Prediction of Property Prices of Dublin Housing Market using Ensemble Learning

MSc Research Project
Data Analytics

Vani Mirg
Student ID: x19211538

School of Computing
National College of Ireland

Supervisor: Majid Latifi

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Vani Mirg
Student ID:	x19211538
Programme:	Data Analytics
Year:	2022
Module:	MSc Research Project
Supervisor:	Majid Latifi
Submission Due Date:	31/01/2022
Project Title:	Prediction of Property Prices of Dublin Housing Market using Ensemble Learning
Word Count:	7879
Page Count:	22

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:	
Date:	31st January 2022

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):	

Prediction of Property Prices of Dublin Housing Market using Ensemble Learning

Vani Mirg
x19211538

Abstract

Dublin Housing Market is considered to be one of the most decisive factors contributing to the economy's growth. Over the recent years, the Property Regulatory Authority has noticed a difference in the market price and the selling price of the residential houses. This study tries to predict the prices of residential properties using the data from PSRA (Property Services Regulatory Authority) from 2010 through 2021 and determine the actual market value of any residential house. The research focuses on building a model using Ensemble Learning and traditional methods of Machine Learning and determining the best model for predicting property prices. The final results reveal the performances of the five models developed: Multiple Linear Regression, K Nearest Neighbour, Decision Tree Regressor, Random Forest Regressor and Gradient Boosting Regressor. The results suggest that the Gradient Boosting Regression Model accurately predicts the price and has an R-Square value of 75.22

1 Introduction

The capital of Ireland, Dublin, has been home to 544,107 people. Over the last few years, home buyers have faced difficulty buying residential houses. The Central Statistics Office (CSO) has also claimed that there has been a shortage of houses, as the demand increased along with high prices. The statistics were performed based on data of 2018 by Central Statistics Office Ireland [1], only the top 25% of earners would be in a position to buy a house with a 90% mortgage. According to the latest report by Central Statistics Office Ireland, property prices have more than doubled across the country in the past eight years, increased by 12.4%. Even reported by the Irish Examiner in one of its news articles [2], the property prices have risen over the past years, and only high earners are likely to afford a house. The recent years have been a bumpy ride for the home buyers.

All of this information does not cover everything about the housing situation of Ireland. The home buyers have also claimed a difference in the market price and the selling price of the houses. It is mainly because the houses are bought mainly by a mediator, which is either daft, rent.ie or a letting company. Therefore, to better understand the housing market of Dublin, it is essential to explore its background that led to the fluctuations of property prices.

[1] https://www.cso.ie/en/statistics/prices/residentialpropertypriceindex/

[2] https://www.irishexaminer.com/news/arid-40745031.html

The research aims at investigating the residential property prices across Dublin using the data published by the Property Services Regulatory Authority that controls the prices of the registered commercial and residential houses in Ireland. This research aims to help the direct and indirect stakeholders like home buyers, letting companies, property agents, and academic students. The study focuses on analyzing the data from 2010-2021 recorded by PSRA that contains information like Date of Sale, Price, Address, Market Price, VAT Details, Property Description and Land Area. Using the information provided by PSRA, the research is based on building a predictive model by exploring the state-of-the-artwork and extending it further with this research's methodology.

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

- **RQ**: Which technique is the best technique to predict the market price of houses in Dublin?
- **Sub-RQ**:Which factors affect the price of a residential properties?

The main Research Objectives of the Research Project are:

- **Study on the state of the art works**: The exploration of recent work done in a similar field would contribute to knowing the possible methods of predicting price. Section 2 summarizes the papers studied to understand the topic better.
- **Propose a Research Method**: The second objective highlights the methodology of the work done. It involves stages of data collection, data preprocesssing, data mining, data visualisation, training the data. This stage refers to shaping the data into knowledge.
- **Implementation of models**: Running a model on a dataset is the most crucial task to do hence, this research objective holds the truthfulness of the work done. The proposed work aims to implement a model using Ensemble Learning and traditional machine learning models. All the implementation has been done in Python, specifically on Anaconda.
- **Evaluation of the developed model**: The Evaluation stage is the final stage of the research which deals evaluating the model on parameters like R-Square, RMSE and MAE.

The research topic is quite common, and many researchers have worked on a similar idea. Although, the chosen area of interest holds great importance. The study not only focuses on a general house price prediction model but also explicitly targets Dublin's housing market, which is, in reality going through a housing crisis due to price inflation. The main contributions to the research are as follows:

- This research was capable of identifying the nature of the Dublin Housing Market. It does it by analyzing each of the relationships of the attributes with the Price.
- The research talks about the challenges of the chosen dataset and how they were tackled during the model building phase.
- Lastly, the research compares the performances of the best model with the state-of-the-art work and underlines the advantages and the shortcomings.

The rest of the document is divided into sections as follows. The following section, Section 2, outlines the literature survey focusing on the background of work done. Section 3 describes the methodology for the building of the model followed by Section 4 defines the design tools. Section 5 showcases the steps followed during the implementation, Section 6 briefly compares the chosen models and outlines the best technique, and Section 7 highlights the conclusion and the future work.

2 Related Work

This section covers an essential part of the framework which resonates with the latest findings for the problem statement. In the last ten years, there has been quite a lot of research conducted on housing price prediction of the various parts of the world, and Ireland's market condition has explicitly showcased the different independent factors contributing to the economy and underlined their shortcomings. The survey has been structured from the earliest methods to the latest methods on selling price evaluation of the Irish Market. The survey is divided into further subsections based on: Factors Affecting the Irish Housing Market, Statistical Techniques to Analysis House Prices in Dublin, and Machine Learning and Deep Learning Techniques to Predict House Prices.

2.1 Factor Affecting the Irish Housing Market

This section introduces the surveys led by many researchers based on the exploration of the factors directly or indirectly affecting the housing condition of Ireland. In Lyons (2018), the author examines the veracity of how the house price ratio was throughout the harsh housing market years of 2010-2012. A time-series analysis of quarterly-based data was used to get the results, which considered housing units' sale and rental prices. In line with the author's approach, the model explained more than 86% of the variance in the housing price ratio and had a root mean square value of 0.012%. It suggested an annual growth of 9% between 2000 and 2016. It was also established by this study that home prices have a fragile relationship with credit conditions. However, the study's conclusions were solely confined to credit circumstances, and it may have been enhanced by including more basic criteria such as the price of a house in its analysis.

Irish Market is also affected by employment, claims the authors Agnew and Lyons (2018). The study reveals that the housing market's rate are dependent on the job opportunities in a region. The data for carrying out the survey was taken from daft.ie and government site that handles the annual list of employment in Ireland. The cause of employment is noted from 2009-2013, and it is seen that FDI jobs have 1% and 2% effect on rent and sales prices, which means that the wealth increase in owner-occupied real estate by 48 million euro and 8 million in rental sector. Furthermore, the study revealed that after 1-2 years, rents are supposed to increase by 0.5%-1% in the rental sector and at least 2% for sales. In all, these results confirm that the housing market and employment are dependent on each other on one of the core intuitions. In other words, the location of offices has a great influence of housing marker in Ireland.

The authors, Stanley et al. (2016) demonstrated empirically that the energy performance rating of Dublin houses affects the market price of the properties. The hedonic prices are derived from a list of 2792 houses on daft.ie, which includes information about the houses' location, a number of bedrooms, type, and size. The findings indicated that energy efficiency positively correlated with housing prices in Dublin. A 50-point increase in the EPI resulted in a 1.5 percent increase in the list price. The EPI coefficient indicated 10% increase in the EPI is associated with a 0.87% increase in the list price of Dublin properties. Following the price and EPI relationship, regression results indicated that detached homes sell at a premium, as size plays a significant role in determining the price, and older buildings, in particular, haves the highest price. This study contributed to our understanding of the factors that influence price determination; however, it could not determine the percentage of houses dependent on BER ratings, as more than 40% of 2792 listings did not receive BER ratings. Nonetheless, this study demonstrates that electricity is a significant factor in determining the price of a home in Dublin.

The impact of Airbnb on the Greater Dublin Area's housing market has also been studied in the paper by Lima (2019). Airbnb has attracted a diverse audience due to the increased demand for short-term housing suitable for students, tourists, and those looking to avoid legal formalities. The increase in demand for short-term housing has impacted the demand for rental spaces, increasing rental space prices. One of the researchers, Dupre (2020) used a different strategy in determining the urban and socioeconomic characteristics that influence the property price in Dublin. The study's objective was to gather data from the PSRA platform using geocoding to find the landmarks that impact the pricing and compute the forecast using Extreme Gradient Boosting. The experimental result revealed a variance of 43%, indicating that distance to the city, monuments, tourist attractions, businesses, and embassies may be a determining factor in determining property prices in Dublin. However, the primary limitation is that none of the attributes in the dataset were utilized, except for the address.

In one study based on the indirect factors affecting the housing condition, one of the authors, Zhang et al. (2021) constructed a dynamic-based approach by building a network and calculating the correlation between the places situated in the UK, precisely Europe. The study revealed that the Northern Ireland part is least integrated with the housing market of the UK, which implies that its neighboring region, the UK, does not impact Ireland's housing situation. All of the information gathered by the mentioned articles explained one thing: the housing condition is not only affected by the parameters of a house but also by factors like employment, electricity rates, tourism, and price of neighboring regions.

2.2 Statistical Techniques to Analyse House Price in Dublin

The literature in this section is based on statistical approaches used by researchers to analyze the Dublin market. The studies included in this section belonged to when mathematical computing was at its peak, and machine learning was not practiced. The author Roche (2001) uses a regime-switching model to examine the growth in housing prices in Dublin. He uses several methods to break down Dublin's housing prices into fundamental and non-fundamental components. Statistical Bulletins of Consumer Price Index and the Department of Environment and Local Government's Annual and Quarterly Housing Bulletins were used in the research. Method A-D and its coefficients of t-statistics, probability values, and Wald tests are used to determine the findings. The findings reveal that market fundamentals solely determine prices, regardless of the technique employed to evaluate the non-fundamental price. The second piece of evidence examined whether the overall regime-switching model's predicted intercept and slope coefficients were accurate. Different regimes have different intercept and slope coefficients suggested by the bubble model. The third evidence verified if the calculated slope coefficients of the regime-switching model were significant. As shown by the three pieces of proof mentioned earlier, a speculative bubble in Dublin's housing market is evident. With an economic boom and reduced mortgage rates expected in the next century, the demand for property in Dublin is expected to rise significantly.

Between 2006 and 2016, the researchers Corrigan et al. (2019) examined the distribution of housing expenditures across different kinds of households in Ireland. The research employed a micro-level dataset to analyze income and living conditions across different household types. The study yielded key conclusions: Housing costs accounted for one-fifth of family income in 2016. The SILC statistics revealed that, although house affordability concerns are not universal, many groups face significant barriers. Private renters, those who live in Dublin (and its commuter regions), and those on low incomes have the most severe difficulties. Indeed, housing costs accounted for between two-fifths and more than half of income for families in the poorest 25% of the income distribution, versus one-fifth for other households. Although the total payment-to-income ratio for housing increased between 2006 and 2016, the repayment-to-income ratio for low-income households increased dramatically between 2008 and 2016. Additionally, the authors discovered that low-income households (those in the bottom 25% of the income distribution) in the private renting sector have historically faced high housing expenses. The second evidence spoke about households that spend more than 30% of their income on housing costs and are in the bottom 40% of the income distribution. Most families in this 30/40 generation were private renters with very little residual income once housing costs were met. The last empirical contribution was to determine if the strict 30/40 cut-offs are appropriate parameters if international recommendations are used to construct a realistic definition of high housing expenses in Ireland. The authors did this by examining how much monthly income households have after housing expenditures are deducted. The experimentation showed that residual earnings for persons with high housing costs did not begin to increase significantly until the 60th percentile of the income distribution; thus, focusing only on families in the bottom 40% may overlook those with similar residual earnings. According to their reported results, salaries seem to be a critical factor in deciding which households face Ireland's most severe affordability challenges.

The researcher, Lyons (2015) offered information in their study about a descriptive examination of the Irish housing market from 1980 to the early 1920s. The study examined how economics and other variables such as neighboring nations, owner-occupied housing demand, housing supply, trends, and modeling have influenced Ireland's housing expansion. The results indicated that between 1995 and 2000, Ireland's housing market was influenced by fast growth in the process, high levels of demand, and limited supply, all of which impacted the expansion in family wealth, increasing personal spending. Moreover, statistics cited in the OECD Economic Survey Ireland 2001 indicated that Ireland had a substantially lower home mortgage than any other European nation. Their research noted that the current house prices were determined by land values and other house-related characteristics and by the economic circumstances of the historical period. Also, the study presented by Moro et al. (2013) examined whether or not the Greater Dublin housing market reflected cultural heritage. The researchers assessed several elements in a hedonic pricing equation to determine whether the distance and density of the cultural heritage site are factored into Greater Dublin house costs. The data was extracted from a Geographical Information System (GIS) dataset that included the location and attributes of homes acquired in Ireland between 2001 and 2006. According to the descriptive statistics extracted from the dataset, most residences in Dublin have a garden and parking. The second data set was given by Sherry FitzGerald, Ireland's leading property advisory business and auctioneer; it included transactions from 2001 to 2006. Using statistical approaches and geocoding, it was shown that some historical places had a positive spillover effect on surrounding homes, while churches had a negative effect on property values. The existence of historical culture had a net influence on the pricing of 0.4 percent 0.6 percent. This work helped establish a correlation between Dublin residential properties and economic reports.

Another statistical approach pioneered by academics, McCord et al. (2016) was used to investigate the affordability of the Northern Ireland property market. The tool used for the analysis were Principal component analysis, Granger causality test, and Johansen co-integration tests. Using PCA initially, the authors found the inter-relationships between the dataset features, reducing the dimensionality by maximizing the variance between the variables. Following that, on the same variable set, co-integration modeling was conducted in two phases: verifying the number of integrations and performing the unit root test, also known as the Augmented Dickey-Fuller (ADF) test. The investigation revealed various divergent economic and financial indicators influence the housing market. As a result, policymakers should delve further into the underlying reasons for home affordability issues. The researcher, Lyons (2019) conducted an in-depth examination of whether or not property prices and list prices in Ireland varied. He analyzed two population-level datasets on the Irish property market from 2006 to 2012: online listings and mortgage drawdown. The prices of homes were broken into four components: selection spread, matching spread, counteroffer spread, and drawdown spread; these phases indicate the price differential between searching for property and finalizing the purchase. The data indicated that although the counteroffer spread was substantial near the conclusion of the price bubble, the matching spread was by far the most significant contributor during the down market. The investigation discovered a discrepancy between the house prices known to policymakers and the registered list price. Although this scenario in Ireland seems feasible in theory, the study had a flaw in that the data utilized for the analysis consisted only of property prices and not of house attributes. As a result of these factors, our study could not provide a clear image of the property price.

The researcher's analysis Gupta et al. (2015) examines the correlation between home prices throughout the Eurozone. Co-integration and fractional integration were used in the methodologies. According to the research, individual log-real price indices showed orders of integration over one, signifying long-term growth rates. Furthermore, it was shown that the Euro area is co-integrated with Belgium, Germany, and France. The findings drawn from exploring the statistical tools provide clear evidence that the housing market has been dependent on house buyers' heritage culture, income, and financial stability.

2.3 Machine Learning and Deep Learning Techniques to Predict House Price

This section covered the recent technologies being practiced to predict the price. Different authors have used different techniques like a hybrid, deep learning, time series for predictions. The authors Wu and Wang (2018) in their study built a house prediction model using Random Forest, a machine learning approach, and compared the model's accuracy with a widespread yet effective linear regression model. The data contained nearly 26800 records having features such as the lot size, year, quality of the house, distance from the city, and many others. The approach adopted by the researchers was first to use linear regression and compare the score with the random forest model's result. Then, the test was repeated on different sets of features related to the house by comparison, and they successfully showed how Random Forest could take advantage of that information and give better estimation. For the model building, the whole data set was divided into a 7:3 ratio which means 70 percent as the training set and 30 percent as the testing test. The results were based on R-Square and RMSE values. The R-Square and RMSE values for each set of features using Random Forest varied, as it meant that every non-linear relationship between the features was added to the information. The authors concluded from their research that the most common approach, i.e., linear regression, is not always the best in describing prediction problems as sometimes even the non-linear variables are helpful and add to the advantage of the prediction.

The authors' Tran et al. (2017) analysis was ranked first in one of Kaggle's most considerable Data Science Challenges for forecasting the home price index for residential properties. The information contained characteristics such as housing type, size, and proximity to amenities. The authors first preprocessed the data, designed features, and used a hybrid Lasso and Gradient Boosting Regression. They constructed a Lasso and Ridge regression model, followed by a Gradient Boost Algorithm with a fixed learning rate of 0.1 and 1000 estimators. The study established that hybrid regression is superior to a single regression technique. The best hybrid model had 230 characteristics and a score of 0.11260, with 65% Lasso and 35% combination. The students Limsombunc et al. (2004) of American University, conducted comparative research on the property price prediction. Their research was based on comparing the hedonic pricing model and the artificial neural network. Their research asserted that, although the hedonic pricing model is the most often used approach, it does not accurately reflect the actual world or geography. On the other hand, Artificial Neural Networks functioning is similar to a human brain and forecast prices more accurately in real-world scenarios. Their findings support the same premise, demonstrating that the artificial neural network model predicted prices more correctly than a hedonic pricing model with an R-Square greater than 75%. Although this technique had several disadvantages, the home price utilized was not the actual sales price. The other is that the model was built having only one year-based record.

One of the researchers, Lu et al. (2019) presented a novel hybrid approach focused on forecasting property prices using neural networks and machine learning. The authors' analysis is distinct from previous work in that it does not use the conventional technique of estimating the price using lot size, land area, and home area. Instead, this technique used CNN (Convolutional Neural Network) and XGBoost as output layers. The framework uses photos from the AVA dataset, a collection of over 225,000 photographs spanning over 50 architecture, landscape, and interior dwelling categories. The model is composed of a pre-trained model trained on the AVA dataset, a Multiplayer Perceptions (MLPs) model with ReLu activation, another CNN model with ReLu activation, and an XGBoost component's final output layer. The model attained an MAE of 0.0332 and a

MAPE of 8.70% throughout the trial. The study differs from prior practices and adds to the presented approach. Another similar study based on machine learning was performed by students Rana et al. (2020) who tried to predict the housing price scenario using a Kaggle dataset which consisted of 13,320 records and 9 features. A comparative study showed the models using Decision Tree, Support Vector Regressor, Random Forest, and XGBoost. On performing the evaluation, the results showed that out of all the four algorithms, XGBoost had the highest training and test accuracy of 0.63 and 0.90. Although, the only main drawback to the study was the use of a small size dataset, as most real-world housing data have thousands of records compared to the data set used.

The study Gupta et al. (2020) computed the predictive power for four different models: multiple linear regression, Lasso and Ridge Linear Regression, Support Vector Machine, and Extreme Gradient Boosting. The methodology was similar to the traditional approaches, with a clear indicator that XGBoost had the most stable predictions with R-square for train and test set as 0.7869 and 0.75. More so, this paper introduced a new regression technique that other authors have not used, i.e. Lasso and Ridge Regression. The experimentation Tang et al. (2019) uses Random Forest and Bagging algorithm to predict the house price range using a class using a classification model. Ensemble learning is combined with multiple base learners for excellent performance. The authors have used a decision tree, random forest, and single decision tree as the base learners and judged the performance by the optimal depth using a verification curve to achieve the results. The model performance proves that an ensemble learning algorithm has higher prediction accuracy than a single decision tree. However, in terms of MAE and MRE, the prediction accuracy of random forest is better than bagging. The model developed shows the error between the actual houses and the prediction is low with an ensemble model and proves that bagging methods can be used to attain practical accuracy when using a large-scale dataset.

The author, Zhongyun (2019) have developed a model for the prediction of house prices in China using Neural Network. They have extracted the data using parsing and then applied a multi-layer feed forward neural network trained by inverse propagation algorithm. The most optimal model was achieved by using six hidden layers and the 'Relu' activation function of Back Propagation Neural Network with a learning rate of 0.001. The model predicted the housing rates quite well, having only a 5% error rate and accuracy as 95.59%. A similar study Li (2021) was put forward for the prediction of the House Price Index using two machine learning models and one neural network. The data was sourced from Kaggle, and the features included year, frequency, HPI flavour, HPI type, and the models developed were Lasso Regression, Ridge Regression, and BP Neural Network, nearly having close results. Out of all the three, BP Neural Network was slightly better; however, the tested choice of techniques could not predict the prices fully as stated by the authors.

All of the above-studied papers added new information about the Irish Market. However, the techniques to study Ireland's housing market in the late 90s were more based on theoretical concepts. The recent studies by various authors are majorly performed using machine learning techniques showcasing state of the art work. A comparison is made highlighting the most relevant techniques and their pros and cons.

2.4 Comparison of state of the art work

In Table 1 shows briefly all the important articles and the work done by the authors and the method they have used for similar work.

Table 1: Summary of the most related key works

Criteria	**Authors**	**Algorithms**	**Advantages**	**Limitations**
A new Machine Learning Approach to house price estimation	Limsombunc et al. (2004)	Random Forest,Linear Regression	Less Variance Score for Random Forest	Could not handle non-linear relationships
A Hybrid Regression Technique for House Price Prediction	Wu and Wang (2018)	Lasso and Gradient Boosting Regressor	Hyperparameter tuning	Modelled for small dataset
House Price Prediction: Hedonic Price Model vs. Artifical Neural Network	Limsombunc et al. (2004)	KNN and BP Neural Network	Neural network improved the model performance	Concentrated to smaller dataset
Deep Learning with XGBoost for Real Estate Appraisal	Agnew and Lyons (2018)	CNN and XGBoost	High accuracy	Lots of Training data required
House Price Prediction Using Optimal Regression Techniques	Stanley et al. (2016)	Decision Tree, SVM , Random Forest, XGBoost	XGBoost performed better than all	Poor Value of R-Square
Prediction Housing Price Based on Ensemble Learning	Tang et al. (2019)	Random Forest and Bagging algorithm	Weak Learners increase performance	OverFitting

3 Methodology

The research methodology discusses the methods used to achieve the study aim outlined in Section 1. The suggested research methodology does not resonate with a generic data analysis methodology, but it is designed according to the sequence of activities completed to determine the optimum residential housing prediction strategy. The Figure 1 emphasizes the sequence of procedures to be taken in order to achieve the desired result. Before developing the research plan, the primary guideline was to investigate the state-of-the-art work and identify the approaches with their benefits and limitations, as indicated in the table 1. However, the procedures in developing the research strategy are unique because the data set utilized in the project was distinct from the previous work. The strategy used in this study is thus wholly original since no such approach on the chosen dataset has ever been adopted previously.

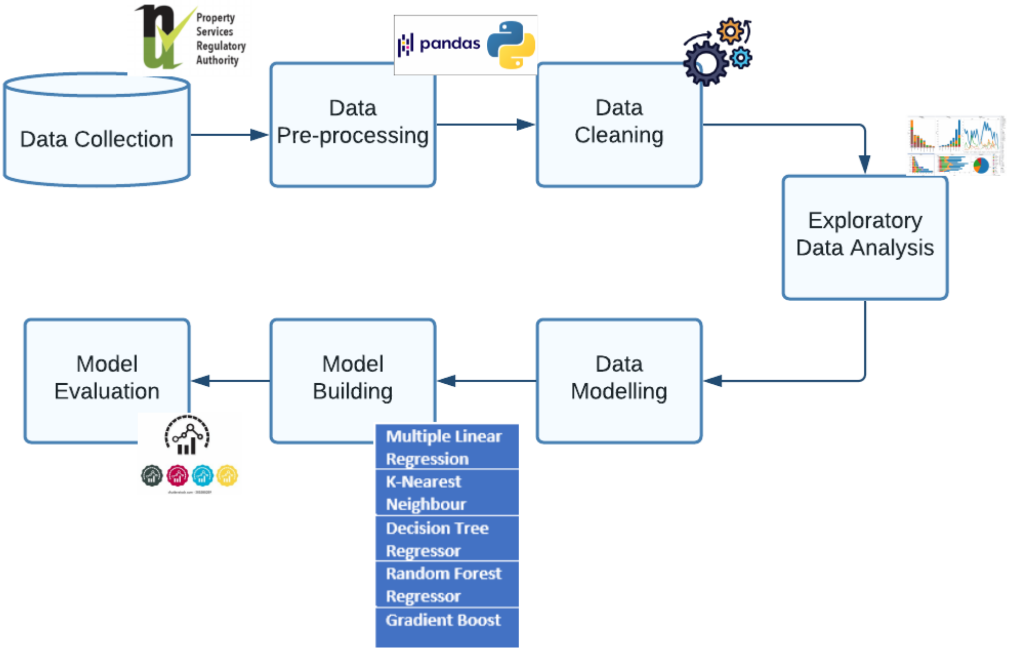


Figure 1: Proposed Research Methodology for Prediction of Dublin Housing

3.1 Data Collection

Data collection is a critical stage that leads to the first step towards the research objective. The data set used for the research comes from PSRA, a government-run platform that keeps track of Ireland's residential and commercial property prices. [3]. The site offers statistics spanning a decade for various prominent counties in Ireland, including Dublin, Wicklow, Kildare, Cork, and Maynooth. The data set selected for the study is based on the residential houses and spans over the years from 2010 to 2021 and comprises roughly 506020 rows. The conventional download CSV approach was used to get the data in this circumstance. When obtaining huge volumes using the API, a runtime error occurred. As a result, the data is downloaded as a CSV file and then processed using Python into a data frame. The dataset's characteristics are shown in Table 2.

3.2 Data Prepossessing

It is relatively uncommon in real life for the collected/original dataset not to be in a suitable format. Because the existence of missing information, noise, and outliers results in erroneous predictions, preprocessing is a critical component of the approach. The raw data is first translated into a structured format, and then each attribute is verified to see whether any of its values are null or if any of its values are not null. Each attribute in the dataset has been subjected to data preparation in order to facilitate further investigation. Data preparation staged operations included anything from altering the data type to manipulating the properties of the data. The detail techniques are discussed in Section 5.

[3] http://psr.ie/en/psra/pages/what_is_the_psra

Table 2: Summary of the Attributes in the Dataset

Attribute	**Data Type**	**Description**
Date of Sale	Numerical	Date when the property was sold
Address	Object	Address of the place
Postal Code	Object	Postal code with respect to the Address
County	Object	Part of Ireland it belongs to
Price in Euro	Float	Selling Price of the property
Not Full Market Price	Categorical	Whether the price is the full market price or not
VAT Exclusive	Categorical	Whether VAT is excluded or not
Description of Property	Categorical	Description of the kind of property: New or Second-hand
Property Size Description	Categorical	Description of Size of the property

3.3 Data Cleaning

Outliers and noise can cause inaccurate predictions, therefore it needs to be removed from the preprocessed data. The data can only be properly trained if it is tested for missing values, noise, and outliers. Section 5 discusses how to deal with missing values and outliers in the data collection, which comprises around 506020 rows, since it is impossible to assume that the data does not have any.

3.4 Data Exploration/Visualisation

Exploratory data analysis is carried out on the clean dataset, which results in interactive visualizations and charts. Data Exploratory Analysis is done using building charts and interactive plots. Without performing any Exploratory Analysis, it is impossible to understand the relation with the independent and the dependent features by drawing plots like scatter plot, histogram, bar plots demonstrate their relationship. Data Exploration in this project is supported using advanced Python data visualization libraries such as Altair, Plotly, Folium, Geopandas, and many more. Using these packages for visualizations contributed to interactive plots, making information better to understand. The other purpose that Data Exploration solves is answering the questions that arise by looking at the dataset. Data exploration with the help of visualizations is clearly explained in Section 5.

3.5 Data Modelling

Refining the data enhances features through feature engineering and aids in developing a more robust prediction model. There are many ways accessible for feature engineering; nonetheless, choosing the correct methodology for more precise outcomes is critical. This section of the project outlines the approaches employed in the model's construction. The data modeling step explicitly outlines the set of attributes for model training and how they were converted into binary form. Data Modelling carries a challenge in this research, as most attributes are categorical, and there were limitations to each when converting into binary form. Finally, the Implementation section explains the detailed logic and technique.

3.6 Model Building

After a deep understanding of the target and independent variables, the last stage is to identify the algorithms suitable for predicting price. The choice of techniques to build the model depends on two factors: based on the nature of the target variable, i.e., price (a continuous variable) and the state-of-the-art work. Most of the research done on similar problems has been studied using ensemble learning techniques, KNN, linear regression, and Random Forest. Therefore, this research aims to use five algorithms based on these two factors: Multiple Linear Regression, KNN, Random Forest Regressor, Decision Tree Regressor, and Gradient Boosting Algorithm and identifies the best out of the chosen five.

3.7 Model Evaluation

The model assessment is the last phase of the proposed process, in which the models are assessed using performance measures. Because this is a regression-based subject, the parameters for assessment would be R-Square, Root Mean Square Error (RMSE), and Mean Absolute Error (MAE). The models would be compared based on how close their forecasts are. The evaluation is covered Section in 6, and prediction charts for each model are included with the code artifact.

4 Design Specification

- This section highlights the tasks performed in sequential order according to the proposed methodology. The design specification presents a process flow diagram shown in Figure 2 to solve the proposed research objective. The stages described are the steps performed: data collection, data preprocessing, data visualization, feature engineering, data modeling, Model Building, and Evaluation implemented using Python.
- The data is initially collected into a csv file, which is then imported into Python. The dataframe is preprocessed and cleaned, and additional features are incorporated. It is then followed by selecting a collection of characteristics from which to construct the test and train set for applying the five models shown in Figure 2. Each model is then tested using performance metrics and a prediction graph.

5 Implementation

This project stage talks about how the data set was transformed for model building and how the chosen five techniques: KNN, Multiple Linear Regression, Gradient Boost, Decision Tree Regressor, and Random Forest Regressor, were implemented along with the results.

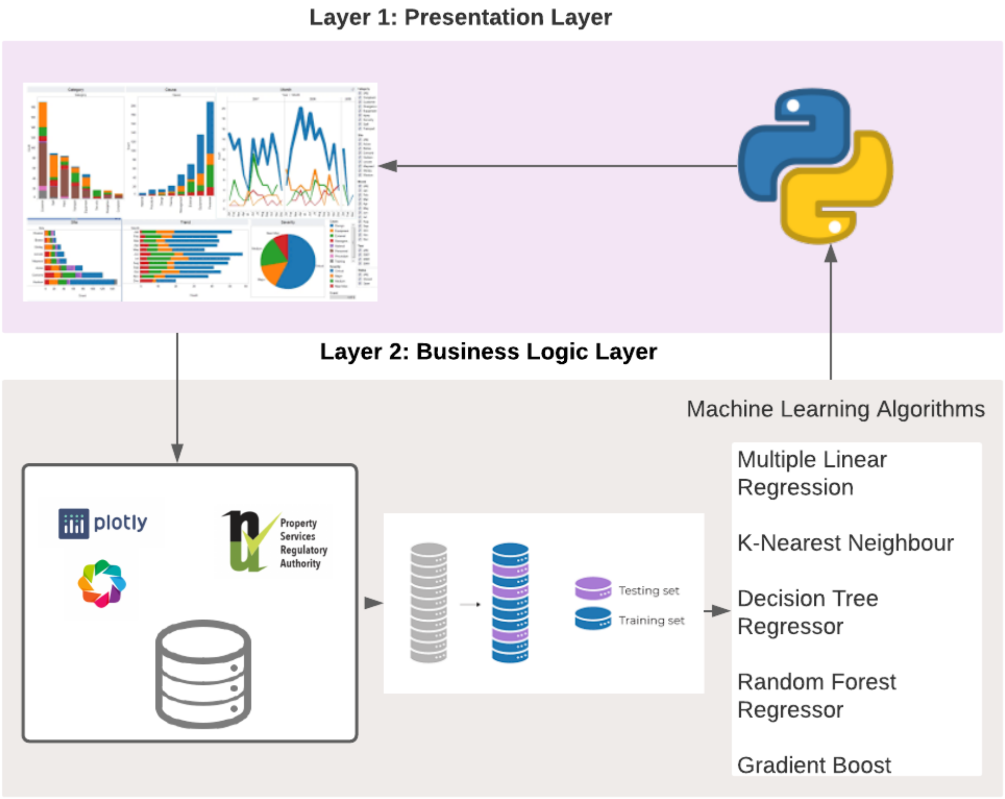

Layer 1: Presentation Layer
Layer 2: Business Logic Layer
Machine Learning Algorithms
Multiple Linear Regression
K-Nearest Neighbour
Decision Tree Regressor
Random Forest Regressor
Gradient Boost
plotly
Property Services Regulatory Authority
Testing set
Training set

Figure 2: Design Specification Architecture

5.1 Data Preprocessing

The data set for this project was collected from the official website of the Property Services Regulatory Authority, which included records of allowed property sales in every county in Ireland between 2010 and 2021. The data set for all counties during the previous 11 years was obtained as a csv file comprising more than 506020 entries for the analysis. Due to the large volume of information, data preparation was an intriguing and challenging phase of the project. The raw data was imported into Python 3.7 and then put in a dataframe for analysis using pandas. Each attribute in the dataset has been subjected to data preparation in order to facilitate further investigation. Data preparation staged operations included anything from altering the data type to manipulating the properties of the data. The preprocessing processes that were done on each attribute are described in brief below

- **Reviewing the columns**: After seeing that there were spaces in all but one of the stated nine columns, I decided to change the names of those columns to underscores instead. Initially, the 'Price' column featured the Euro currency sign, causing the values to be incorrectly translated. For this reason, the currency sign was omitted.

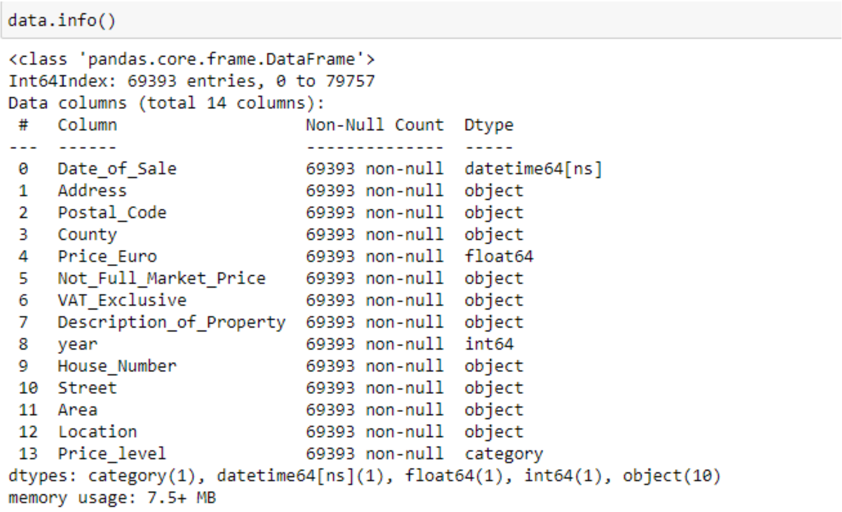
- **Adjustment of Data types**: Due to the fact that all columns were initially treated as object data types, this might have resulted in inaccurate analysis. As a result, the cardinality and character of each column were altered. Qualitative and quantitative attributes were used to classify the attributes.

- **Derivation of new attributes**: Certain characteristics in the dataset, such as Address, Postal Code, and Date of Sale, include some missing data. To have a better grasp of these traits, additional columns are formed from the three existing ones based on the data. To begin, a property called 'Year' is extracted from Date of Sale, allowing us to easily trace the change in house values over time. Second, a new property called 'Location Dublin' is produced from the 'Postal Code' element, which distinguishes the Northern and Southern Postal Areas. Finally, the address characteristics include the whole of the address, which cannot be interpreted in any other way. As a result, the address is converted to 'House Number, 'Street', and 'Area'. The development of additional qualities aided in determining the link between the features, which is then utilized for visualisations and model building.

- **Geocoding**: Using geopandas, the coordinates are extracted from the address attribute and new columns are constructed as 'latitude', 'longitude', and 'point'.

- **Binning**: The attribute 'Price Euro' consists of prices which are infinite prices which are hard to interpret. Using binning on price attribute can help with categorising it into levels by creating bins as low, medium, high. The value range for each category is taken by using 'linspace' function in pandas.

The final dataset after sucessful preprocessing and cleaning resulted into a dataframe as shown Fig 3



```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 69393 entries, 0 to 79757
Data columns (total 14 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Date_of_Sale             69393 non-null  datetime64[ns]
 1   Address                  69393 non-null  object
 2   Postal_Code              69393 non-null  object
 3   County                   69393 non-null  object
 4   Price_Euro               69393 non-null  float64
 5   Not_Full_Market_Price    69393 non-null  object
 6   VAT_Exclusive            69393 non-null  object
 7   Description_of_Property  69393 non-null  object
 8   year                     69393 non-null  int64
 9   House_Number             69393 non-null  object
 10  Street                   69393 non-null  object
 11  Area                     69393 non-null  object
 12  Location                 69393 non-null  object
 13  Price_level              69393 non-null  category
dtypes: category(1), datetime64[ns](1), float64(1), int64(1), object(10)
memory usage: 7.5+ MB
```

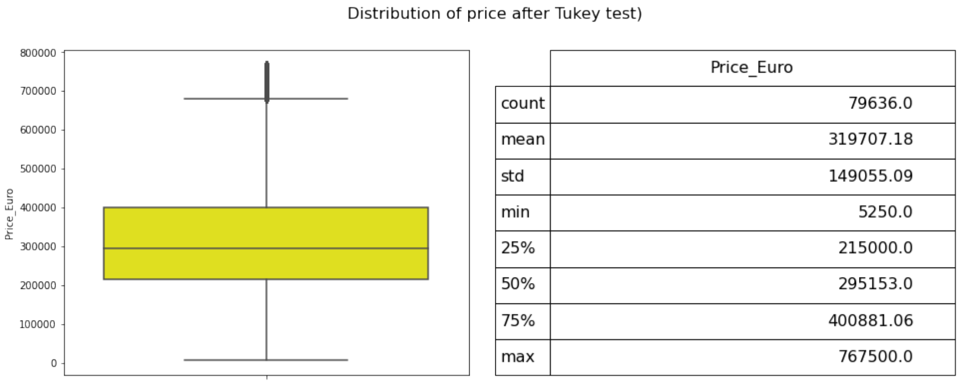
Figure 3: Pre-processed Dataset

5.2 Data Cleaning

At this stage, identification of missing values take place. Removal of missing values and outliers can reduce the chance of incorrect predictions and more accurate results

- **Treatment of Missing Values**: Only two characteristics in the dataset are lacking values, namely Postal Code and Property Size Description. 'Property Size Description' had around 88 percent missing data, which did not seem to be significant; hence, it was eliminated. The other property, 'Postal Code,' had roughly 30% of its values missing, which was dealt by eliminating just the NaN values.

- **Handling Outliers**: An outlier is a single observation that seems to diverge significantly from the rest of the sample. The data set included substantial outliers in 'price'. As a result, the outliers were identified using the Tukey Test. This test calculates the upper and lower gates using interquartile ranges. All outliers were eliminated using the Tukey test, as seen in Fig 4



	Price_Euro
count	79636.0
mean	319707.18
std	149055.09
min	5250.0
25%	215000.0
50%	295153.0
75%	400881.06
max	767500.0

Figure 4: Outlier Treatment

5.3 Exploratory Data Analysis

Since the data set has large rows, it is not easy to look at the features and understand their relationship with the other attributes. Therefore, EDA on the formatted data set is done to understand the effect of price, type of housing, affordable locations in Dublin. Some of the basic patterns discovered from the visualizations are listed as follows:

- **Visualisation 1**: The bar charts highlights how the prices have varied over the last 11 years. It is clear from the Figure 5 that the prices have increased in Dublin by 15%- 17% percent every year up to year 2019 and the it started to fall. The reason of fall of prices seems practical as after the pandemic, the sales market have seen a fall in prices by 10%. As said by a housing market analyst Rohan Lyons, an associate professor in economics at Trinity College Dublin that the number of homes coming onto the market has fallen by about 40% in some parts of Dublin.`https://amp.rte.ie/amp/1210158/`

- **Visualisation 2**: The vertical stacked bar chart shown in Figure 6 demonstrates the ratio of second-hand property to new apartments across all the postal areas in Dublin. The insights derived from the visualization are: Firstly, most property listings are in Dublin 15, followed by Dublin 18 and Dublin 24. Secondly, every postal code is more than 70% of the total property listings as Second-hand Dwelling. This observation tells us about the nature of people willing to buy property; they prefer buying second-hand property as the prices are relatively lower than brand new apartments.

- **Visualisation 3**: The graphical visualization represents in Fig 7 a pie chart showcasing the distribution of price level i.e. low, mid, high prices, and a vertical bar chart showcasing the count of listings in each region: North and South. It is evident that approximately 50% house prices fall under the medium-range price category

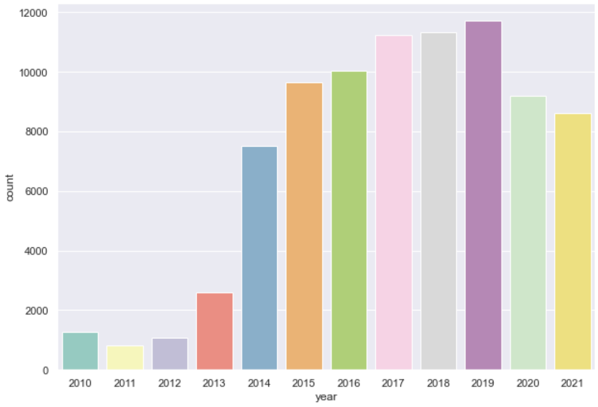


Figure 5: Sales from 2010-2021

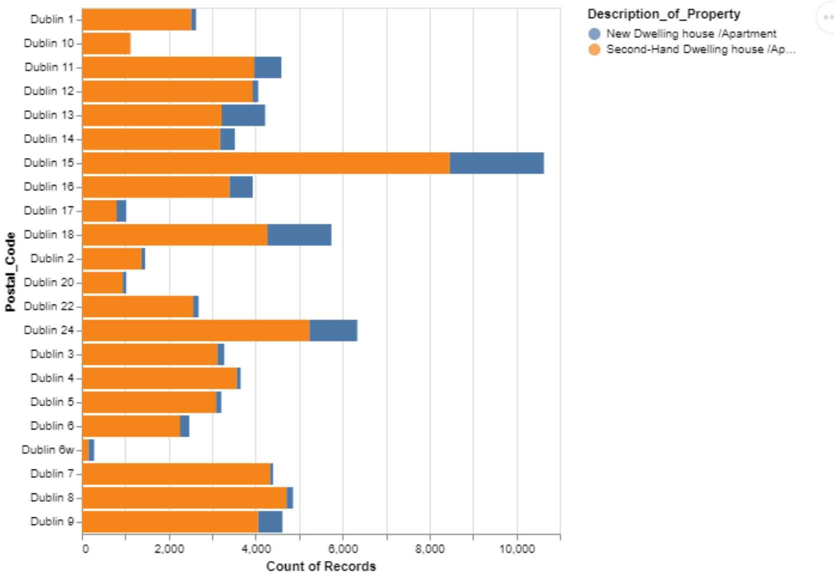


Figure 6: Distribution Second-hand and New property across Dublin

and hardly 11% under the high range. It seems practical as we know that the nature of buyers in Dublin is to go for second-hand housings. Furthermore, the bar chart shows that most of the houses sold are in Dublin's South region; South regions have an even digit at the end of the postal codes like Dublin 6, Dublin 8, Etc., whereas odd digits represent North regions.

- **Visualisation 4**: Figure 8 shows a map highlighting the parts of Dublin with the property listings. The map is formed by geocoding the addresses in the data set into geographical coordinates.

All the above visualization adds to knowing more of the Dublin Housing Market. First, it is pretty evident from these figures that Dublin Housing Market as suffered a significant fall in property prices during and after the pandemic, due to which the demand of home buyers have increased, and more realistically, at this point, people in Dublin are facing a housing crisis as properties are only available for renting purpose and not for buying. The second information added is that prices in Dublin are based on the region; the North is supposed to have lower prices than the South.

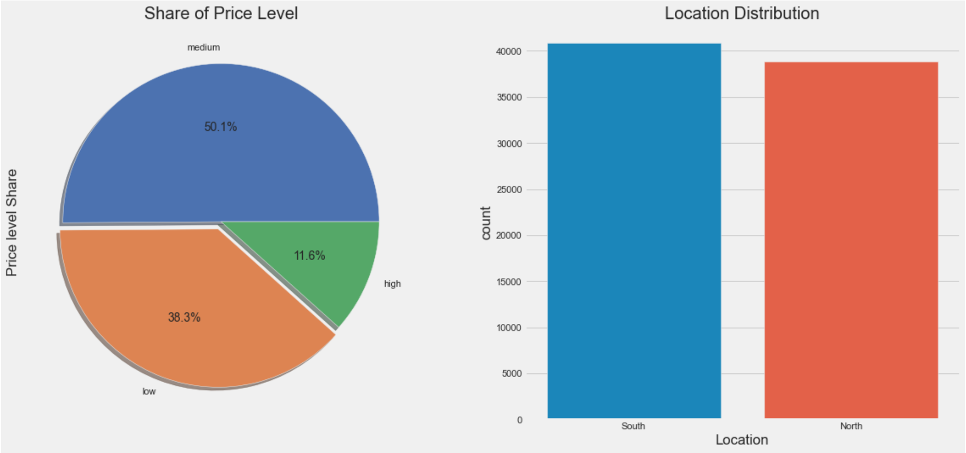


Figure 7: Price level and Region Distribution

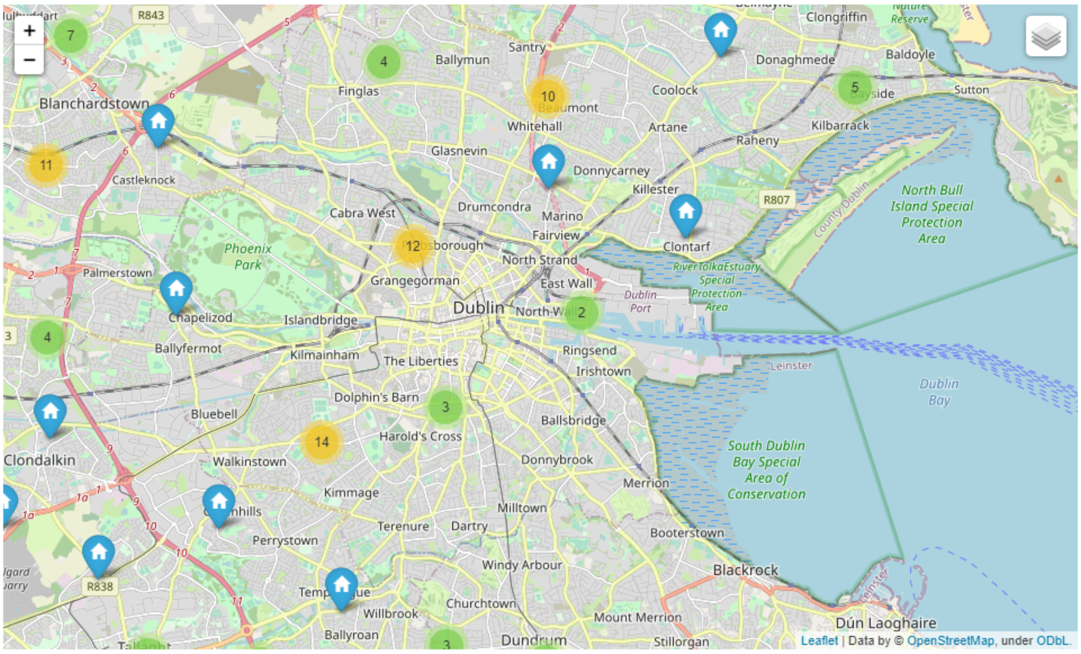


Figure 8: Map demonstrating the listings in different regions

5.4 Feature Engineering

Feature Engineering refers to transforming the essential features into trainable form. Most of the features present in the data set are categorical and cannot be ignored while building a model. Therefore, it is essential to restore the information of the categorical variables into integer values. For this purpose, all the categorical variables are transformed using the given techniques

- **Label-Encoding**: Label-Encoding is one of the most effective techniques of converting categorical variables into numerical values. While modeling the data, all the categorical variables must be converted into binary form for the algorithms to interpret it correctly. The categorical variables that were converted using Label-Encoding are: 'Not Full Market Price','Price Level', 'VAT Exclusive', 'Location'.
- **Split Function**: Split function is a technique in which the categorical values sep-

arated by ‘,’ can be further divided. It is used when the attribute has two or more information given in one column. For instance, the complete address of property listed has three components: lane address, town, and Area. So, when building a model, this information can be used using the split function.

5.5 Model Selection

Five algorithms were selected based on a review of the literature on Machine Learning approaches for price prediction, in Section 2.3. Numerous writers discussed classic, hybrid, and ensemble learning approaches; the table Table 1 highlights the work closest to state of the art. For the Model Selection step, the algorithms considered are Multiple Linear Regression, K Nearest Neighbour, Decision Tree Regressor, Random Forest Regressor, and Gradient Boost. Before developing the models, the train and test set are separated into 80% and 20%, respectively.

5.6 Implementation of Models

- **Multiple Linear Regression**: Multiple Linear Regression is one of common techniques applied when dealing with regression problem. To predict the price of the property, there could be more than one independent variables present in the data set, to evaluate the affect of the features, multiple linear regression can be a great tool.The model for the multiple linear regression consists of ’X’ having features like ’Postal Code’, ’County’, ’Not Full Market Price’, VAT Exclusive’,’Year’, ’Location’, and ’Price level’ and ’Y’ consists of the target variable: ’Price’. The model dealt with the assumptions of multicolinearity, Normality and homoscedasticity.

- **K Nearest Neighbour**: K Nearest Neighbour model is based on predicting the price by measuring the Euclidean distance. It is very important to choose the right number of neighbours fore achieving accurate score. The model created is first tested using random value of n, say 5 and then the best value of N neighbours is chosen by iterating through a loop. According to the results, it is observe that at N= 7.

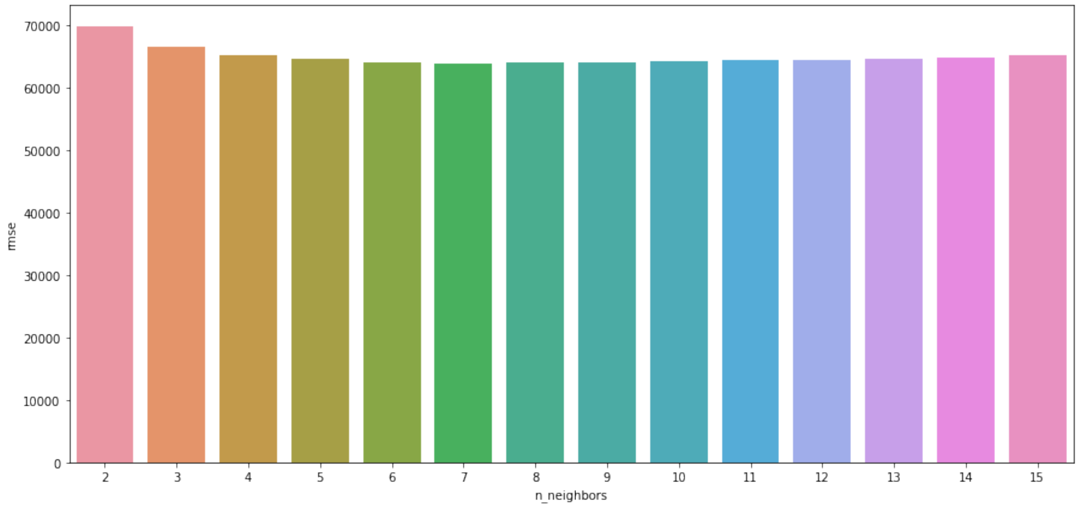


Figure 9: KNN Neighbour

- **Decision Tree Regressor**: The Decision Tree Regressor is a strong supervised learning technique for prediction. The advantage of building a model using Decision Tree is that due to non-linearity, the model's performance is not affected. Using sklearn tree class, Decision Tree Regressor is imported and fit on the X and Y train set. The results are predicted using regression.predict function and is assinged to 'YPred'.

- **Random Forest Regressor**: To overcome the problem of overfitting of Decision Tree, another technique known as Random Forest Regressor is applied to the training and test sets. Random Forest is an Ensemble Learning technique, specially a Bagging technique which uses random samples of data with replacements that reduces the bias. Random Forest works om the strategy of bagging, which means it makes each model run separately and then combines the performance to predict the performance.

- **Gradient Boost Regressor**: Another, very effective technique other than Bagging is Boosting which is also an Ensemble Learning technique. Trees are added in sequential order to improve the prediction made by the prior model.For initialization on Gradient boost, three important parameters to improve the performance are taken into consideration which are: Depth and the number of tree, learning rate and random subset of fitting tree. During the implementation, the number of estimators chosen were 5000 and the learning rate was set to 0.02

6 Evaluation

This section provides an overview of the all the five models used by evaluating them on the basis of the metrics. The metrics for the evaluation of the models depend on three aspects: RMSE (Root Mean Square Error), R-2 Square and MAE.

- **Multiple Linear Regression Model**: The first tested model was built using Multiple Linear Regression. Despite of being aware of the facts that Linear regression models cannot deal with non-linear relationships, this model was implemented to see how much better it performs than the state of the art work models on similar dataset. The model recorded to have a R2-Square value of 67.83, indicate that the predicted data points are nearly 68% close to the predicted value.

- **K Nearest Neighbour Model**: A KNN model scored a variance score of 62.81% which was lower than a Multiple Linear Regression model. The main reason of low performance of this model could be a large dataset. Otherwise, this model worked averagely good.

- **Decision Tree Regression Model**: Decision Tree Regression model is used to solve the problem of infinite possible values of price. Even though the training time required to train the model was high, the results of a decision tree model were better than the KNN and Multiple Linear Regression Model. For instance, a house with a price of worth 215000 Euro, the model predicted its price to be 200705 Euro. This model have an overall R-Square value of 70.98.

- **Random Forest Regression Model**: This model which is based on Ensemble Learning Technique performed really well. The performance metrics proves it as it has a R Square of 73.13; greater than the above implemented models.
- **Gradient Boosting Regression Model**: Ideally, this model is assumed to have a better performance than the rest as it is also based on using Ensemble Learning. Gradient Boost performance was recorded to have a R-Square of 75.22. The reason of increased score is the use of Boosting technique which make multiple weak learners into strong learners.

The Table 3 depicts the performance of the implemented models. By looking at the performances of each model, it can be observed that Ensemble Learning approach gives better prediction of the residential houses than the traditional approaches. However, out of the two Ensemble Learning models, Gradient Boosting Regression achieves the best result.

Table 3: Performance Metrics

Model	**R Square**	**Mean Squared Error**	**Mean Absolute Error**
Multiple Linear Regression	67.83	59511.67	48981.31
K Nearest Neighbour	62.81	63984.20	48245.11
Decision Tree Regressor	70.98	56521.59	43228.32
Random Forest Regressor	73.31	54199.99	42001.01
Gradient Boosting Regressor	75.22	52197.72	40700.01

To validate the performance of each model, the actual and predicted prices were calculated on the testing test. The detailed plot is attached to the code artifact. Table 4 is drawn here to showcase the predicted values of residential houses. The predicted value for each model is close to the actual value, but out of all five, Gradient Boost gives a near-to-actual value, i.e., for a residential property priced at 430000 Euro, it predicts the price to be 402268.72 Euro. Therefore, the evidence concludes that Gradient Boosting Regression would be the best technique to determine the prices of residential houses in Dublin.

Table 4: Actual Vs Predicted Prices

Model	**Actual Value**	**Predicted Value**
Multiple Linear Regression	161000.00	172057.033
K Nearest Neighbour	190000.00	128942.73
Decision Tree Regressor	215000.00	208785.94
Random Forest Regressor	430000.00	343626.68
Gradient Boosting Regressor	430000.00	402268.72

7 Conclusion and Future Work

The research aims to find the best strategy for estimating the values of residential properties in Dublin, and it turns out that Gradient Boost is the best method. The contribution to the study was not just limited to constructing a model but also to improving the dataset's properties and investigating the nature of the Dublin Housing Market. The majority of the effort has gone into elaborating each characteristic and obtaining additional information. It was challenging to collect data from 5,06,020 records with restricted attributes and extract them. During the process of Exploratory Data Analysis, this research presented several additional elements such as Property Number, Street, Town, Price Level, Location, and geo coordinates for each house. The highlight of the study has been the progressive improvement in performance by addressing the inadequacies of the existing ones. The final finding of this study shows that the best strategy for predicting on changed datasets is Gradient Boost Regressor, which has an R-Square of 75.22.

The work may be improved further by deploying a model adopting the best method, which would aid in research and provide home buyers with a platform on which to verify pricing when acquiring a property. Furthermore, since the data is held by the Price Regulatory Services Authority of Ireland, the visualizations created might be valuable to the Advisory in assembling its yearly reports. As data science is vast, improving the provided work is limitless. However, establishing an end-to-end residential property price prediction system will significantly contribute to the research.

References

Agnew, K. and Lyons, R. C. (2018). The impact of employment on housing prices: Detailed evidence from fdi in ireland, *Regional Science and Urban Economics* **70**: 174–189.

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

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.

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

Gupta, R., J, M. and bajaj, N. (2020). 2nd international conference on innovative mechanisms for industry applications (icimia 2020) : conference proceedings : 5-7 march, 2020.

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

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

Limsombunc, V., Gan, C. and Lee, M. (2004). House price prediction: Hedonic price model vs. artificial neural network, *American Journal of Applied Sciences* **1**: 193–201.

Lu, S., Qin, Z. and Yang, X. (2019). 2019 ieee symposium series on computational intelligence : Ieee ssci 2019 : December 6-9, 2019, xiamen, china.

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

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.

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

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.

Moro, M., Mayor, K., Lyons, S. and Tol, R. S. (2013). Does the housing market reflect cultural heritage? a case study of greater dublin, *Environment and Planning A* **45**: 2884–2903.

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.

Roche, M. J. (2001). The rise in house prices in dublin: bubble, fad or just fundamentals.

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.

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

Tran, D., Shetty, G. and Chao, Y. (2017). Ieee ieem2017 : 2017 ieee international conference on industrial engineering engineering management : 10-13 dec, singapore, **8**.

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

Zhang, D., Ji, Q., Zhao, W. L. and Horsewood, N. J. (2021). Regional housing price dependency in the uk: A dynamic network approach, *Urban Studies* **58**: 1014–1031.

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.