A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Student Full Name* | Santhosh Shanmugasundaram |
| *Student Number* | Sba24100 |
| *Module Title* | Strategic Thinking |
| *Assessment Title* | Rent Predictor: A Machine Learning Approach to Forecast Dublin Home Rent |
| *Assessment Due Date* | 18th May 2025 |
| *Date of Submission* | 18th May 2025 |

**Declaration**

By submitting this assessment, I confirm that I have read the CCT policy on academic misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source.

I declare it to be my own work and that all material from third parties has been appropriately referenced.

I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Rent Predictor: A Machine Learning Approach to Forecast Dublin Home Rent

Contents

[1 Introduction 1](#_Toc198722924)

[2 Problem Definition 1](#_Toc198722925)

[3 Project Scope 1](#_Toc198722926)

[3.1 Business Understanding 2](#_Toc198722927)

[3.2 Data Understanding 3](#_Toc198722928)

[3.3 Data Preparation 3](#_Toc198722929)

[3.4 Modeling 4](#_Toc198722930)

[3.5 Evaluation 5](#_Toc198722931)

[3.6 Deployment 5](#_Toc198722932)

[4 Details of Dataset 6](#_Toc198722933)

[5 Key Challenges 6](#_Toc198722934)

[6 Machine Learning Algorithms 7](#_Toc198722935)

[6.1 Random Forest Regression 7](#_Toc198722936)

[6.2 Linear Regression 9](#_Toc198722937)

[6.3 ARIMA 10](#_Toc198722938)

[7 ML Techniques 10](#_Toc198722939)

[7.1 Cross Validation 11](#_Toc198722940)

[7.2 Mean Square Error (MSE) & R2 11](#_Toc198722941)

[7.3 Hyperparameters 11](#_Toc198722942)

[7.4 GridSearchCV 12](#_Toc198722943)

[8 Model Implementation 12](#_Toc198722944)

[8.1 Exploratory Data Analysis (EDA) 12](#_Toc198722945)

[9 References 18](#_Toc198722946)

[10 GitHub Link 18](#_Toc198722947)

# Introduction

Rental homes are an integral part of the housing ecosystem, providing an alternative to homeownership. A properly functioning housing market is an essential ingredient of a properly functioning economy and society. (IPOA and IPAV)

From an economic, social and political perspective, housing is currently the greatest challenge faced by Irish policymakers. The housing challenge is manifested in a lack of supply of owner-occupier and rental properties; prohibitively high house prices and rents; and a serious problem of homelessness. (IPOA and IPAV)

# Problem Definition

The inability to provide an adequate supply of suitable and affordable housing supply for those who want to rent has very negative economic and social consequences. These consequences include:

* High and rising rents take spending power out of the economy and render it very difficult for aspiring house buyers to build up a sufficient deposit.
* High and rising house rents put upward pressure on wages, and this undermines national competitiveness.
* The availability of an abundant supply of high-quality housing to rent or purchase at affordable prices is a necessary condition for labour mobility within a country and between countries. For Ireland, inward migration is an essential part of the economic model, and housing can act as a major impediment to such labour flows.  
  (IPOA and IPAV)

# Project Scope

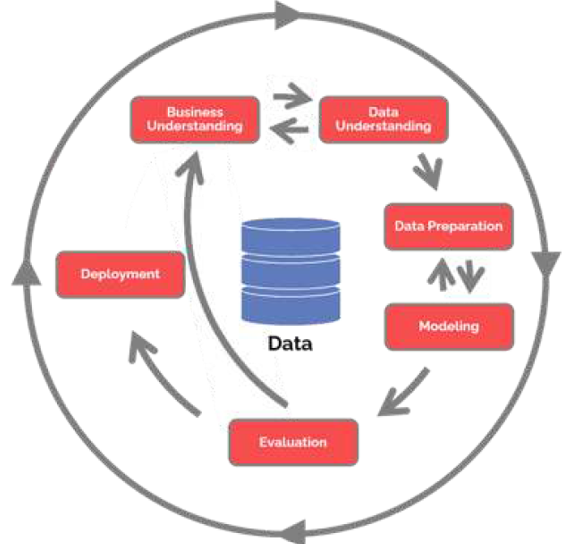
The scope of this project is to understand the **Project Management** strategy using the average price to rent a home at Dublin County during every quarter for three different types of homes namely apartment, terrace house and semi-detached house. The target of this study is to help the renters to decide on the future rent prices in Dublin so that expenses are planned.

Use the AI technology along with the Machine Learning models like ARIMA, Linear Regression and Random Forest algorithm to predict the future rent at Dublin based on the **CRoss Industry Standard Process for Data Mining** (CRISP-DM).

This is a process model that serves as the base for a data science process that includes 6 phases as below.

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modelling
5. Evaluation
6. Deployment

(Hotz)



## Business Understanding

Any good project starts with a deep understanding of the customer’s needs. Data mining projects are no exception and CRISP-DM recognizes this.

The Business Understanding phase focuses on understanding the objectives and requirements of the project. Aside from the third task, the three other tasks in this phase are foundational project management activities that are universal to most projects:

1. **Determine business objectives:** You should first “thoroughly understand, from a business perspective, what the customer really wants to accomplish.” ([CRISP-DM Guide](https://web.archive.org/web/20220401041957/https:/www.the-modeling-agency.com/crisp-dm.pdf)) and then define business success criteria.
2. **Assess situation:** Determine resources availability, project requirements, assess risks and contingencies, and conduct a cost-benefit analysis.
3. **Determine data mining goals:** In addition to defining the business objectives, you should also define what success looks like from a technical data mining perspective.
4. **Produce project plan:** Select technologies and tools and define detailed plans for each project phase.

While many teams hurry through this phase, establishing a strong business understanding is like building the foundation of a house – essential.

## Data Understanding

 Next is the *Data Understandin*g phase. Adding to the foundation of *Business Understanding*, it drives the focus to identify, collect, and analyze the data sets that can help you accomplish the project goals. This phase also has four tasks:

1. **Collect initial data:** Acquire the necessary data and (if necessary) load it into your analysis tool.
2. **Describe data:** Examine the data and document its surface properties like data format, number of records, or field identities.
3. **Explore data:** Dig deeper into the data. Query it, visualize it, and identify relationships among the data.
4. **Verify data quality:** How clean/dirty is the data? Document any quality issues.

## Data Preparation

A common rule of thumb is that 80% of the project is data preparation.

This phase, which is often referred to as “data munging”, prepares the final data set(s) for modeling. It has five tasks:

1. **Select data:** Determine which data sets will be used and document reasons for inclusion/exclusion.
2. **Clean data:** Often this is the lengthiest task. Without it, you’ll likely fall victim to garbage-in, garbage-out. A common practice during this task is to correct, impute, or remove erroneous values.
3. **Construct data:** Derive new attributes that will be helpful. For example, derive someone’s body mass index from height and weight fields.
4. **Integrate data:** Create new data sets by combining data from multiple sources.
5. **Format data:** Re-format data as necessary. For example, you might convert string values that store numbers to numeric values so that you can perform mathematical operations. (Data Science PM)

## Modeling

What is widely regarded as data science’s most exciting work is also often the shortest phase of the project. Here you’ll likely build and assess various models based on several different modeling techniques. This phase has four tasks:

1. **Select modeling techniques:** Determine which algorithms to try (e.g. regression, neural net).
2. **Generate test design:** Pending your modeling approach, you might need to split the data into training, test, and validation sets.
3. **Build model:** As glamorous as this might sound, this might just be executing a few lines of code like “reg = LinearRegression().fit(X, y)”.
4. **Assess model:** Generally, multiple models are competing against each other, and the data scientist needs to interpret the model results based on domain knowledge, the pre-defined success criteria, and the test design.

Although the [CRISP-DM Guide](https://web.archive.org/web/20220401041957/https:/www.the-modeling-agency.com/crisp-dm.pdf) suggests to “iterate model building and assessment until you strongly believe that you have found the best model(s)”, in practice teams should continue iterating until they find a “good enough” model, proceed through the CRISP-DM lifecycle, then further improve the model in future iterations.

## Evaluation

Whereas the *Assess Model* task of the *Modeling* phase focuses on technical model assessment, the *Evaluation* phase looks more broadly at which model best meets the business and what to do next. This phase has three tasks:

1. **Evaluate results:** Do the models meet the business success criteria? Which one(s) should we approve for the business?
2. **Review process:** Review the work accomplished. Was anything overlooked? Were all steps properly executed? Summarize findings and correct anything if needed.
3. **Determine next steps:** Based on the previous three tasks, determine whether to proceed to deployment, iterate further, or initiate new projects. (Data Science PM)

## Deployment

Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.

A model is not particularly useful unless the customer can access its results. The complexity of this phase varies widely. This final phase has four tasks:

1. **Plan deployment:** Develop and document a plan for deploying the model
2. **Plan monitoring and maintenance:** Develop a thorough monitoring and maintenance plan to avoid issues during the operational phase (or post-project phase) of a model
3. **Produce final report:** The project team documents a summary of the project which might include a final presentation of data mining results.
4. **Review project:** Conduct a project retrospective about what went well, what could have been better, and how to improve in the future.

Your organization’s work might not end there. As a project framework, CRISP-DM does not outline what to do after the project (also known as “operations”). But if the model is going to production, be sure you maintain the model in production. Constant monitoring and occasional model tuning is often required. (Data Science PM)

# Details of Dataset

Residential Tenancies Board is an independent public body that regulates the private rented sector. It is responsible for multiple activities like tenancies registration, dispute resolution, investigation and sanction of landlords. It has also published many numbers of reports and data. (Citizensinformation.ie)

For the specific case study, I will use the dataset available in the below link.

<https://data.cso.ie/table/RIQ02>

This dataset contains the average rental price at Dublin city based on number of bedrooms (1, 2, 3) and type of property (semi-detached house, terrace house, apartment) starting from fourth quarter 2020 till first quarter 2024.

Below are the columns in the datasheet

1. STATISTIC Label – Description of the dataset
2. Quarter – Time period indicating the data time
3. Number of Bedrooms – Indicate the number of bedrooms in the house. It has 3 values namely – One bed, two bed and three bed.
4. Property type – 3 possible values namely, semi-detached house, Terrace house and Apartment
5. Location – In this dataset, this will contain only ‘Dublin’
6. Unit – Indicates the currency and this will contain only ‘Euro’
7. Value – Numerical value indicating the rent amount.

# Key Challenges

Collecting valid, meaningful and up-to-date data from a trusted source, poses a significant challenge. Moreover, selecting the right dataset for analysis with correct filters and formatting the data before consumption takes significant amount of time.   
  
Exploratory Data Analysis (EDA) has been performed to make sure there is no duplicates entries nor null values is present. Also, scaling the data by eliminating unnecessary columns also had to be done.   
  
Hyperparameters in machine learning are extremely crucial and significant because they greatly affect the performance and efficiency of the selected model. I faced several typical difficulties with hyperparameters, which are:

Incorrect tuning of the parameters can result in either overfitting or underfitting. There are numerous hyperparameters for machine learning models, and selecting the appropriate one with the optimal configuration is quite challenging. Sometimes the models take longer to run during training, which complicates their use with larger datasets.

This dataset contains the average rental price at Dublin city based on number of bedrooms (1, 2, 3) and type of property (semi-detached house, terrace house, apartment) starting from fourth quarter 2020 till first quarter 2024.

# Machine Learning Algorithms

Three different types of algorithms were used to evaluate and predict the data namely,

1. Random Forest Regression
2. Linear Regression
3. ARIMA

## Random Forest Regression

Random forest is a supervised learning algorithm that uses an ensemble of decision trees to predict outcomes. It operates by creating multiple decision trees from random subsets of the data using a technique called bagging.

They are widely used for classification and regression task. It is a type of classifier that uses many decision trees to make predictions. Each tree makes its own prediction, and the model aggregates these predictions to generate a result.

In regression tasks, random forest predicts continuous target variables, reducing variance and improving accuracy by combining the outputs of several trees. This approach allows it to handle larger datasets and capture more complex relationships than individual decision trees.

Imagine asking a group of friends for advice on where to go for vacation. Each friend gives their recommendation based on their unique perspective and preferences (decision trees trained on different subsets of data). You then make your final decision by considering the majority opinion or averaging their suggestions (ensemble prediction). (GeeksforGeeks, “Random Forest Algorithm in Machine Learning”)

A diagram of a tree

AI-generated content may be incorrect.

Process starts with a dataset with rows and their corresponding class labels (columns).

* Then - Multiple Decision Trees are created from the training data. Each tree is trained on a random subset of the data (with replacement) and a random subset of features. This process is known as **bagging** or **bootstrap aggregating**.
* Each Decision Tree in the ensemble learns to make predictions independently.
* When presented with a new, unseen instance, each Decision Tree in the ensemble makes a prediction.

The final prediction is made by combining the predictions of all the Decision Trees. This is typically done through a majority vote (for classification) or averaging (for regression).

## Linear Regression

**Linear Regression** helps to find the relationship between two features (input variable/independent variable and target variable/dependent variable). Linear Regression is a supervised learning algorithm which is used for predicting a continuous target variable. In this model, best fit linear equation is used to reduce the chances of error between the predicted values and the actual target values.

It uses a linear relationship between the independent variables and the dependent variable. The main objective of the Linear Regression is to find the best fitting line that predicts the actual target variable based on the predicted values.

Linear Regression model can be represented as below:

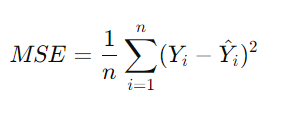
Y = β0 + β1X + ϵ

where β0​ in the above equation is the intercept, β1​ in the above equation represents the slope (coefficient) and ϵ represents the error term which needs to be as minimal as possible.

We need to make certain assumptions which are listed below:

* Linearity: Relationship between the X and the Y values should be linear.
* Independence: Observations should be independent.
* Homoscedasticity: Constant variance of errors.
* Normality: Errors should be normally distributed.

Linear Regression model finds the line that minimizes the Mean Squared Error (MSE):



## ARIMA

**ARIMA** (AutoRegressive Integrated Moving Average) is powerful time series forecasting model which is used to investigate and predict time series data. It is very useful for data that exhibits trends and forecasting. It combines the components of autoregression (AR), differencing (I), and moving average (MA) into a single model.

* Autoregressive (AR) part: Refers to the relationship between the observation and the previous observations. The Autoregressive order is denoted as p and this indicates the number of lagged values considered.
* Differencing (I): This helps to transform a non-stationary series into a stationary one by calculating the difference. For example, this will subtract the previous observation from the current one. The order of differencing is d and this indicates how many times the difference is applied.
* Moving Average (MA) part: Refers to the dependency between an observation and a residual error from a moving average model. The MA order is denoted as q.

To summarise, the ARIMA model is represented by p, d and q where:

* p: Order of autoregressive terms.
* d: Order of differencing.
* Q: Order of moving average terms.

The model predicts the future value of the time series by considering both past values (AR) and past forecast errors (MA), adjusted for non-stationarity through differencing (I).

While comparing each of these models serves distinct purposes, tailored to specific types of data and goals: Linear Regression for continuous predictions, Logistic Regression for binary classification, and ARIMA for time series forecasting.

# ML Techniques

A successful implementation of machine learning algorithm involves usage of multiple techniques like cross validation, hyperparameters, Grid search CV, Ridge Regression, Lasso Regression.

Similarly multiple validation methods are used to evaluate the model performance. Some of which are Mean Square Error (MSE), Mean Absolute percentage and R2.

## Cross Validation

Cross validation is a technique used to avoid overfitting scenario by training the model with multiple subsets of train-test data. To achieve this, k-folds is used which is a numeric value that is passed as input and based on which actual dataset is split into multiple smaller units and model is trained and tested with each unit. Using this approach, model is trained with all the possible data and tested multiple times. The average of all the result is taken as final measure.

(Simplilearn)

## Mean Square Error (MSE) & R2

Mean Square Error (MSE) is the measure of average squared difference between actual value and predicted value. Smaller the value, better the prediction. In simple terms, smaller the number, smaller the error.

R2 is a statistical measure that indicates proportion of the variance with the target. R2 value ranges from 0 to 1, where 1 being perfect prediction and 0 being poor prediction. If R2 value is less than 0, then model is very poorly predicting. (Rowe)

## Hyperparameters

Hyperparameters are nothing but configurational parameters that are provided for algorithms. Each algorithm has their own configuration parameters that control the behaviour and performance of the model.   
For example, the configurational parameters pertaining to Random Forest Regression algorithm are – n\_estimators, max\_depth, min\_sample\_split, max\_features, max\_leaf\_node, max\_samples, etc.

Each and all parameters have a specific setting that controls how the algorithm is executed.

Thus, finding the optimal hyperparameters would help us achieve the best-performing model. There are several techniques for choosing a model’s hyperparameters, including Random Search, GridSearchCV, Manual Search, and Bayesian Optimization. Among these, GridSearchCV is widely recognized for its efficiency in tuning parameters.

## GridSearchCV

GridSearchCV (Grid Search Cross-Validation) is a technique used in machine learning to search and find the optimal combination of hyperparameters for a given model. We provide GridSearchCV with a set of Scoring parameter to experiment with, and it systematically explores the hyperparameter values, creating a “grid” of possible combinations. For every combination, it evaluates the model’s performance by testing it on various sections of the dataset to gauge its accuracy. This technique replaces the manual trial and error approach.

After exhaustively trying out all the combinations, GridSearchCV returns the combination of settings that yielded the most favourable outcomes.

Since GridSearchCV tries all the possible combination, the execution time is high, and it could be higher as the complexity increases. To utilize the parallelism, GridSearchCV provides a parameter namely n\_jobs. Setting a value of -1 means the algorithm can use all the processors at once to complete the task.

(Shah)

# Model Implementation

With the dataset containing the average rental price at Dublin city, Rental price of Dublin homes has been predicted using 3 different machine learning algorithms namely – Random Forest Regression, Linear regression and ARIMA.

## Exploratory Data Analysis (EDA)

The relevant data is downloaded in the ‘CSV’ file and uploaded to the python programming module using library named ‘PANDAS’.

Once the data is loaded, the same can be displayed as-is using data frame function namely ‘HEAD’.

A screenshot of a computer

AI-generated content may be incorrect.

As part of the EDA, we examine the data and document its surface properties like data format, number of records etc.

To check the number of rows and columns in the dataset, we use the data frame built in parameter ‘shape’ and to display the format of each data, we use the data frame function ‘info’.

A screenshot of a computer

AI-generated content may be incorrect.

It is important to understand and inspect the data quality. To do that, we need to check the unique values in each of the column and check for any missing or duplicate records in the dataset.

A screenshot of a computer code

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

The above screen shot taken from jupyter notebook shows all the unique values corresponding to each column in the datasheet. This provides insight into what kind of data to expect in the outcome.

A white rectangular box with red and blue text

AI-generated content may be incorrect.

We must make sure that there are no duplicated entries in the dataset because this affects the model results. Thereby, we check for it and delete if any exists.

A dataset may contain lot of information that may not help a machine learning algorithm in predicting the result. Therefore, it is important to find the columns that constitute no meaningful information and remove it. As a rule of thumb, any column that contains just 1 value for all records adds no value to the modelling. Therefore, such columns can be removed using data frame function ‘drop’.

A screenshot of a computer

AI-generated content may be incorrect.

For the dataset used in this project, I have removed 3 columns that has no real use for machine learning and the final columns that are relevant has been displayed and shown the screen shot above.

It is very common for a dataset to contain information in strings and text. But for a machine learning model to work, it is very important to provide all the information in numeric format. As a common approach, we replace the unique text values into a simple meaningful number and same has been done as shown in the below screen shot.

A screenshot of a computer

AI-generated content may be incorrect.

In this project, I have converted the number of bedrooms and property type into simple 0 to 3 number and same has been displayed above.

From now on, above naming conventions apply to the values in column ‘number of bedrooms’ and ‘property type’.

As part of data understanding phase, it is critical to look deeper into data and draw a relationship between each column. To perform this, we derive the correlation factor associated with each column and find the most important data controls the result.

A screenshot of a diagram

AI-generated content may be incorrect.

Before we start on building a machine learning model, it is important to get a holistic view of the complete dataset and to do that we draw plots using library ‘matplotlib’.

For this project, I have created an individual plot based on the type of property and each plot will contain 3 graphs corresponding to bedroom numbers 1, 2 and 3. The X-axis corresponds to quarterly timeline and y-axis corresponds to the rental amount in Euro.

A graph of a rental cost

AI-generated content may be incorrect.

A graph of a rental cost

AI-generated content may be incorrect.

A graph of a rental cost

AI-generated content may be incorrect.

# References

(IPOA and IPAV)

IPOA, and IPAV. *THE IRISH PRIVATE RENTAL MARKET*. IPAV.ie, June 2022.

Available at: <https://www.ipav.ie/sites/default/files/ipav_ipoa_jim_power_updated_report_june_2022.pdf>

(Hotz)

Hotz, Nick. “What Is CRISP DM?” *Data Science Project Management*, 2024, www.datascience-pm.com/crisp-dm-2/.

# GitHub Link

<https://github.com/santhosh-sba24100/SBA24100_CA3_Strategy_Thinking>