Impact of the high-frequency public transport on the performance of the Machine Learning model for predicting the rental price in Dublin

MSc Research Project
MSc in Data Analytics

Neha Kulkarni
StudentID:x22166165

School of Computing
National College of Ireland

Supervisor: Taimur Hafeez

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Neha Kulkarni
Student ID:	X22166165
Programme:	Msc in Data Analytics
Year:	2024
Module:	MSc Research Project
Supervisor:	Taimur Hafeez
Submission Due Date:	31/01/2024
Project Title:	Impact of the high-frequency public transport on the performance of the Machine Learning model for predicting the rental price in Dublin
Word Count:	6551
Page Count:	21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Neha Kulkarni
Date:	31st January 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☑
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☑
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☑

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Impact of the high-frequency public transport on the performance of the Machine Learning model for predicting the rental price in Dublin

Neha Kulkarni
X22166165

Abstract

In years the rising population, in cities has made it increasingly important to find affordable rental houses that match their true value. To address this concern our research employs a machine learning technology to predict rental house prices in the Dublin housing market. By acknowledging that multiple factors influence housing prices we adopt an approach that combines algorithms to enhance model accuracy and prediction. This study incorporates machine learning algorithms such, as Decision Tree, Gradient Boosting and K-Nearest Neighbour(KNN). This study predicts the prices of residential rental properties using the data from RTB (Residential Tenancies Board) which is open source data. The dataset has various features like location, property features etc. Additional features have also been added to the dataset by looking into the connectivity of the rental property with the high-frequency public transportation of Dublin. The high-frequency public transportation in Dublin is the luas and the darts. The luas have two different lines one is the green line luas, and another is the red line luas. We will also add more features like a number of luas and dart stops and also identified and then classify the postcodes of Dublin into North and South of Dublin. This study aims to provide accurate rental price prediction and also explain the connection between the various factors. After applying the regression machine learning algorithm, we also predicted the rental house price from the year 2023 to 2026 by applying the ARIMA Model Grid Search and Forecasting.

1 Introduction

Dublin City the capital of Ireland is an expanding and lively urban hub. This city perfectly demonstrates the complexities associated with rental housing in the present time. It is very difficult to find a suitable house with reasonable rent. As the country welcomes people from all around the world, housing crises has increased in the capital city Dublin which has led to an increase in rental price. Dublin being a dynamic metropolitan and economic hub of Ireland abbreviates the opportunities and challenges in the rental housing market. In this situation, understanding and predicting the rental house price in Dublin is essential to stimulate sustainable living and it will also facilitate the tenants and landlords for making informed decisions. This will overall assist in guiding and following effective housing policies. This research seeks to contribute to this essential by applying advanced ensemble machine learning methods to predict rental house prices in the Dublin house market. This research will help multiple stakeholders like tenants, landlords, property agencies and students.

[1]

[2]This research aims to investigate the rental housing price in Dublin by using data published by the Residential Tenancies Board (RTB).

This data is controlled by the Central Statistics Office(CSO) which is Ireland's national statistical institute. This research is focused on the data for the year 2022 for Dublin City. The dataset has multiple features like the year, number of bedrooms, property type, and location. Figure 1, shows the high-frequency public transportation in Dublin are the luas and the darts. The luas have two different lines which are the green line luas, and another is red line luas. The study focuses on various areas of Dublin with high frequency transportation which are connected with luas and dart.

Hart, R., Singleton, C., Figure 1. Dublin Luas and Dart stop Map. [Online Image] [Accessed on 14th Dec 2023]
https://www.dublinpublictransport.ie/dublin-train-map

The proposed research tries to investigate a solution to answer the following research question.

- **Research Question -** What is the most effective machine learning method for predicting house rental prices in Dublin using a variety of influencing factors?

[1] https://www.rtb.ie/

[2] https://www.dublinpublictransport.ie/

[3] https://www.cso.ie/en/index.html

- **Sub-Research Question** – What would be the trend of housing rent for the next four years from 2023 to 2026 in Dublin?

The main objectives for this research are –

- Study of the contemporary work – The exploration in recent work to predict the house price would contribute to the knowledge of ensemble methods for rental house price prediction.
- Comprehensive Study for additional features – To gain a complete understanding of the multiple factors which influence the rental housing price in the Dublin market.
- To find out the impact of high-frequency transport (luas and dart) connectivity on the rental house price.
- Research Method – Research methods focus on various steps starting from data collection to building a model. The steps are data collection, data pre-processing, data cleaning, and splitting the data. The data is prepared to be used by the model.
- Model Building –We will split the dataset into a test and the train data for the model. This step involves the actual implementation of the regression machine-learning algorithm. After building the model we will also add more additional features like a number of luas and dart stops and also identify and then classify the postcodes of Dublin into North and South of Dublin.
- Evaluation of Model – After adding the additional features we re-ran and model and got improved performance and accuracy of the model. This is the final stage of the research where we evaluate the model based on parameters like R-square, RMSE and MAE.

Predicting the rental house price was a novelty of this research which also includes adding influencing features to the dataset like high-frequency transport(luas and dart) connectivity and finding which postcodes of Dublin are in North and South of Dublin. This study gave a detailed idea about various areas of Dublin and public transportation. The luas are divided into two different lines which are red line luas and green line luas.

The primary contributions to the research are as follows:

- This research successfully identified the attributes of the Dublin rental house price market by examining the relation between each attribute and its rental price.
- The research discusses the challenges of fewer existing features in the dataset and how it was addressed by adding additional features of high-frequency transport connectivity and north/south geographical classification of the Dublin postcode.
- Lastly, in this research, we compared the performance of the best model and also implemented a time series ARIMA model to predict the rental house price from the years 2023 to 2026 for different types of houses and the number of bedrooms.

The rest of the document is structured into sections as outlined below. The subsequent section outlines the literature review which focuses on the background and idea of the research done. Section 3 describes the methodology for model building which is followed by section 4 defining the designing tools. Section 5 describes the steps which were followed during the implementation. Section 6 highlights the comparison of the implemented models and outlines the best technique followed by the application of the time-series model using ARIMA to predict the rental house price from the year 2023 to 2026. Section 7 contains the conclusion and future work.

2 Related Work

In this section a crucial part is covered which has the most recent results for the problem statement. In the past few years there are many research that has been lead for various parts of the world on house price prediction. Rising house price is the most crucial challenge that different cities are facing, be it rental or sale. Irish economy has highlighted many factors that contributed to the economy and showcased the shortage. There were multiple research and survey that has been conducted for house price evaluation.

2.1 Factors Affecting House Price Dynamics :

In this study, the researcher Maurice (2001) talks about a noticeable increase in the Dublin house price. He mentioned that this increase in house prices is because of the increasing demand. Here the author did not use any machine learning technique to predict the house price however a statistical approach is being used. The author divided the house price into fundamental and non-fundamental components considering different measures. In this research, a regime-switching model is evaluated and tested to check if the speculative bubbles, fads, or fundamentals impact Dublin's real estate value. A-D method and the values like probability and coefficient are calculated to get the findings.

Research by Hurley, A.K., Sweeney, J. (2022) addresses the importance of property valuation especially for the Irish real estate market. There is a unique feature in the dataset which is descriptive values. The variable has valuable information which includes the descriptive features of the property and its surroundings locations. The spatial hedonic regression method performed better as compared to other flexible approaches using GAMs methods while considering factors like location. The research suggests extending the model to the Republic of Ireland which would result in improved computation of property tax by using valuation technique. The non-linear association between the price columns and other various features is considered. This also allows to include the Gaussian Process smooth method to estimate the values for location. The hedonic model shows a reduction in median absolute percentage error and increasing model complexity which is 12.10%.

2.2 Analysis of House Price Prediction in Different Geographic Regions :

Truong, Q., Nguyen, M., Dang, H., Mei, B. (2020) explored house price prediction using machine learning techniques. They also highlighted the importance of features in predicting the house price. They compared machine learning methods which are Random forest, XGBoost and LightGBM and two different techniques Hybrid regression and Stacked Generalized Regression. The authors state that each model has different weaknesses and strengths XGBoost and LightGBM have good accuracy and lower time complexity. Due to generalization hybrid regression model has performed well and on the other side sacked generalization regression needs more time to execute however excels in accuracy.

In this study by Phan, T.D. (2018), machine learning techniques are used to predict the house price in Melbourne city of Australia by using historic property transactions. The study reveals remarkable price variation between expensive and affordable areas of Melbourne. This study also follows the CRISP-DM standard for data mining as they have historical data. This study is conducted by applying a combination of stepwise and Support Vector Machines (SVM) and it shows dynamic output. This study proposes the applicability across different cities of Australia using transactional data.

The researchers, Park, B., Bae, J.K. (2015) studied house price prediction using the housing data of 5359 townhouses in Fairfax County in Virginia. This data was collected by the Multiple Listing Services. The house price prediction model was developed by applying multiple machine learning algorithms such as C4.5, RIPPER, Naïve Bayesian, and AdaBoost were applied and their accuracy performance was compared. The researchers explain that the RIPPER algorithm outperformed the other models based on accuracy.
Oshodi et al.(2019) conducted research using a neural network model to estimate the residential property rental prices in Cape Town, South Africa. There were a total of 14 attributes in the dataset. The deep analysis shows that features like balcony and floor area had a significant impact on the rent price of the property. The least influential factors were parking space and a swimming pool. Other features like the garden and proximity to the police station had less influence as compared to the balcony. The model accuracy was 78.95%. The researcher mentioned that the results had limitations as the dataset was very less and focused on that a good amount of data is needed to get accurate results and less data is a challenge in property valuation.

2.3 Innovative Methodologies in Housing Economics Research :

In this study conducted by researcher Lorenz(2015) focused on analysing house prices in Berlin. What made this study unique, in the field of housing economics was the use of decomposition and estimation techniques. One interesting finding was that when looking at price points there were variations in coefficients providing insights into average regression boundaries. The study emphasized the significance of changes in coefficients than property features when explaining variations in prices. It also recommended investigation into understanding the contributions of different factors and how demand impacts rental rates.

In a study conducted by Li (2021) the aim was to predict the House Price Index using two machine learning models and a neural network. The data used for the study was obtained from Kaggle. Included features such, as year, frequency HPI flavour, HPI type. The models that were developed for this purpose were Lasso Regression, Ridge Regression and BP Neural Network. These models produced results, with BP Neural Network performing better. However it should be noted that despite these techniques being tested they were not able to predict the prices according to the authors findings.

Gupta et al. (2015) conducted a study to investigate the relationship between home prices in the Eurozone. In their approach they employed co integration and integration methods. The study revealed that several log real price indices exhibited integration orders than one indicating long-term growth rates. Additionally, it was discovered that France, Germany and Belgium are integrated members of the Eurozone. Overall findings from the analysis emphasize that variables like stability, income level and heritage significantly influence the property market, for buyers.

2.4 Summary of the Research Studied

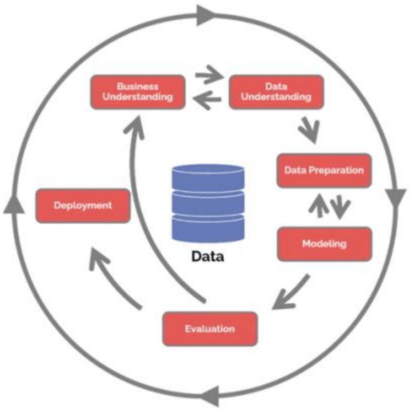
Table 1 shows all the important articles the work done and the methods they have used for similar work.

Table 1: Summary of Most Related Work

Year	Title	Method Used
2001	The Rise in House Prices in Dublin: Bubble, Fad or Just Fundamentals. Economic Modelling.	A-D method and the values like probability and coefficient are calculated to get the findings.
2022	Irish Property Price Estimation Using A Flexible Geo-spatial Smoothing Approach: What is the Impact of an Address?.	spatial hedonic regression method
2020	Housing Price Prediction via Improved Machine Learning Techniques'	XGBoost , LightGBM, hybrid regression
2018	Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia	SVM(Support Vector Machine) CRISP-DM framework followed
2015	Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data	C4.5, RIPPER, Naïve Bayesian, and AdaBoost

3 Methodology

There are two research methodologies which are regularly used one is the Knowledge Discovery Database (KDD) and the other is Cross Industry Standard Process for data mining (CRISP-DM). CRISP-DM allow more flexibility as its steps are reversible as shown in fig.2.



Hotz, N., Figure 2 CRISP-DM Lifecycle. [Online Image]
[Accessed on 14th Dec 2023]
https://www.datascience-pm.com/crisp-dm-2/

The methodology section delves into the strategies employed to accomplish the research objectives outlined in the study aim.[1]. The proposed research approach is not compatible with a one-size-fits-all approach to data analysis; rather, it is structured around the steps needed to identify the best method for rental residential housing forecast for Dublin. The steps that need to be followed in order to get the desired outcome are highlighted in Fig.3.

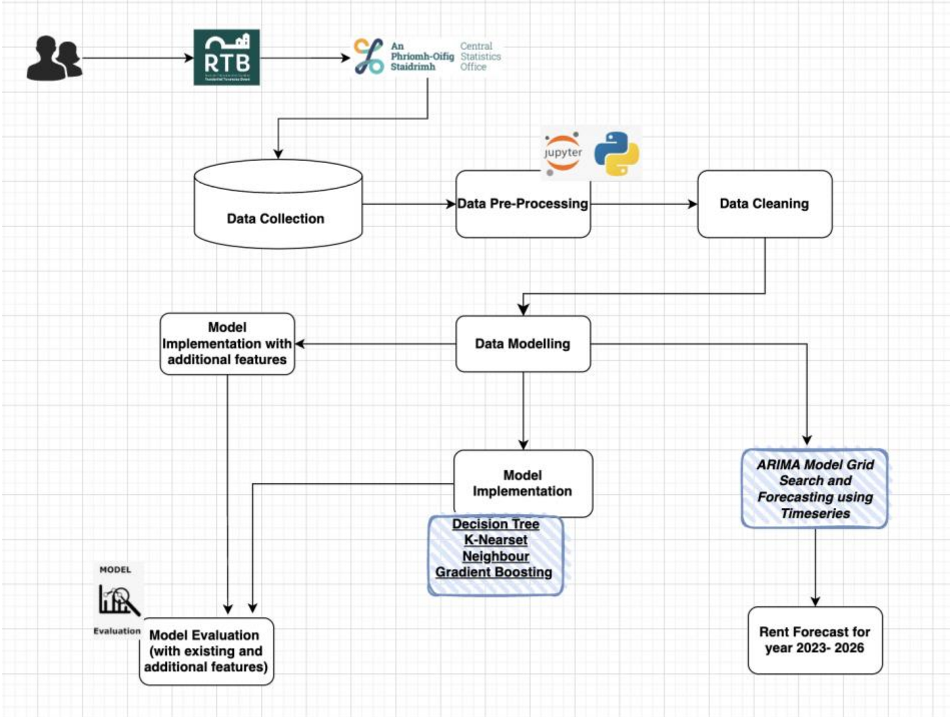


Fig. 3: Research Methodology for Rental House Price Prediction Dublin

3.1 Data Collection

Data Collection is the first and most important part of the process. Data used for this research is used from the CSO website (Central Statistics Office). This dataset is an open-source data which is maintained by the government of Ireland. The data is first fed into the RTB (Residential Tenancy Board) by the people of Ireland. Then RTB transfers this data to CSO which hosts and maintains the dataset on their website. RTB is an important regulatory agency called the Residential Tenancies Board (RTB) that supervise tenant-landlord relationship. In Ireland, we have two agencies, namely the Central

Statistics Office (CSO) and the Residential Tenancies Board (RTB) each with its own unique function. However, they are closely connected through their involvement in housing-related statistics and data. In addition to its responsibility of managing the tenancy registration system and resolving conflicts between tenants and landlords, the RTB also collects valuable information about rental properties, tenancies and rental prices.

3.2 Data Pre-processing

Data preparation is a step, in the data analysis and machine learning pipeline. It involves cleaning and converting the data to ensure its accuracy, quality and suitability for analysis. The success of data analysis and the performance of machine learning models depend greatly on proper data pre-processing. We take measures to ensure that the data is clean and doesn't contain any undesirable values. In Section 5 we delve into a technique for pre-processing the data.

3.3 Data Cleaning

Data cleaning, also known as cleansing or scrubbing involves identifying and rectifying errors, discrepancies and inaccurate information in databases. By improving the quality of the data through cleaning processes we make it more accurate, reliable and ready for machine learning purposes. When cleaning the data we address issues such, as values duplicate entries, redundant information, inconsistent datasets and standardizing the format of the data.

3.4 Data Modelling

Data modelling is a part of designing a database because it entails constructing a representation of how the database or system is organized and functions. By employing data modelling techniques we can improve the efficiency, consistency and integrity of data storage and retrieval processes.
This phase of the project provides information, on the methods used to create the model. It outlines the attributes that were transformed into form and the subset used for training purposes. Since many features are categorical in nature and binary forms have limitations, data modelling in this study poses challenges.

3.5 Model Building

Model building refers to the process of developing, training and refining a model. This process involves training algorithms to identify patterns, within input data and using that knowledge to make predictions or decisions. Constructing models is a part of any machine learning process.
In research related to home prices scholars have primarily focused on learning, linear regression, KNN (K Nearest Neighbours) and Random Forest techniques. As a result, this research will use Decision Tree, Gradient Boosting Algorithm and KNN as the algorithms. Moreover, we have also employed the ARIMA Model Grid Search and Forecasting technique along, with the regression approach to estimate the prices of Dublin dwellings for the years 2023 to 2026.

3.6 Model Evaluation

When comparing regression models it is crucial to examine how well each model aligns with the given data. As a step, in the recommended approach model assessment involves evaluating the models using performance metrics. Since regression analysis plays a role in this scenario we would employ the following metrics, for evaluation; R Square, RMSE and MAE. When comparing these models we would assess their R Square, RMSE and MAE metrics. Additionally, the code component includes prediction charts corresponding to each of the evaluations presented in Section 6.

4 Design Specification

In this research Figure 4 showcases the Three Tier Architecture that was utilized. The foundation of the structure is the Data Layer, for collecting data points. The information for this layer is sourced from Rental house price data published by the Central Statistics Office Ireland portal in.xlsx format. Above the data layer lies the business logic layer where different machine learning models are applied. For example, regression models such as Decision Tree, KNN and Gradient Boosting are used to compare house price predictions. Additionally, a Dublin rental forecast for the years 2023 to 2026 is conducted using grid search ARIMA. It's worth noting that Mean Absolute Error (MAE) Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) values are stored for each model type. These values play a role, in determining the machine learning model. The final outcome of combining these two layers is stored in the Data visualization layer, where PowerBI extracts and visualizes the data stored in Excel.

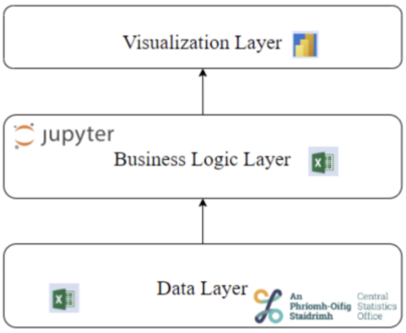


Figure 4. Three Tier Architecture

5 Implementation

This stage of the project explains how the data is transformed for the model implementations. This process explains every step in detail from data loading to model building. We have implemented regression model which are K-Nearest Neighbour (KNN), Decision Tree and Gradient Boosting for the rental house price prediction. Also, we have implemented ARIMA Model Grid Search and Forecasting using time series for forecasting Dublin rent for the year 2023 to 2026. All the steps covered in the implementation process is highlighted in Fig.5. All the steps involved are as follows

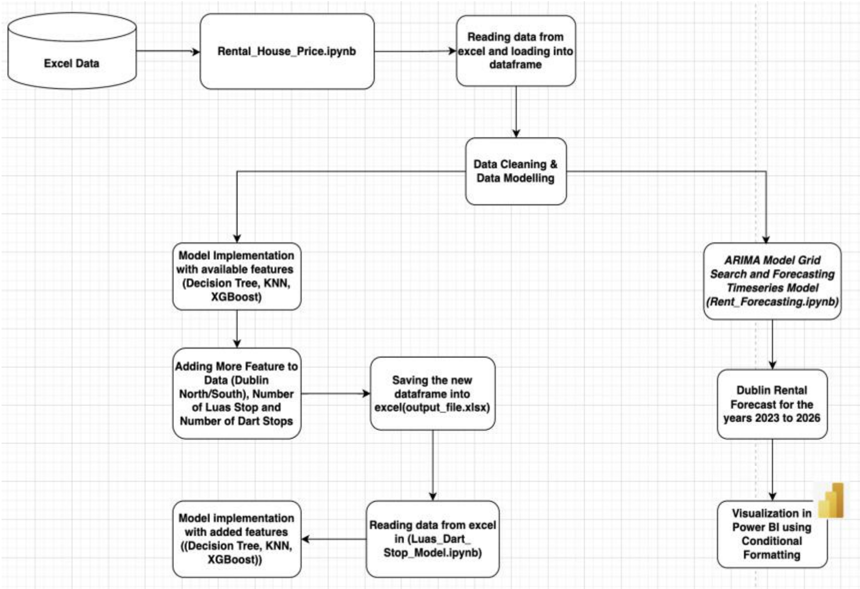


Fig.5 Implementation Flow Chart

5.1 Data Collection

Here in this stage, we collect the data which we will use in the project. This is the first and most important step to start with the implementation. For this project, we have used the open-source data available on the CSO website (Central Statistics Office) which is maintained by the government. The data is first uploaded to the RTB portal by the landlord or the agents. As an agency the CSO compiles and publishes housing market statistics using data provided by the RTB.
This site has historical data for various counties of Ireland. The data is downloaded from the CSO website in an excel format which has 280980 rows. This data is read by the python code in a data frame to perform the data pre-processing and data cleaning.

5.2 Data Pre-Processing

The project encountered a challenge because of the amount of data that needed to be processed and prepared. After importing the data into Python we created a dataframe to arrange it. We utilized pandas to analyse the values. We paid attention to preparing all the attributes in the dataset, which would aid in research. Tasks such, as modifying attribute values and adjusting data types were crucial, during the data preparation stage. We diligently documented a summary of the procedures implemented for each attribute.

- **Deriving New Columns**

The first step, in pre-processing a dataset is to apply filters based on the year. The number of bedrooms. In this study we have specifically chosen data from the year '2022'. After applying these filters we are left with a total of 10,704 rows of data. As part of preparing the data we have. Replaced values in the 'Location' column. When filtering the "Dublin" dataset based on location certain variables were excluded. Through searching of Dublin Postcodes we have created a mapping. Additionally new features like "postcode Dublin" have been generated using existing data. To further narrow down the dataset and focus on Dublin postcodes relevant, to our research scope another round of filtering has been applied. You can refer to Figure 6 for a view of the processed dataset.

```
print("Info of DS1:", filtered_DS1.info())
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 574 entries, 264996 to 279784
Data columns (total 8 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   STATISTIC Label     574 non-null    object
 1   Year                574 non-null    int64
 2   Number of Bedrooms  574 non-null    object
 3   Property Type       574 non-null    object
 4   Location            574 non-null    object
 5   UNIT                574 non-null    object
 6   VALUE               574 non-null    float64
 7   postcode_dublin     574 non-null    object
```

Fig.6 Pre-Processed Dataset

- ## Adding New Features to the Dataset

  We have added new features to the dataset by doing additional research from Transport for Ireland(TFI) website and Dublin Public transport website. The two new features that we added to the dataset are Luas and Dart. We achieved this by doing reverse engineering by first taking all the Dublin station in an excel and then later adding the number of Luas and Dart station for it. Fig. 7 shows the analysis and steps involved for adding the new feature.

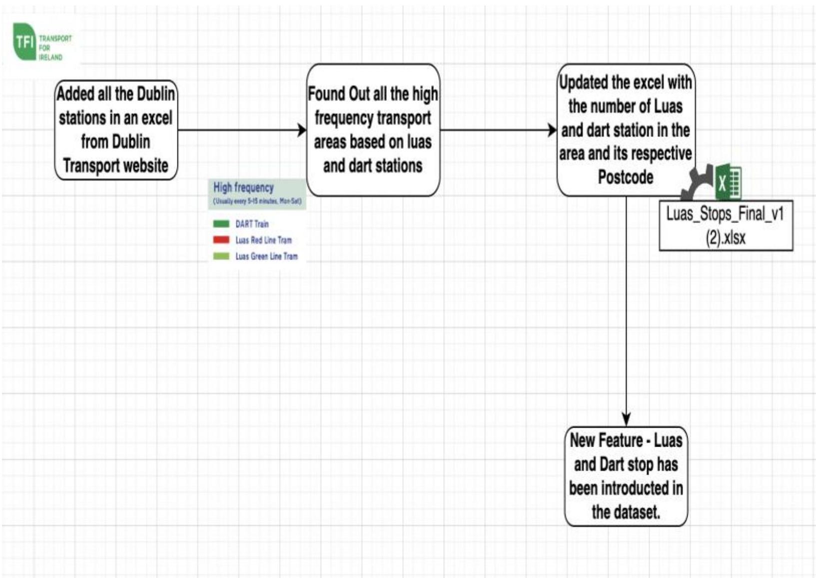


Fig.7 Steps of adding new feature in the dataset

The below Fig.8 shows the two new features that have been added to the dataframe.

	Postcode	Luas	Dart
0	Dublin 01	12	1
1	Dublin 02	8	3
2	Dublin 03	0	1
3	Dublin 04	0	3
4	Dublin 05	0	1
5	Dublin 06	4	0
6	Dublin 07	9	0
7	Dublin 08	6	0
8	Dublin 12	3	0
9	Dublin 13	0	6
10	Dublin 14	3	0
11	Dublin 15	0	0
12	Dublin 16	1	0
13	Dublin 18	11	1
14	Dublin 22	1	0
15	Dublin 24	10	0
16	Dublin A94	0	4
17	Dublin A96	0	5
18	Dublin A98	0	1
19	Dublin K36	0	1
20	Dublin K45	0	0

Fig.8 Additional New Features in the dataset

We have utilized the data from Excel Luas_Stops_Final_V1(2).xlsx and created the visualization for Luas and Dart stops in Dublin using Power BI. Below Figure 9 and Fig.10 shows the bar graph representation.

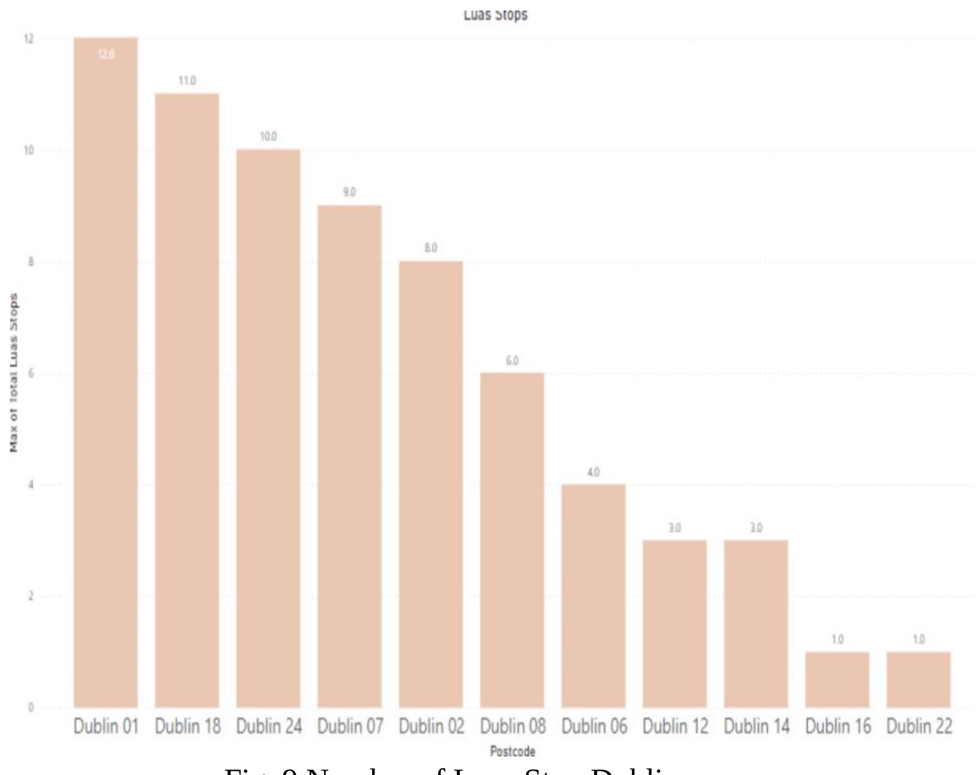


Fig. 9 Number of Luas Stop Dublin

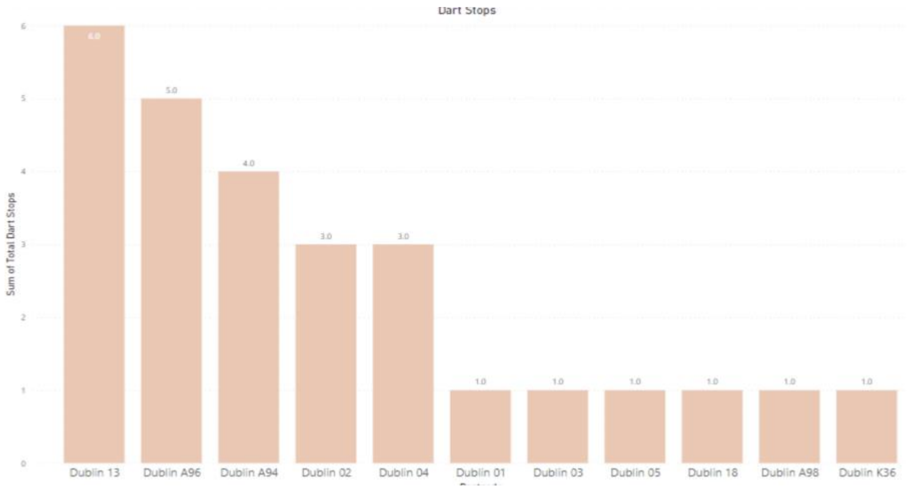


Fig.10 Number of Dart stops in Dublin

- **Classifying Dublin North & South**
  We have created one more additional column in the dataset based on the postal code. We have classified the North/South of Dublin. Fig. shows the postcode which are in North of Dublin and all the other postcodes apart from that are marked as South Dublin.

Postcode	North/South Classification
Dublin 1	North
Dublin 3	North
Dublin 5	North
Dublin 7	North
Dublin 9	North
Dublin 11	North
Dublin 13	North
Dublin 15	North
Dublin 17	North
Dublin K32	North
Dublin K36	North
Dublin K45	North

Table 2. Classification of North/South Dublin

The below fig.11 shows all the columns in the data frame with the new additional columns. Originally there was 7 columns in the dataset and we added 3 more feature to the dataset and total columns are 10 in the dataset.

```
merged_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 574 entries, 0 to 573
Data columns (total 11 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   STATISTIC Label      574 non-null    object
 1   Year                 574 non-null    int64
 2   Number of Bedrooms   574 non-null    object
 3   Property Type        574 non-null    object
 4   Location             574 non-null    object
 5   UNIT                 574 non-null    object
 6   VALUE                574 non-null    float64
 7   postcode_dublin      574 non-null    object
 8   North_South_Dublin   574 non-null    int64
 9   Luas                 267 non-null    float64
 10  Dart                 267 non-null    float64
```

Fig.11 New Features in the data

- **Data Scaling**
  Min-Max Scaling is also applied on a few of the columns to normalize the data. We have also used one-hot encoder on the feature columns to convert in from categorical data to binary matrix.

5.3 Data Cleaning

By improving the quality of the data through cleaning processes we make it more accurate, reliable and ready for machine learning purposes. Here the data cleaning included replacing values and creating a mapping for the location columns. Rows which has missing values in the 'VALUE' columns which show the rental price have been removed from the data. By performing a quick check if there is any Null it has been removed from the data. The column Location has missing postcode in it which is taken care of by creating a mapping and updating the postal code through python code.

5.4 Splitting Train and Test Data

For any machine learning model, the data needs to be split into test and train. We have split the train and test data to the ratio of 80:20 meaning 80% is training data and the rest 20% can be used for testing. The training set will be used to train the machine learning model that we have implemented. We have implemented the Decision Tree Regressor, K-Nearest Neighbour (KNN) and Gradient Boost regressor, all three models are trained on the training data which leads to better accuracy. Splitting the data is also a crucial step as the model's performance also depends on it. And it helps to train the model better.

5.5 Model Implementation

We have implemented three regression algorithms which are Decision Tree, K-Nearest Neighbour (KNN) and Gradient Boost regressor based on the literature review. We have used Python language in the Jupyter Notebook framework for all the data pre-processing, data cleaning and model implementations. The complete coding is done in Python using the Jupyter Notebook framework.

- **Model Implementation with Existing Features**

  Here we have implemented all three models with the existing features in the dataset

  1. **Decision Tree Regressor**

Decision tree is a widely used machine algorithm for regression tasks. Also, it can be used for classification problems. The DecisionTreeRegressor class, from the scikit learn library is utilized to choose and train a model for Decision Tree Regression. To manage the complexity of the tree and to make sure that its ability to adapt to data a maximum depth of 3 is set (max_depth=3). To make sure the reproducibility of results the random_state=0 parameter is used. By employing the method the model is trained using the provided training data (X_train for features and y_train for the target variable) which enables it to learn patterns within the data. Later predictions (y_pred) generated by the model are applied to the test set (X_test) after training.

  2. **K-Nearest Neighbour**

K-Nearest Neighbour is a widely used machine algorithm for regression tasks and also for classification problems. Here we have implemented the K-Nearest Neighbour model using KNeighborsRegressor class from scikit-learn library. Here the regression number is set to 5, so during prediction the number of neighbour data point is 5. The model is trained on (x_train and y_train) by using fit method.

  3. **Gradient Boosting Regressor**

Gradient boosting is widely used in Regression tasks. Here we have implemented the gradient boosting model using the XGBRegressor. The model is created using specific parameters where n_estimators=100, learning_rate=0.1, max_depth=3. This model perform better as compared to the other machine learning models.

- **Model Implementation with Additional Features**

  Here we have implemented all three models with the additional features in the dataset like Luas, Dart, and Postcode. We implemented the same machine learning model to identify which type of implementation performs better i.e., with the additional feature or the existing features. The three model that we implemented are listed below

  1. **Decision Tree Regressor**
  2. **K-Nearest Neighbour**
  3. **Gradient Boosting Regressor**

- **ARIMA Model Grid Search and Forecasting**

Here we have applied the ARIMA (Autoregressive Integrated Moving Average) which is a time-series model. We are implementing ARIMA using grid search which defines a function grid_search_ARIMA and perform a grid search over the (p, d, q) parameters and to find the best set based on (Akaike Information Criterion)AIC. We have used this for forecasting Dublin rental house prices for four years which is from 2023 to 2026. We have used the same dataset extracted from the CSO website and considered the historical data for all the available years in this research. Based on the available data the model forecasted the rental house price in Dublin from the year 2023 to 2026. To perform the grid search we have used these ranges p_values = range(0, 3), d_values = range(0, 2), q_values = range(0, 3) for grid search. We have done the forecasting for three different types of property for the entire Dublin location. The numbers of bedrooms that we considered in this research are 'One Bed', 'Two Bed' and 'Three Bed' and forecasted their rental house price for the duration of four years i.e.2023 to 2026. The forecasted rent visualization graphs are created using PowerBI.

6 Evaluation

This section provides a comprehensive analysis of the three implemented models: Decision Tree, K-Nearest Neighbour, and Gradient Boosting. The model's evaluation metrics depend on these three aspects MAE(Mean Absolut Error), RMSE(Root Mean Squared Error), and R-2 Square. The lower the MAE greater the accuracy.

6.1 Decision Tree Model (With Existing Features)

The first model that we implemented is the Decision Tree where we used the existing features in the d ata to build and train the model.
Mean Absolute Error (MAE): 243.27 , Mean Squared Error: 114375.54 , R-squared: 0.65

6.2 K-Nearest Neighbour (With Existing Features)

The second model that we implemented is the K-Nearest Neighbour(KNN) where we used the existin g features in the data to build and train the model.
Mean Absolute Error (MAE): 205.23 , Mean Squared Error (MSE): 90391.71 , R-squared (R^2): 0.73

6.3 Gradient Boosting (With Existing Features)

The third model that we implemented is the Gradient Boosting where we used the existing features in the data to build and train the model.
Mean Absolute Error (MAE): 156.22 , Mean Squared Error (MSE): 53077.48 , R-squared (R^2): 0.84

6.4 Decision Tree Model (With Addition Features)

Here we applied the Decision Tree model again, however, with the new additional features to check and understand if adding new features to the data results in better performance.
Mean Absolute Error (MAE): 248.28 , Mean Squared Error: 123987.72 , R-squared: 0.63

6.5 K-Nearest Neighbour (With Additional Features)

Here we applied the K-Nearest Neighbour model again, however, with the new additional features to check and understand if adding new features to the data results in better performance.
Mean Absolute Error (MAE): 250.98 , Mean Squared Error (MSE): 126372.53 , R-squared (R^2): 0.62

6.6 Gradient Boosting (With Additional Features)

Here we applied the Gradient Boosting model again, however, with the new additional features to check and understand if adding new features to the data results in better performance.
Mean Absolute Error (MAE): 144.90 , Mean Squared Error (MSE): 43550.23 , R-squared (R^2): 0.87

6.7 Regression Model Result Interpretation

The below table shows the performance and accuracy of all the models for Dublin Rental house price prediction – with existing features and with additional features of Luas, Dart and Postcode. Overall Gr adient Boosting has performed better in both - with existing features and additional features data. The output shows that the Gradient Boosting with additional Features is the best-performing model. This model has the lowest MAE (Mean Absolute Error) and MSE (Mean Square Error) and the highest R^2 of 0.87%. This model is outperforming others. The additional feature has positively contributed to the model's prediction capability. This answers the research question that Gradient Boosting is the best machine learning to predict Dublin rental house prices.

Model	MAE	MSE	R^2
Decision Tree (Existing Features)	243.27	114375.54	0.65
KNN (Existing Features)	205.23	90391.71	0.73
Gradient Boosting (Existing Features)	156.22	53077.48	0.84
Decision Tree (Additional Features)	248.28	123987.72	0.63
KNN (Additional Features)	250.98	126372.53	0.62
Gradient Boosting (Additional Features)	144.90	43550.23	0.87

Table 3. Model Output

6.7 Dublin Rent Forecasting Visualisation

The visuals are created in Power BI and conditional formatting is used on all the visuals to show the difference in the forecasted value and present & historical value. The forecasted values are in a light colour.
Below fig.12 is the Forecasted rent visualisation in Dublin for One Bedroom from the year 2023 to 2026.

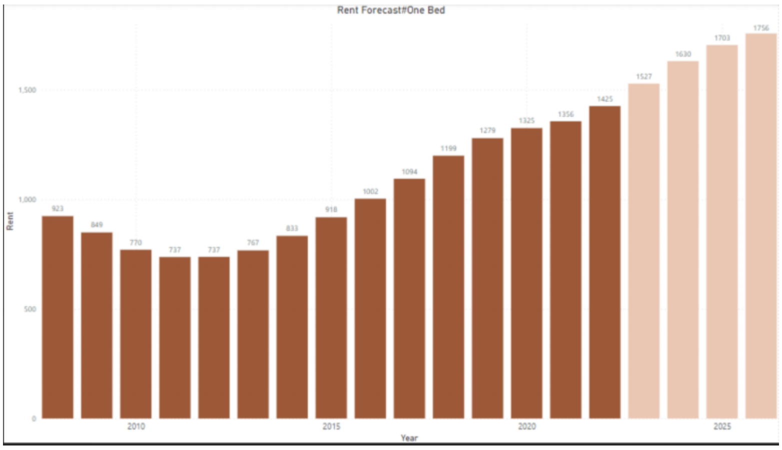


Fig.12 Dublin One Bedroom Rent Forecast from 2023-2026

Below fig.13 is the Forecasted rent visualisation in Dublin for Two Bedrooms from the year 2023 to 2026.

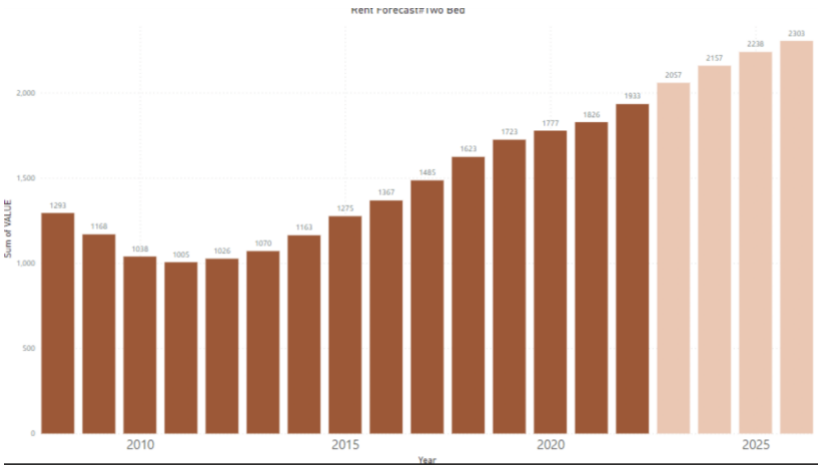


Fig.13 Dublin Two Bedroom Rent Forecast from 2023-2026

Below fig.14 is the Forecasted rent visualisation in Dublin for Three Bedrooms from the year 2023 to 2026.

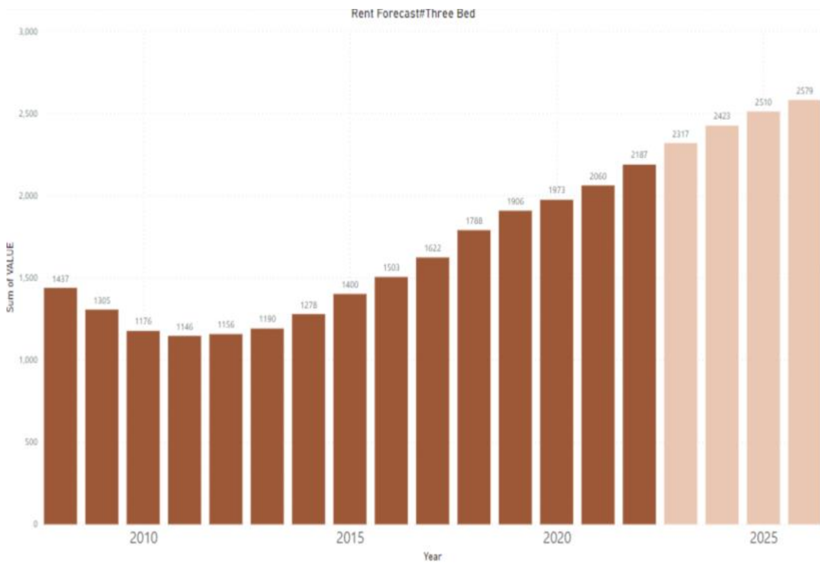


Fig.14 Dublin Three Bedroom Rent Forecast from 2023-2026

All three visualisation shows the trend of increase in rent for all the bedroom types. This answers the sub-research question

6.8 Discussion

The main objective of this research is to determine the approach, for estimating the values of rental properties in Dublin. After analysis, it has been concluded that Gradient Boost is the suitable method. The research primarily concentrated on transportation with high frequency including red line Luas, green line Luas and Dart. A significant amount of effort was dedicated to gathering information about Luas and Dart stops along with their postal codes. To simplify this process we adopted a reverse engineering approach by first identifying the Luas and Dart stops and then determining their respective postal codes. This approach proved efficient compared to the method which would have been challenging and time-consuming. During the Exploratory Data Analysis phase, several additional elements were introduced in this research such as Luas stops, Dart Stops and Postal Codes which improved performance by addressing existing limitations.

7 Conclusion and Future Work

The main objective of this research is to determine the approach, for estimating the values of rental properties in Dublin. After analysis, it has been concluded that Gradient Boost is the suitable method. This study is not focused only on creating a model however it also involves enhancing the dataset and exploring the dynamics of the Dublin Rental Housing Market. The research primarily concentrated on transportation with high frequency including red line Luas, green line Luas and Dart. A significant amount of effort was dedicated to gathering information about Luas and Dart stops along with their postal codes. One notable aspect of this study is its enhancement in performance by addressing existing limitations. The final outcome reveals that the Gradient Boost Regressor is indeed the strategy for predicting changes in datasets, with a R Square value of 0.87. In conclusion, this research not only provides a predictive model, for estimating rental property values but also enhances the methods used to improve datasets. By focusing on transportation analysis streamlining data collection and refining the model this study becomes a tool, for understanding and predicting the dynamics of Dublin's Rental Housing Market.

To improve the work further one could think about integrating a model that incorporates the method. This would benefit research efforts and will also provide tenants with a dependable platform to verify the price of the property before renting.

This work can be extended by studying other public transportation in Dublin and incorporating those details in the rental house price prediction model by various machine learning techniques. Further research can be conducted by studying medium-frequency public transportation and bus connectivity in Dublin. This research can also be carried forward for different countries of Ireland by considering various other impacting factors.

References

Roche, Maurice. (2001). The Rise in House Prices in Dublin: Bubble, Fad or Just Fundamentals. Economic Modelling. 18. 281-295. 10.1016/S0264-9993(00)00040-7.

Hurley, A.K., Sweeney, J. Irish Property Price Estimation Using A Flexible Geo-spatial Smoothing Approach: What is the Impact of an Address?. J Real Estate Finan Econ (2022). https://doi.org/10.1007/s11146-022-09888-y

Truong, Q., Nguyen, M., Dang, H., Mei, B. (2020) 'Housing Price Prediction via Improved Machine Learning Techniques', Procedia Computer Science, 174, pp. 433-442. ISSN 1877-0509.

Phan, T.D. (2018) 'Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia,' 2018 International Conference on Machine Learning and Data Engineering (iCMLDE), Sydney, NSW, Australia, pp. 35-42. doi: 10.1109/iCMLDE.2018.00017

Park, B., Bae, J.K., 2015. Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. Expert Systems with Applications, 42(6), pp. 2928-2934. ISSN 0957-4174

Oshodi, O.S., Thwala, W.D., Odubiyi, T.B., Abidoye, R.B. and Aigbavboa, C.O. (2019), "Using neural network model to estimate the rental price of residential properties", Journal of Financial Management of Property and Construction, Vol. 24 No. 2, pp. 217-230. https://doi.org/10.1108/JFMPC-06-2019-0047

Thomschke, L. (2015) 'Changes in the distribution of rental prices in Berlin,' Regional Science and Urban Economics, Volume 51, pp. 88-100. ISSN 0166-0462.

Ma, Y., Zhang, Z., Ihler, A., & Pan, B. (2018) 'Estimating Warehouse Rental Price using Machine Learning Techniques,' International Journal of Computers Communications & Control, 13, 235-250

Zhongyun, J. (2019). Prediction of house price based on the back propagation neural network in the keras deep learning framework; prediction of house price based on the back propagation neural network in the keras deep learning framework

Gupta, R., Andr e, C. and Gil-Alana, L. (2015). Comovement in euro area housing prices: A fractional cointegration approach, Urban Studies 52: 3123-3143.

Zhang, H., Weng, S., & Zhou, X. (2007) 'Housing price fluctuations across China: An equilibrium mechanism perspective,' Tsinghua Science and Technology, 12(3), 302-308.

Zhan, C., Wu, Z., Liu, Y., Xie, Z., & Chen, W. (2020) 'Housing prices prediction with deep learning: an application for the real estate market in Taiwan,' in 2020 IEEE 18th International Conference on Industrial Informatics (INDIN), Warwick, United Kingdom, pp. 719-724.

Gallin, J. (2008) 'The Long-Run Relationship Between House Prices and Rents,' Real Estate Economics, 36, pp. 635-658.

Stanley, S., Lyons, R. C. and Lyons, S. (2016). The price effect of building energy ratings in the dublin residential market, Energy Efficiency 9: 875–885.

Rana, V. S., Mondal, J., Sharma, A. and Kashyap, I. (2020). House price prediction using optimal regression techniques, Institute of Electrical and Electronics Engineers Inc., pp. 203–208.

McCord, M. J., Davis, P. T., Haran, M. and McCord, J. (2016). Analysing housing market affordability in northern ireland: towards a better understanding?, International Journal of Housing Markets and Analysis 9: 554–579

Lyons, R. C. (2018). Credit conditions and the housing price ratio: Evidence from ireland's boom and bust, Journal of Housing Economics 42: 84–96.

Li, Z. (2021). Prediction of house price index based on machine learning methods, Institute of Electrical and Electronics Engineers Inc., pp. 472–476.

Lyons, R. C. (2015). Measuring house prices in the long run: Insights from dublin, 1900-2015.

Dupre, D. (2020). Urban and socio-economic correlates of property prices in dublin's area, Institute of Electrical and Electronics Engineers Inc., pp. 556–562.

Corrigan, E., Foley, D., Mcquinn, K. and Slaymaker, R. (2019). Exploring affordability in the irish housing market.

Lima, V. (2019). Towards an understanding of the regional impact of airbnb in ireland, Regional Studies, Regional Science 6: 78–91.

Lyons, R. C. (2019). Can list prices accurately capture housing price trends? insights from extreme markets conditions, Finance Research Letters 30: 228–232.

Wu, H. and Wang, C. (2018). A new machine learning approach to house price estimation, New Trends in Mathematical Science 4: 165–171.

Tang, Y., Qiu, S. and Gui, P. (2019). Predicting housing price based on ensemble learning algorithm, Institute of Electrical and Electronics Engineers Inc.

Mirg, V. (2022) Prediction of property prices of Dublin housing market using Ensemble Learning, NORMA@NCI Library. Available at: https://norma.ncirl.ie/id/eprint/6229