted to buy a house in one of most expensive cities of world, so they can have an idea in the future, what price should be and they can plan their financial services, save money etc.

In future work, we can use support vector machine, gradient boosting, and neural networks machine learning algorithm to increase accuracy of our model predictions. Further, we can use Power bi to create a dashboard to visualise our insights in one page. We can also analyse why house prices in London and Dublin has becoming unaffordable for low-income employee. Does prices of house will change based on locations.

REFERENCE.

[1] Thamarai, M. and Malarvizhi, S.P. (2020). House Price Prediction Modeling Using Machine Learning. International Journal of Information Engineering and Electronic Business, 12(2), pp.15–20.

[2] Burke-Kennedy, E. (n.d.). House price inflation expected to hit 12% before falling back in 2022. [online] The Irish Times. Available at: https://www.irishtimes.com/business/economy/house-price-inflation-expected-to-hit-12-before-falling-back-in-2022-1.4655269

[3] Zhang, Q. (2021). Housing Price Prediction Based on Multiple Linear Regression. Scientific Programming, 2021, pp.1–9. Available at: https://www.hindawi.com/journals/sp/2021/7678931/

[4] twin (2019). Data Mining: How Companies Use Data to Find Useful Patterns and Trends. [online] Investopedia. Available at: https://www.investopedia.com/terms/d/datamining.asp

[5] rstudio-pubs-static.s3.amazonaws.com. (n.d.). Time Series Analysis and Forecasting with the TSstudio Package. [online] Available at: https://rstudio-pubs-static.s3.amazonaws.com/464590_529f604d55674bd3a046d7e76f862a1f.html
[Accessed 24 Dec. 2021].

[6] Patil, P. (2020). International Research Journal of Engineering and Technology (IRJET). [online] Available at:
https://www.irjet.net/archives/V7/i3/IRJET-V7I31123.pdf
[Accessed 24 Dec. 2021]

[7] sarsam, S. (2018). What are the Transformation as a step of KDD can include? [online] Available at: https://www.researchgate.net/post/What_are_the_Transformation_as_a_step_of_KDD_can_include
[Accessed 24 Dec. 2021].

[8] Wikipedia Contributors (2019). Lasso (statistics). [online] Wikipedia. Available at:
https://en.wikipedia.org/wiki/Lasso_(statistics)

[9] Bevans, R. (2020). Simple Linear Regression | An Easy Introduction & Examples. [online] Scribbr. Available at:
https://www.scribbr.com/statistics/simple-linear-regression

[10] www.youtube.com. (n.d.). LASSO Regression in R (Part One). [online] Available at:
https://www.youtube.com/watch?v=5GZ5BHOugBQ&t=4108s
[Accessed 25 Dec. 2021].

www.youtube.com. (n.d.). LASSO Regression in R (Part Two). [online] Available at:
https://www.youtube.com/watch?v=NUfbl7ijZ0Q

infovis-wiki.net. (n.d.). Knowledge Discovery in Databases (KDD) - InfoVis:Wiki. [online] Available at: https://infovis-wiki.net/wiki/Knowledge_Discovery_in_Databases_(KDD)
[Accessed 25 Dec. 2021].

Property Services Regulatory Authority (2011). Residential Property Price Register - Home Page. Propertypriceregister.ie. [online] Available at: https://propertypriceregister.ie/Website/NPSRA/pprweb.nsf/page/ppr-home-en
[Accessed 16 Apr. 2020].

kaggle.com. (n.d.). Housing in London. [online] Available at:
https://www.kaggle.com/justinas/housing-in-london
[Accessed 25 Dec. 2021].