Contents

[Abstract 2](#_Toc159281950)

[Introduction 2](#_Toc159281951)

[Research motivation 4](#_Toc159281952)

[Motivation 4](#_Toc159281953)

[Objectives 4](#_Toc159281954)

[Literature review 4](#_Toc159281955)

[An overview of models for predicting Airbnb accommodation prices 6](#_Toc159281956)

[Linear regression 6](#_Toc159281957)

[Ensemble learning 7](#_Toc159281958)

[Neural Network 9](#_Toc159281959)

[Multi-modality data to predict rent prices 10](#_Toc159281960)

[Environmental noise and air quality on accommodation price 10](#_Toc159281961)

[Room prices and points of interest 11](#_Toc159281962)

[Conclusions 12](#_Toc159281963)

[Methodology 12](#_Toc159281964)

[Primary research 13](#_Toc159281965)

[Ethics 14](#_Toc159281966)

[GDPR 15](#_Toc159281967)

[Research methodology 15](#_Toc159281968)

[Literature Review 16](#_Toc159281969)

[Business Understanding 16](#_Toc159281970)

[Data Understanding 16](#_Toc159281971)

[Data Preparation 17](#_Toc159281972)

[Modelling 17](#_Toc159281973)

[Linear regression 17](#_Toc159281974)

[Support vector machine regressor 18](#_Toc159281975)

[Decision Tree 19](#_Toc159281976)

[Random Forest 21](#_Toc159281977)

[Feedforward neural network 22](#_Toc159281978)

[Evaluation 23](#_Toc159281979)

[Implementation 23](#_Toc159281980)

[Primary data 23](#_Toc159281981)

[Group one: Noise and Air quality monitoring experts 24](#_Toc159281982)

[Group two: property and tourism experts 24](#_Toc159281983)

[Key decisions record 25](#_Toc159281984)

[Data Architecture 25](#_Toc159281985)

[Datasets 26](#_Toc159281986)

[Data understanding 27](#_Toc159281987)

[Sensor data 27](#_Toc159281988)

[Airbnb Data 28](#_Toc159281989)

[Data preparation 29](#_Toc159281990)

[Exploratory Data analysis 31](#_Toc159281991)

[Model Results 36](#_Toc159281992)

[Time series analysis 36](#_Toc159281993)

[Regression Results 41](#_Toc159281994)

[Linear regression (R2 0.58) 41](#_Toc159281995)

[Support vector regressor (R2 0.59) 42](#_Toc159281996)

[Decision tree (R2 0.51) 42](#_Toc159281997)

[Random Forest (R2 0.70) 43](#_Toc159281998)

[Neural network (R2 0.79) 43](#_Toc159281999)

[Conclusions 43](#_Toc159282000)

[Limitations 43](#_Toc159282001)

[Future steps 43](#_Toc159282002)

[References 43](#_Toc159282003)

[Annex 43](#_Toc159282004)

# Abstract

# Introduction

One of the key aspects of the modern urban landscape is the escalating challenge posed by rapid urbanization which calls for innovative solutions to improve the overall quality of life for city inhabitants; living conditions are subjected to many factors, in this context city sensors play a pivotal role in measuring and predicting multiple aspects of living conditions.

The same is particularly true with smart cities: smart cities are characterized by the integration of technologies to optimize urban functions; therefore, sensors are at the hearth of smart cities development: air quality and environmental noise sensors play a crucial role in the monitoring of urban environments as they provide real-time data which is essential for informed decision-making and policy formulation.

One of the main applications for sensors data relies on public health, with real time capabilities is possible to predict, assess and tackling pollution and develop efficient interventions for mitigating risk factors, air quality data are mainly used for this purpose as excellent indicators.

Collecting noise data is also a useful tool for safeguarding noise pollution level, by deepening the understanding or urban noise, policy makers can implement different traffic management strategies improving the overall well-being of their citizens.

Sustainability is another important key aspect of the use of sensors; continuous monitoring is one of the most important aspects for measuring air quality and noise level indices and maps contributing to an holistic understanding of the urban environment, overall sensors data can contribute to evidence-based decision making guiding policy makers.

When it comes to urban planning, sensor data can facilitate the design of infrastructures and new developments, as an example data from these sources are essential for define layouts, highlights possible challenges and optimizations of commercial and urban zones: identifying pollution sources and making sure urban populations will not be exposed to pollutants.

When it comes to the impact of sensor data in economics understanding of urban areas, sensor data is capable of capturing the influence of key factors in many aspects of the city living: sensors are particularly useful in the realm of real estate, the tracking of environmental factors is hugely important for measuring and predicting property prices as it’s easier to imagine how noise and air quality levels discrepancies in the same urban areas can determine differences in prices, in many cases properties in quiet areas can correlate to higher prices as noise can indicate more attractive areas for families to buy a house in.

Prediction of rents prices can also avail the help of sensor data, correlated to property prices sensor data is a useable information to predict rent prices, they can also be use for identifying popular areas by using noise, accurate and up-to-date sensor data can provide a nuanced understanding of the desirability of different neighbourhoods based on environmental quality; for example, some areas with consistently low pollution and minimal noise disturbances may be perceived as more attractive but also high levels of noise during weekends can detect the most vibrant parts of city night life which might be attractive for the opposite reason.

All this reason can be summarised in property evaluation models which can be used from investors and property agents along side traditional factors for prices such as location and amenities.

In summary the utilization of noise and air quality sensors in smart cities extends beyond mere data collection: the integration of noise and air quality sensors into smart cities has implications for the economic dynamics of urban living. By incorporating environmental data into property valuation and rent-setting processes, cities can create more responsive and fair real estate markets, ultimately contributing to a more sustainable, liveable, and economically vibrant urban environment.

# Research motivation

## Motivation

The primary motivation behind this research stems from the fundamental question: Does the environment significantly influence Airbnb prices? While existing knowledge indicates that factors like proximity to attractions and ambient noise affect residential property prices, the impact on short-term touristic rentals remains uncertain.

## Objectives

The question that leads to this research is as follows: Does the environment affect Airbnb prices? It is known that, for residential properties, distance from places of interest and noise level can affect the price; is the same true for touristic short-term rentals; and what other factors can affect the price of short-term rentals; this research is set to answer this question by accomplishing the objectives below:

* To evaluate the impact of various features on Airbnb prices (property related features, host related features, noise level and air pollution levels, distance from different points of interest).
* To evaluate if a neural network can outperform best performing models used in previous literature (random forest and XGBoost).
* To contribute to predict the price of Airbnb apartments for Dublin city.
* To predict noise level using time series analysis on sensor data to help modelling for future prices.

# Literature review

In recent years, the landscape of the sharing economy has undergone substantial growth, propelled by platforms like Airbnb that have revolutionized how people access accommodations. Within the sharing economy, individuals collaborate by sharing resources, goods, or services through online platforms or communities, optimizing the use of assets that might otherwise be underutilized. Airbnb, a key player reshaping the accommodation sector, introduces a noteworthy challenge due to the dynamic pricing nature. Unlike traditional lodging options, Airbnb prices can fluctuate based on factors such as location, demand, seasonal variations, and unique listing attributes. This inherent unpredictability poses a specific dilemma for hosts and guests as they seek to strike a balance between cost-effectiveness and value.

To grapple with this challenge, the integration of data science has assumed increasing importance. Researchers can leverage the wealth of historical data available on Airbnb to extract insightful information on pricing trends, customer preferences, and market dynamics. These insights serve as the groundwork for constructing accurate models capable of predicting Airbnb listing prices.

The accurate prediction of Airbnb listing prices carries significant implications for both hosts and guests immersed in the sharing economy. For hosts, precise price predictions empower them to fine-tune their listings by setting competitive prices that align with market dynamics and property attributes. This, in turn, maximizes their revenue potential and elevates overall business performance. Moreover, it furnishes hosts with insights into the factors influencing price fluctuations, aiding informed decisions regarding property upgrades, amenities, and other listing characteristics.

Conversely, accurate price predictions benefit guests navigating the sharing economy by fostering transparency and enabling them to make well-informed decisions based on their budget, preferences, and desired amenities. Within the sharing economy framework, where resources are shared among individuals, precise price predictions enable guests to select the most fitting accommodation options and plan their trips with efficiency. This transparency heightens their overall satisfaction and enriches their experience of engaging in the sharing economy through the Airbnb platform.

While predicting prices of accommodation has certainly been a well-researched area there still is room for improvement, as mentioned sharing economy prices are subject to fluctuations from different factors highlighting the need for well rounded approaches that cannot simply rely on accommodation-based modelling.

This research will focus on the concept of using different features mostly related to geolocation of the accommodations to create a reliable and accurate price prediction strategy.

## An overview of models for predicting Airbnb accommodation prices

Many papers have explored the prediction of room prices for both long- and short-term accommodation; many models were chosen for this objective.

### Linear regression

In this context linear regression can be used for modelling the relationship between a dependent variable and one independent variable by fitting a linear equation to the observed data. The goal of linear regression is to find the best-fitting line (or hyperplane in the case of multiple independent variables) that minimizes the differences between the observed and predicted values.

Linear regression offers many advantages: it can perform better than other models when the relationship between the variable follows a linear pattern, it is easy to compute and easier to explain than more complex algorithms.

However, it needs to be handled with caution as linear regression is susceptible to overfitting; also, noise and outliers can dramatically reduce model’s performance, when handling linear regression, it is important to keep in consideration how to deal with multicollinearity (a state when independent features are correlated within themselves).

When it comes to utilizing linear regression for predicting accommodation prices in tourism or real estate several authors already approached the subject.

Yu et al. (2018) in a paper called "Real Estate Price Prediction with Regression and Classification" [1] developed a set of models and used linear regression as benchmark for forecasting the ideal sale price of a home by using a dataset of residential houses sold in Ames, Iowa from 2006 to 2010; the datasets consisted of 79 features and more than 1000 houses, the authors used Root mean squared error (RMSE) as a commonly used performance measure for evaluation of price predictions, overall the paper reached a score of 0.55 by using simple linear regression with no regularization.

To reduce loss the paper also applied two regularization methods: Lasso and Ridge: both regularizations help prevent overfitting enhancing the generalization of the model, Lasso works by adding a penalty term to the function discovered by the regressor therefore avoiding overfitting.

Lasso regression often shrinks down the coefficient of a particular variable down to zero therefore implying a feature selection technique were only the relevant features have a coefficient that is different from zero, Ridge regression also similarly to Lasso adds a penalty term and shrinks down the coefficient but it never reaches exactly zero.

Yu et al. Reached a similar RMSE score with ridge and lasso regression of 0.54 compared to 0.55 on simple linear regression, in another paper Chapman et al. “PREDICTING LISTING PRICES IN DYNAMIC SHORT TERM RENTAL MARKETS USING MACHINE LEARNING MODELS” [2] found similar scores using Lasso and ridge regression: the performances of the two models were extremely close regardless of regularization techniques; they found a R-squared of around 0.52 for Ridge and 0.51 for Lasso attributing the small difference to the ridge regularization factor handling better the multicollinearity of the features.

The paper also explored, among others, Elastic Net algorithm: Elastic Net can combine both L1 and L2 regularization techniques; the methodology of the paper revolved around comparing the results for each algorithm on two datasets, one with and another without feature selection. Elastic Net performed quiet similarly to Lasso regression for a standard dataset; however, it performed worse on the dataset when feature selection was applied, the authors identified Elastic Net as unnecessary complex leading to slightly worse performance than Lasso and Ridge.

Another paper by Ziyue Huang “Logistic Regression in Rental Price and Room Type Prediction Based on Airbnb Open Dataset” [3] confirmed previous results with very similar success using Lasso regression and hyper parameter turning showing a positive correlation between the accuracy or the model and the regularization parameter.

Overall, the literature seems to do frequent use of linear regression techniques, however, the algorithm is mostly used as a benchmark to test other algorithms which proven more effective than linear regression.

### Ensemble learning

When it comes to price predictions in accommodation random forest and other ensemble learning models are one of the most used; ensemble learning can be explained as a technique which takes in consideration various model and their prediction and combines them into a more robust prediction by compensating the weakness of a single prediction improving the overall performance.

Ensemble learning utilizes different methods to achieve an accurate prediction, the two main approaches are bagging and boosting: bagging is the process of train the same type of models on different batches of the same dataset, each model is trained independently and then the output of the models are combines into a single prediction; boosting instead works by developing models in a sequential fashion and compensating the mistakes made by the previous model by putting emphasis on the errors made by the previous model, again the prediction are then combined at the end for a final prediction.

Overall ensemble learning models results more generalizable than traditional models for the aforementioned reasons, over the years multiple papers have explored the use of ensemble learning models of the price prediction of accommodation.

Hu et al. in “Prediction and Analysis of Rental Price using Random Forest Machine Learning Technique” [4] explored the impact of different features when predicting the prices of rents for two major cities in China Wuhan and Shanghai, the dataset used was comprehensive of multiple features regarding location, amenities and quality of the accommodation.

The result of the paper indicates a good R-squared score, the model worked especially well for Wuhan compared to Shanghai, the researchers stressed the importance of selecting the correct features such as city district which was found as particularly important.

Adetunjia et al. in “House Price Prediction using Random Forest Machine Learning Technique” [5] also implemented a random forest model with good results, the data was collected from a dataset of 500 homes in Boston collected in 1978 with 14 features such as location, closeness to the main river, number of rooms.

The methodology applied consisted of using bootstrap as sampling method and 500 trees in total, as for previous literature the R-squared score was set at 0.9 showcasing a high degree of accuracy.

XGBoost, or extreme Gradient Boosting, stands out as a potent ensemble machine learning algorithm, widely recognized for its outstanding predictive capabilities. Functioning as a boosting algorithm, XGBoost amalgamates the outcomes of numerous weak learners, typically decision trees, resulting in a robust and precise model. It systematically constructs trees with a focus on rectifying errors from previous trees. XGBoost uses regularization techniques to prevent overfitting and Its efficiency is further elevated through parallel processing and hardware optimization.

In the already mentioned paper from Chapman et al. “PREDICTING LISTING PRICES IN DYNAMIC SHORT TERM RENTAL MARKETS USING MACHINE LEARNING MODELS” XGBoost even if comparing better than other models it falls short performing slightly worst than a random forest regressor used on the same dataset; the authors outlined how the iterative nature of the model made the predictions susceptible to noise and outliers reducing the generalization of the trained model to the testing data, however the ensemble learning regressors used in the paper outperformed linear regression.

### Neural NetworkTop of Form

In recent years neural networks have emerged as powerful tools in predicting financial market prices, revolutionizing the landscape of quantitative analysis and forecasting. These computational models draw inspiration from the human brain, consisting of interconnected nodes that process information and learn patterns from historical data. In the realm of predicting prices, neural networks excel at capturing complex relationships and non-linear dependencies within vast datasets.

The ability of neural networks to adapt and optimize their parameters training makes them particularly well-suited for predicting price movements. By ingesting historical price data and relevant features, neural networks can learn intricate patterns, identify trends, and adapt to changing market conditions. This adaptability is crucial in the dynamic and ever-evolving world of financial markets.

Neural networks are employed in various financial applications, ranging from stock and commodity markets to cryptocurrency and foreign exchange. Their capacity to analyse large datasets and discern subtle patterns enables them to uncover hidden correlations and exploit predictive signals that may elude traditional forecasting methods.

When it comes to price prediction for accommodations neural networks have been deployed with relative success in Airbnb or other rent datasets: Peng, et al. in “Leveraging Multi-Modality Data to Airbnb Price Prediction” [6] used a deep neural network with two hidden layers having nine and ten neurons each with ReLU activation function; the methodology of the paper had them using Principal component analysis and selecting six components as inputs for the algorithm, overall the model compared similarly to XGboost performing better than linear counterparts.

However, in another paper from Ahuja et al. “Predicting Airbnb Rental Prices Using Multiple Feature Modalities” [7] by using a similar approach on multi-modality data found how even with no hyper parameter tuning LightGBM and XGboost outperformed both Ridge regression and DNN.

## Multi-modality data to predict rent prices

Previous research has been optimistic on the use of external and multi-modality data to enhance and better predict the prices of an accommodation: Peng, et al. in “Leveraging Multi-Modality Data to Airbnb Price Prediction” [6] employed a wide range of data inputs for their predictions demonstrably reaching higher accuracy than single-type data; The findings of this study indicate that customer reviews, house features, and geographical data serve as effective predictive factors for Airbnb rentals.

Ahuja et al. followed a similar methodology in “Predicting Airbnb Rental Prices Using Multiple Feature Modalities” [7] this work comprehensive preprocessing steps for a rental price dataset, focusing on geospatial, visual, categorical, numerical, and temporal features. The exploration of geospatial features involves reverse-geocoding and hierarchical clustering, emphasizing the significance of neighbourhoods. Visual features, incorporating listing images, were explored using a neural network but were ultimately deemed non-informative for modelling. Categorical features were managed through one-hot and label encoding to prevent sparseness. Numerical features underwent scaling, outlier handling, and a log transformation for normalization. Temporal features were introduced to capture host experience and listing newness. This preprocessing framework serves as a valuable contribution to understanding and modelling rental prices, especially considering its emphasis on diverse feature types.

## Environmental noise and air quality on accommodation price

Many papers highlighted the effect (mostly negative) that environmental noise and air quality has on rent prices, however not much literature has been explored for investigating the impact of noise on tourism accommodation specifically, Kemiki et al. in “THE IMPACT OF NOISE AND DUST LEVEL ON RENTAL PRICE OF RESIDENTIAL TENEMENTS AROUND LAFARGE CEMENT FACTORY IN EWEKORO TOWN, NIGERIA” [8] utilizes a hedonic model to assess the influence of noise and dust levels on the rental prices of residential properties in Ewekoro, Nigeria, with a focus on the impact of a nearby cement factory. The findings indicate that both noise and dust levels significantly affect housing rents, The study suggests that tenement rents decline with increasing distance from the cement factory due to the severity of dust and noise; These findings underscore the importance of addressing environmental factors in housing markets and their broader implications for community well-being and development.

The impact on noise in rent and property prices have been validated all across the globe with different research for many cities: as expressed in “The Influence of Traffic Noise on Apartment Prices on the Example of a European Urban Agglomeration“ by Szczepanska et al [9] “Acoustic discomfort cannot be fully eliminated in urban areas, but noise levels vary in different locations and, consequently, influence property prices.”

When it comes to the relation between property prices and noise in cities most of the research shows that one of the highest contributors for devaluation is in fact traffic noise: different papers evaluated the impact of traffic noise on prices in percentage points which may different from city to city.

Other than traffic noise, previous literature highlighted how other source of noise can have a negative impact on property prices and rents: according to Jun et al. “Noise Pollution Loss Value Evaluation of Railway Transportation Based on Hedonic Price Method- The Case of Taiyuan City” [10] There is a notable adverse correlation between residential unit prices in proximity to railways and the noise generated by railway traffic. The economic impact of traffic noise pollution is quantified at 50.8 yuan/(m2 ·dB), and its influence extends to approximately 500 meters.

### Room prices and points of interest

In the dynamic realm of the tourism industry, the determination of accommodation prices is intricately tied to the surrounding points of interest. These points of interest, ranging from cultural landmarks and natural attractions to vibrant urban centers, serve as key influencers that shape the perceived value of accommodations. The proximity and accessibility of these points of interest often become pivotal factors for travelers when choosing where to stay. Whether it's the allure of iconic landmarks, the tranquility of scenic landscapes, or the vibrancy of local communities, these attractions contribute significantly to the overall experience for tourists. As a result, accommodations strategically positioned in close proximity to points of interest not only enhance the desirability of the stay but also command a premium in pricing. Understanding how points of interest intersect with the pricing dynamics of tourism accommodations is essential for both industry stakeholders and travelers seeking an enriched and tailored experience.

A study conducted in Sanya city from Hu and Song “Analysis of Influencing Factors and Distribution Simulation of Budget Hotel Room Pricing Based on Big Data and Machine Learning from a Spatial Perspective“[8] explained how the spatial landscape is indeed fundamental for determing hotels and accommodations prices; the research used different machine learning models (XGBoost, Linear regression, random forest and Multilayer perceptron) and identified three main categories for points of interest and their impact on accommodation prices:

1. Traffic related features: density of parking lots, density of coach stations, density of bus stops
2. Business related features: density of restaurants, density of markets.
3. Public Service related features: density of schools, density of parks.

It is important to note that every city is unique, with distinct characteristics and points of interest. As such, any research in this field must consider the specific points of interest and contextual factors relevant to each city. The dynamic interplay between accommodations and points of interest requires a nuanced understanding of the local environment to accurately assess and predict pricing dynamics.

### Conclusions

Different methodologies and machine learning models have been tested with success for predicting the price of a property in Airbnb, however only few researchers have explored the impact of features such as environmental noise, air quality and closeness to points of interest; research from other fields demonstrated how these features can be effective for enhancing the accuracy of a price predictor.

# Methodology

This section aims to elucidate the methodologies employed in this study, with a specific focus on two key aspects: primary research conducted through expert interviews and the overall research methodology. These methodologies are carefully crafted to guarantee the relevance and currency of the data, ensuring compatibility and impartiality in the study.

## Primary research

Among the many possible sampling available, this study selected In-depth interviews with Nonprobability sampling as main research method for primary data; Non-probability sampling implies that the sample does is not taken by applying a random probability sampling, the sample do not intend to represent the whole population instead it aims to select specific unit selected because of their unique or specific characteristics.

A comprehensive grasp of the research objectives is pivotal in determining an effective sampling strategy. This study seeks to forecast rental prices for tourism accommodations in Dublin by employing various machine learning algorithms. The emphasis is placed on leveraging sensor data and its predictive capability in determining rental prices, incorporating a multitude of features therefore requiring inputs from different experts.

The groups selected for the research are selected for gathering different perspectives on the matter offering a holistic view of the problem area, both populations have been selected as subject matter experts for their respective fields, the two populations are as follow:

* Group one: Noise and Air quality monitoring experts.
* Group two: Tourism and property management experts.

Noise and Air quality monitoring experts, such as those at the Dublin city council, can easily provide useful insights regarding the regulatory and legal standards regarding noise and air quality. On the other hand, property manager professionals can avail of their hands-on experience in property and rental price for identifying trends and causes.

The selected sampling approach is purposive sampling, primarily chosen to gather insights from individuals who possess significant expertise and experience in monitoring sensor data and managing property.

Units can be found through different means: by reviewing academic literature, reading professional reports or by researching for specific skills it’s easy to find highly knowledgeable and influential interviewees or subject matter experts, all these methods can ensure that the information gathered is relevant, detailed, and comprehensive of the topic.

However, it’s important to check for selection Bias when utilizing purposive sampling; aim of this research is provide a representative sample of the population of interest to avoid bias and ensure accurate results.

If a sample does not reflect accurately the population, the research can lead to incorrect conclusions or creating results which cannot be easily extended to the larger population, therefore minimizing or deleting the impact of the research: for minimizing the risk of selection bias some conditions are applied to the work:

* Firstly, from a group of selected units in both groups random selection will be applied to diminish possible biases.
* The overall final audience of experts must come from different backgrounds both in and out the group for minimizing bias.
* To establish minimum criteria for interviewees to take part.

Minimum Criteria for both groups:

1. To have at least 9 months of experience in the subject or related area.
2. To have worked in the relevant area (Dublin) for at least one year.
3. To have lived in the relevant area (Dublin) for at least one year.

The selection process involves thoroughly reviewing academic and professional background for both groups and documents for group one in order to identify the correct individuals who can provide detailed and highly relevant data to inform the research.

In total three in-depth interviews have been organised and their data gathered: 1 for group number one (Noise and sensor data) and 2 for group number two; the selected method for interviews was a Google form designed with open-ended questions for encouraging participants to share their opinions, experience, thoughts and perceptions on the matter in detail.

### Ethics

This research ethical framework is guided by complete obedience to a set of defined principals and ethics considerations that have the aim to ensure the research will be valid, reliable and have strong integrity.

The main principles are anonymity, informed consent and privacy, which will be employed since from the early stages of data gathering for then being constantly applied the handling and storage of all data collected.

One priority would be to ensure that the primary data is free from bias because it could have substantial consequences and undesired outcomes so primary data must be well presented and accurate, also, the storage and security of the data asks for ethical considerations; encryption is required, and access must be restricted for ensuring integrity and confidentiality.

Additionally, the publication of the resulting conclusion and exploitation of the findings represent also ethical considerations: anonymity of the participant must be preserved at all costs in the published research, although without forgetting to acknowledge all the contributors, also being transparent on potential conflicts of interest.

On the potential misuse of the study’s findings, this paper aligns with best practices in research, aiming to adopt credibility and trustworthiness in its results and implications: bias can derive from multiple factors, such as non-representative sampling (as explained previously in sampling strategy), wrong or subjective interpretation, or even biased questioning.

Bias can collect skewed data, and therefore resulting in conclusions which may not accurately represent the wider population or group ultimately affecting the overall validity and reliability of the research: For example, a specific group of experts might be under-represented or over-represented in the sample, therefore the prospective of the paper might not be reflected adequately in the research conclusions.

For eliminating this type of bias during the data collection process, the paper have to be transparent about the used methodology, for adequately represent the population, and ensure that results are trustworthy and can contribute to the study.

On the matter of primary Data being used for other purposes: even if remotely plausible, the results of this research could lead to unexpected consequences for uncontrolled secondary uses of the data, this is why Primary Data must surely be conserved and protected at all costs to prevent misuse or simple misinterpretation of the results; exploitation or any harm of the participants privacy must be negated, which is why obtaining informed consent from the participants is key by clearly stating the purposes of the research and any potential future use of the data.

For this reason, primary data must be securely stored using appropriate encryption and access being subjected to restrictive control measures as using data encryption and security access protection is the primary barrier alongside with anonymising the data by removing any identifiable information or link that could be used to trace the data back to the participants, the researcher will not share any data with other researchers for avoiding any breach in the ethical guidelines of consent conditions.

### GDPR

Whenever research is required to submit a primary dataset with personal data, the authors are subjected under GDPR, it comes that subsequently for storing, processing and publishing this data, if the results could lead to the reidentification of data subjects the appropriate measures have to be set in place:

* Control of access: All primary data is gathered in a private Google drive only accessible by using a combination of passwords and two-set authentication, also the data is not shared, published or present to anyone outside of the researcher providing limited access to it; after collection all data has been stored into a private folder only accessible by the researcher through password in a windows machine.
* Data destruction: At the conclusion of the research, the data will be stored for a specified duration in accordance with institutional guidelines or legal. Following this timeframe, all data will undergo secure and irreversible destruction, employing best practices to guarantee that no information can be accessed or reconstructed.
* Limiting possible exploitations: The conclusions and findings of this research are to be reported with Precision, by avoiding assumptions and defining technical terms clearly, therefore limiting and defining the context of the study to avoid misinterpretations or exploitation.
* Anonymity: All results from primary research will ensure the absolute anonymity and confidentiality of participants, this will be done by removing identifying details either manually or by using coding systems.

## Research methodology

The main framework for this research has been found in the CRISP-DM (Cross-Industry Standard Process for Data Mining) process as it stands as a pivotal tool, very used in the field of data science and analytics.

It was created as a standardized process for data mining projects, today CRISP-DM is being used outside of its initial boundaries and became an invaluable methodology for multiple types of research.

The advantage of CRISP-DM are clear as it provides systematic approach for navigating the landscape of data exploration, preparation, modelling, evaluation, and deployment, this framework facilitates the integration of diverse data sources which are the main point of the research which poses an accent in using sensor data which are not naturally included in property or rent datasets, CRISP-DM also ensure a methodical progression through each stage of research creating a comprehensive understanding of the underlying patterns and insights.

By using CRISP-DM, this research aims to discover the complexity of data-driven exploration and analysis therefore offering a solid foundation for informed decision-making.

The stages of CRISP-DM followed in this research as follows.

## Literature Review

First step for my research is to evaluate the current state-of-the-art by using literature review as main tool to understand the context of the research and previous approaches; on top of that in-depth interviews carried out using google forms are used to evaluate case specific topics related which will be a useful information for carrying out implementation, by Incorporating expert feedback the researcher will be equipped with useful suggestions.

## Business Understanding

The general “business understanding” is done by setting objectives which can be achieved within a limited timeframe (3 months), the objectives are already stated in the relevant section, to summarize:

* To evaluate the impact of various features on Airbnb prices.
* To evaluate if a neural network can outperform past used models.
* To predict the price of Airbnb apartments for Dublin city.
* To predict noise level using time series analysis on sensor data.

## Data Understanding

This phase of the CRISP-DM framework consists of literature review, feedback from the in depth interviews and overall exploratory data analysis to make a holistic understanding of the topic area and available data.

The feedback from professional expertise gathered during the interview phase was extremely valuable helping understanding the topic area and especially for redefining and better shape the objectives of the research.

Thorough exploratory data analysis has been performed in order to explain the features of the dataset and for initially pointing towards the main objectives of the analysis, correlations and insights have been discovered for the role of Noise and air quality feature in predicting prices, for the role of more traditional features in predicting Airbnb prices (room types, availability, etc.); also insights have been useful for the next phase for determine the best data preparation strategies.

## Data Preparation

During this phase the researcher had to prepare the data by dealing with missing values in the Airbnb datasets and mostly developing strategies for incorporating multiple sources of sensor data into a coherent dataframe usable in the modelling phase.

Decisions had to be made about datasets and trade-offs had to be taken in consideration, firstly the very nature of the Airbnb datasets was set as a screenshot of the available prices at a defined point in time which was a limiting factor due to the lack of historical data for the same properties

Big effort has been made for dealing with data normalisation and managing imbalanced and not-normal distributed data.

In this phase the researcher calculated and extracted multiple features based on the sensor data (mean, maximum value, minimum value and many other) for then being added to the property datasets and be evaluated after the modelling phase.

## Modelling

This phase included a thoroughly preprocessing of the datasets by using a specific pipeline in using pandas, Keras (for neural network) and sci-kit learn for Linear regression, Decision tree, Random Forest and Support vector machine.

This phase involved scaling of the features, developing the pipeline for training and testing datasets, scaling of the features (between 0 and 1), developing and testing multiple runs of hyperparameter tuning using Gridsearch CV; in total five models were evaluated, including Random Forest Regressor, a Neural network, a linear regression regressor, a support vector machine regressor and a Decision Tree regressor.

For time series analysis, a random forest auto regressor has been deemed the most appropriate.

### Linear regression

Linear regression is a staple among data science models, it plays and import role whenever there’s a need for predicting continuous outcomes from a set of variables.

The very basic of a linear regression model is the simple linear regression model where there is only one independent variable (x) associated with the dependent variable (y): by having multiple points of x and y the simple linear regression will try to create a line which will pass through the points in a manner of creating the smallest error possible where error is calculated by measuring the distance of the line (predicted values of y) with the actual value of y.

The line created will be composed of two parts: an intercept β0 and slope β1, the formula for creating the best fitted line is expressed as:

Where y is the dependent variable, x is the independent variable, β0 is a constant representing the value of y where x is equal to zero and e being the difference (error) between the predicted value of y and the real observed value of y; mathematically the best fit line is calculated by minimizing the residual sum of squares (RSS).

Having the formula as presented the only thing left for the model is to calculate the optimal values of and this is done by defining a cost function which in the case of linear regression is Mean squared error (MSE), the optimization algorithm used to minimize the cost function is gradient descent which calculates the minimum value of the cost function by changing interactive the values of and .

Being a parametric model linear regression makes assumptions on the underlying datasets, in many cases this assumptions are the reason for a regression model to perform good or poorly; firstly, linear regression assumes independence of residuals: the error should not be dependent on one another and no correlation should be present in residuals, also, the model assumes linearity between the independent and dependent variable, this meaning that the model will not work well if the datasets has non-linear relationships between the two variables; lastly, the model assumes normality distribution in the residuals and equality of variance in the residuals (homoscedasticity).

In the case of this research the linear regression model must be able to utilize multiple independent variables such as how many people the Airbnb property can accommodate, how many bedrooms, how many bathrooms, what’s the distance from multiple attractions, what’s the average noise for the closest sensor and so on, for this reason the model employed could not have been a simple linear regression but a multiple linear regression model.

A multiple linear regression model works similarly to a simple linear regression model, it only makes sure to have multiple slopes values for each independent variable.

When using multiple linear regression models a set of considerations must done: this model is typically subjective of overfitting as the model grows in complexity with a high number of independent variables, the model also captures more noise along with the underlying patter within data; this may lead to poor generalization of the model typically resulting in poor performance on testing data after good performance on the training set.

Another possible problem related to multiple linear regression models is the presence of multicollinearity: multicollinearity is the presence of high level of correlations among the independent variables, this makes it difficult to estimate which variables are contributing towards the prediction and which not.

Typically, the solution to multicollinearity is to be found on reduction techniques as Principal components analysis or by selecting some feature over the others; unfortunately, the scope of this research does not give the research the ability to deploy any of them as one of the objectives require the models to utilize all of the features and later to extract a feature importance score for evaluating if sensor data is effective at predicting Airbnb prices, therefore linear regression models on this research dataset is expected to have a lower accuracy than other models (Random forest) and being used more as a benchmark than a model to be used in a production setting.

### Support vector machine regressor

Support vector machine or SVM is a widely used supervised algorithm for tasks such as image classification, text classification or, even in less popular, regression; overall an SVM aims to find an hyperplane in a space composed of n number of dimensions which can separate and distinguish between two or more classes, overall it tries to maximise the margin between classes and minimizing classification error, SVM is able to handle both linear and non-linear relationships between the features by adopting different kernels.

SVM algorithms are mainly composed of three main parts:

* A hyperplane: defined as the line which is able to separate the different classes in a hyperspace with n dimensions, there can be many hyperplanes for separating the classes; SVM algorithms opts for the hyperplane which maximise the distance (margin) between the plane and the nearest points (supporting vectors).
* A kernel: a function which is able to take the input with a low dimensional space and transform it into a higher dimensional space, by doing what’s called a “kernel trick”, classes which were not dividable linearly in a lower dimensional space can now be divided within a higher dimensional space by using a hyperplane.
* Decision boundaries: can be seeing as the demarcation line dividing the points in the first class from the ones in the other class.

The scope of this research however is to define a regression model, Support vector machine can also be used for predicting continuous target variables such as prices in the case of this analysis; Support vector regression (SVR) models basically flips SVM on its head and instead of maximizing the distance of the hyperplane from the support vectors, the overall scope of SVR is to find the best fitting line which contains the maximum number of points in between the decision boundaries.

Support vector models can offer valuable insights and good performances: one of the main pros of this kind of models is that are highly effective on datasets with high numbers of features; especially when the number of features surpass the number of observations; overall, support vectors models work well when the features create a clear margin between different classes.

For the purpose of this research SVR is expected to work decently good as the dataset consists of a high number of features (78) even though the number of samples are higher, SVR can identifying both linear and non-linear relationships therefore is expected to work better than a linear regression model.

One of the main concerns for using SVR is the possibility of the high degree of noise due to the amount of features, overall SVR is expected to perform decently but from previous literature it has been often surpassed by ensemble learning models.

### Decision Tree

Like Support vector machine, decision tree algorithms pertain to the supervised learning model family; where the model is predicted given train data and provides a prediction to be evaluate and determine the loss.

As well as other models, decision trees can be applied to both regression and classification, anyway it’s very common for decision tree algorithms to be used in binary classification models.

Overall, the model works by using training data to develop decision rules which will be used later for determining prediction values.

A decision tree is made of different parts: the main component of the models is the node; a node can be described as a subset of data points, the act of splitting data from one into two sub-nodes (parent /child node relationship) is called splitting; in a decision tree model the first node which incorporates the entire population is called a root node, from here by splitting a second set of sub-nodes is created and iteratively the same happens for each node, the process then creates multiple branches from a common root node, the process keeps proceeding until the splitting reach a terminal node also called leaf.

In the end, the aim of the model is to split the instances of the classes by understanding the rules that best split the samples into the labelled classification.

In general, decision trees prefer categorical features as for their nature, splitting data by categorical rules (yes/no) is advantageous; however, they can work on any feature and are considered more robust than other models.

Many algorithms are utilized to decide how to best split a node into two child nodes, the main parameter to consider when performing a split is the increase of homogeneity in the resulting child nodes compared to the parent node: the model calculates the split on each variable of the dataset and then proceed with the split which provides the major gain in homogeneity.

Many algorithms which perform these calculations are based on the type of feature the model is trying to predict, for regression such in this case CART is the algorithm which has the best use; as for measuring the homogeneity of the child nodes a typical measure is the Gini index, mean squared error is commonly used for regression application.

The Gini index is a measure of entropy of a set group of items, high entropy means high randomness and with high randomness it gets more and more difficult to draw any conclusion on the matter; the algorithm calculates the best split to minimize randomness and applies it to the model.

Hypothetically a decision tree without limit can reach a 100% accuracy on the training data as per every single data point the model will develop a highly fitted branch for it; this clearly creates problems when the model will need to produce predictions on never seen before data (testing/validation set).

To solve the overfitting problem the researchers can best modelling a decision tree by imposing rules over some of its aspects, for example: the max depth of a tree can be set at a certain value, if not the decision tree will keep splitting data until the minimum number of samples per node is met; minimum number of samples per node is also another parameter which researchers can implement; different criteria such as Gini or MSE can be set up as well as different splitting strategies.

Another way to handle overfitting is to create what’s called a pruning decision tree, pruning happens on a fully grown decision tree and it’s the act of eliminating branches which do not impact the overall accuracy of the model, it’s done dividing the training dataset into two parts and using one as training set and the other as validation; at this point the model is prepared on top of the training set and optimized by leveraging the accuracy on the validation by cutting off the branches which do not result in higher accuracy.

Decision trees algorithms (DT) offer various advantages, among which:

* DT are robust against outliers when they are not overfitted: during the training process, determining the optimal split point involves calculating an average over a sample of the data, if this sample size is sufficiently large, the impact of outliers is likely to be mitigated or suppressed.
* DT are easy to interpret: because of the architecture of the model, which is formulated as a list of steps where at every steps a “question” is presented to the parent node, the “questions” can inform researchers which features are being used to generate a prediction; in favour of explainability many libraries such as sci-kit learn provides visual representations of the model for better interpretability.
* Non-linearity: as non-linear model DT can be applied to a variety of complex problems.
* DT are Nonparametric: being parametric, DT models don’t make assumptions on the underlying dataframe, this is helpful as the model can be applied no matter what to the problem domain.
* DT and categorical values: because of their architecture DT models are able to use both continuous and categorical features.

Even though, Decision trees model can be perceived as a jack of all trades, they still need to be treated carefully as they also maintain some disadvantages, mainly:

* Overfitting: as expressed above, DT models are quite prone to overfitting, pruning and hyperparameter tuning tends to solve for this issue, however DT models need to be correctly handled and maintained for avoiding overfitting.
* DT are subject to noise: because of the greedy approach of the optimizer which maximise the homogeneity of the next split over the overall accuracy of the mode, in presence of noisy data DT optimizers could be set on local minima over global minima.
* DT don’t perform as good in regression: because DT models use a step-based approach where each node is split based on a categorical rule (either < or >) they don’t work well for regression as for classification; which is why, in the context of this research Decision tree regression is mainly set to be a benchmarking tool for evaluating ensemble models (Random Forest)

### Random Forest

In layman’s terms, random forest algorithm is a supervised learning algorithms which utilizes multiple decision trees (trained on various subsets of the same dataset) to increase the overall accuracy therefore solving complex problems, this is done by collecting the results of each tree and deciding the final output by selecting the most frequent answer among the trees.

The process of using multiple models to derive a conclusion is generally called ensemble learning: in ensemble learning every model by using the same data, then by majority voting the overall forest architecture process a prediction which can be either a continuous or categorical variable if the problem is a classification or regression problem; main requirement for ensemble learning is that the error from each tree is independent from one another.

In a random forest algorithm, the data used to train each decision tree is split in different subsets (one per tree), the subsets are extracted from the main dataset by using sampling with replacement, this technique is generally referred to as bagging.

Along with bagging, Random Forest algorithms also use a random selection of features (hence the term random in “Random Forest”) instead of using all features and then calculates which feature gives the best split.

Main advantages of bagging are that the model knows how to handle high dimensionality (many features), however, bagging does not perform well for regression as the output is based on the mean prediction which hurts precision in regression problems, this will be taken in consideration given the problem area requires to maximise regression accuracy.

However, another ensemble technique is widely used in Random forest algorithms called Boosting: Boosting is a technique where the individual learners are trained sequentially and not in parallel as for Bagging; by fitting consecutive models (decision trees in this case) at each step; the first learners will fit simple models and by analysing the error after every step the subsequent model learn and therefore solves for minimizing the error to previous tree.

### Feedforward neural network

A neural network is an artificial intelligence architecture that mimics the human brain by creating a complex web of interconnected neurons and layers, different version of neural networks are used for different applications: the most used are ANN, CNN and RNN:

* ANN: also known as Feedforward neural networks, ANN are less powerful than CNN or RNN, the architecture has a fixed length and they contain hidden layers between the input and output layer
* CNN: CNN are very powerful and mostly used exclusively for image processing, the main difference from other networks is the presence of Convolution layers.
* RNN: Mostly used for time series data, in RNN the output of a node is fed back into the node or into a previous node, one of the most used architectures is Long-short term memory (LSTM)

Specifically, this research is set to use Feedforward neural network, A feedforward neural network is made of a series of neurons divided into different layers, each neuron is connected to every layer of the subsequent layer; the feedforward architecture is the basic and most used type of model in which the inputs are processed only in one direction (hence the name) and can’t flow in the opposite way.

In a feedforward neural network, each input in a layer is multiplied by a corresponding weight; then the individual products are summed, and a bias added, the resulted weighted sum is then processed by an activation function; the scope of the activation function is to introduce non-linearity and helps the network capture complex patterns, at the end the output is compared to a threshold value, and if the sum exceeds the threshold, the neuron is activated.

The same process is repeated for each layer until the final output is obtained, scope of the training is to adjust the weights and biases of each node to minimize the error between predicted values and true values in the validation set.

In regression applications of a neural network the cost function to evaluate predicted values and validation set can be any regression cost function such as MSE or RMSE, by calculating the cost then the model knows how to adjust the weights in the model to minimize the delta between predicted and true results, the algorithm in charge of doing this operation is referred as backpropagation algorithm.

In summary, Feedforward neural network have the advantage of having learning capabilities on non-linear and complex relationships, this make a NN able to adapt to multiple real-life problems; especially, NN are more easily generalized which comes in handy when utilizing unseen data; lastly, due to the many features employed in this research, NN seems to be the best option: NN has the ability to utilize many variables without carrying about the distribution of those, NN are great at handling heteroskedasticity and datasets with high levels of volatility and difference variance levels.

This research is expected to be able to utilize NN to better predict the Airbnb prices, especially in relationship with random forest algorithms as those are the best performing in the literature.

## Evaluation

Evaluation phase was set up to be as simple as possible for better capturing the explainability of the models: due to the scope being a regression problem the coefficient of determination (R2) was selected as main evaluation metric along with the mean absolute percentage error which can be extremely helpful when it comes to explainability because explains the error in percentage of the absolute value, providing additional insights into the model's performance.

Visualisation were developed to better visualize how the models compare with one another.

For evaluating the impact and usefulness of the Noise and Air quality features in predicting Airbnb prices Feature importance scores have been adopted as the main source of evaluation, this decision came with a trade-off as feature importance cannot be applied to neural networks due to their nature itself, which is why feature importance have been only applied to the best performing model excluding the neural network even if it has been proven the most effective.

# Implementation

## Primary data

A series of in-depth interviews through Google forms with property, tourism and policy experts, these discussions offered valuable insights into the realm of the topic, highlighting challenges, limitations and suggestions invaluable for the research.

### Group one: Noise and Air quality monitoring experts

This discussion focused on the possible impacts of Noise and Air quality on property values with a focus to understand if the same impact can be attributed to short-term rentals such as Airbnb accommodations, the expert also pointed useful resources to deepen the understanding of health consequences related to Noise and air quality levels which can indeed influence accommodations.

The expert highlighted the importance of health impacts from Noise and Air quality levels, as expressed in the World health organization Noise guidelines: annoyance and sleep disturbance can impact the enjoinment of an accommodation therefore turning a high value accommodation into a lesser valued accommodation.

The expert pointed out key aspects of the noise distribution in Dublin: most of Airbnb accommodation are located in the city centre area, this area is also louder than suburban counterparts, this meaning that it might be worth using noise sensor data to predict areas that are more attractive than others, overall, more noise means more activities in a urban setting and it may indicate that tourist are paying a premium for being accommodated closer to busy areas.

Finally, the expert pointed out that as well Air quality might be an indicative factor for prices in the city centre.

### Group two: property and tourism experts

These discussions focused on getting a sense of limitations of the research as well as developing an understanding of which points of interest to consider when evaluating the impact of distance to point of interests for the average tourist; in particular, one of the interviews was particularly helpful in understanding the current picture for accommodations in Dublin city centre.

The experts suggested to include a measure of closeness to city centre as, there’s clear signs of higher prices being paid for both properties and short-term rentals when closer to the city centre, also, the experts have pointed out how the current economic outlook has seen prices going up rather quickly after the Covid-19 pandemic with rates being 40% in 2023 compared to 2019 as result of the monetary expansion which led many people to save up during the pandemic and spend now in travels.

Also, short term rentals capacity has been a key-point of discussion: current government policies are aiming to restrict short term rentals in rent pressure zones and urban city in general in an effort to keeping the pressure off regular rents, in this environment short term rentals have dropped to 55% of capacity of what they were in 2019; also, due to recent events the capacity constrains coupled with a drop in overall tourism accommodation capacity taken up or contracted by the government to those without homes, including asylum seekers.

This creates certain limitation to the research because of the peculiar state of short-term rentals in Dublin as the research is conducted on data from December 2023.

Finally, experts have been tasked with specifying the degree of influence that proximity to various tourist attractions would have on accommodation prices.

### Key decisions record

Following discussions with experts, some key decisions have been made for operating the research, these are expressed as follow:

* Distance from tourism attractions: the decision made was to implement the distance from the top tourism attractions by number of visitors in 2022, data from Failte Ireland [https://www.failteireland.ie/Research-Insights/Activities/visitor-numbers-to-attractions-dashboard.aspx], the experts weren’t able to suggest some attraction in particular, therefore a decision has been made to include the attraction which reached at least 400.000 visitors in 2022 (Trinity college, Dublin zoo, Guinness storehouse, National gallery of Ireland, Irish museum of modern art, Saint Patrick cathedral).
* Noise sensor data: a decision has been made to extract the number of times a sensor has reached a level of 60,70 and 80 decibels as noise can lead to noise disturbances especially for high and sustained levels, as referenced in the World health organization guidelines for the European union [https://www.who.int/europe/publications/i/item/9789289053563]
* Distance from tourism attractions: distance from the monument “The Spire” has been included as feature as a proxy for distance from city centre.
* Noise sensor data: a decision has been made to aggregate reading of sensor data by hour, as sleep deprivation has been deemed the most important consequence from urban noise which might indicate differences in accommodation prices.

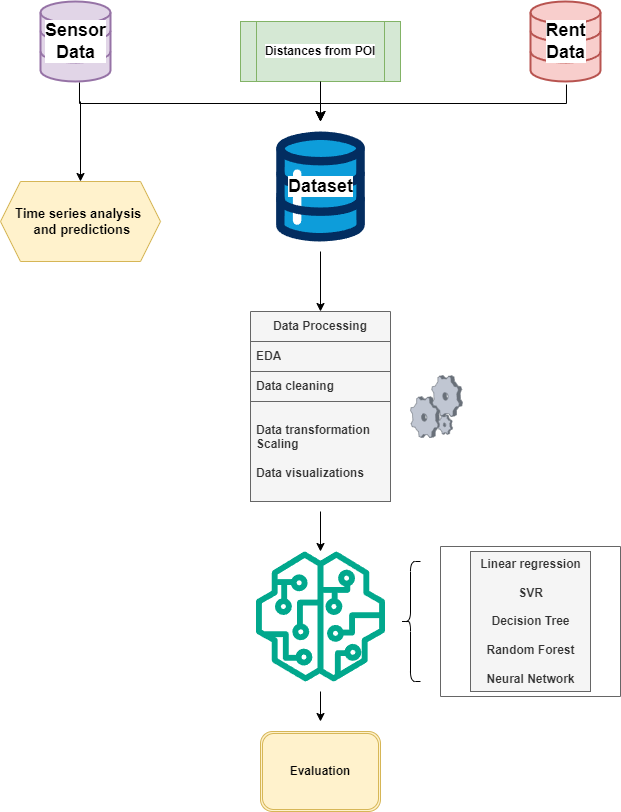
## Data Architecture

This research employed an architecture developed entirely on-premise, the overall data needs were sufficient to be run on a window machine without having to resort to cloud computing.

Personal laptop: Windows 11, RAM 16.0 GB

Version control: GitHub (<https://github.com/alexCCTcollege/Capstone>)

As shown in figure [], the architecture revolved around merging multiple data sources: Sensor data, distances from Point of interest data and rent data into a comprehensive dataset which later has been used for Modelling after careful data processing through Exploratory data analysis, Data cleaning and scaling, evaluation at the end was used for determining the main differences in accuracy of the employed models and by extracting feature importance for sensor and point of interest features as a measure of the effectiveness of those to be used as predictor for Short term rentals prices in Dublin, time series analysis on sensor data has also been done to facilitate the possibility of using sensor data in the future for making predictions.



## Datasets

This research in total has collected three datasets:

* Airbnb Dataset: the data is procured from “Inside Airbnb” [] , the project is a collection of data regarding Airbnb listings for multiple cities, the mission of the project is to add data to debate regarding tourism rental policies; the data for this project is only made of listings in the Dublin areas, Inside Airbnb provides data for a specific timeframe, in this case data is from the date 15/12/2023, de facto providing a screenshot of available Airbnb lettings in Dublin for that date.
* Sensor data: Noise and air quality data is extracted from the Dublin city council ambient and air quality noise website (<https://dublincityairandnoise.ie/>), the website provides real-time data from air quality and environmental noise monitoring stations, including those managed by the Environmental Protection Agency and Dublin City Council; as expressed in the page the data has limitation as It is a work in progress and the data is not fully validated and may experience fluctuations due to calibration, maintenance, or instrument issues; slight variances in data may occur due to delays in synchronizing readings between monitoring stations and websites as well as certain climatic conditions, like foggy weather, can impact the performance of air quality analysers leading to potential inaccuracies in particulate pollution levels.
* Points of interest: Data for touristic points of interest only consists of coordinates for the touristic attractions taken in consideration for the analysis, data has been extracted manually by locating the coordinates on Google maps.

## Data understanding

### Sensor data

Due to the Airbnb dataset being set for the date 15/12/2023, it has been decided to extract reading from sensors for all monitoring station from the 15th of November 2023 until 15th December 2023 as the last month of readings was concluded to be the most relevant in defining the relationship between sensor data and price of the accommodation.

After excluding monitoring stations with missing data, in total, data from 32 stations (divided between air quality and environmental noise) has been analysed.

Air quality stations recorded multiple parameters, however only two parameters were taken in consideration as they were present for all locations:

* PM2.5 (particulate matter 2.5): this measure records the quantity of particles of 2.5 micrometres or smaller suspended in the air, these particles can originate from various sources like construction activities, industrial processes, vehicle emissions, or other natural sources like dust and pollen, PM2.5 is associated with health effects such as respiratory and cardiovascular problems.
* PM10 (particulate matter 2.5): similarly this measure records the quantity of particles of 10 micrometres or smaller suspended in the air, they can come from different sources such as constructions sites or agriculture, these particles are larger than PM 2.5 and can be filtered out by the human nose, however PM10 can still include health effects like eye and throat irritation and respiratory conditions

Noise sensors on the other hand only recorded the noise level in decibels, the data then was reported from the Dublin city council as an aggregate value by hour.

### Airbnb Data

By far the biggest dataset is the Airbnb dataset, listings from Inside Airbnb for Dublin at 15/12/2023 were in total 9020, the dataset is composed of multiple features for each listing which can be divided in multiple themes, the data license can be found at (<https://creativecommons.org/licenses/by/4.0/>) :

1. Host related information:   
   1. host\_id: unique identifier per host.
   2. host\_url: Link to the property on Airbnb.
   3. host\_name: name of the host.
   4. host\_since: registration date of the host on Airbnb.
   5. host\_location: City of the host.
   6. host\_about: Description of the host inserted in Airbnb.
   7. host\_response\_time: average response time with four possible values ('within a few hours', 'within an hour', 'a few days or more', 'within a day').
   8. host\_response\_rate: Percentage of response rate from the host.
   9. host\_acceptance\_rate: Percentage of acceptance rate by the host.
   10. host\_is\_superhost: Super host is a badge earned by hosts on Airbnb if some conditions are met such as a response rate higher than 90%, having completed 100 nights in the platform, maintaining a 1% cancelation rate and a 4.8 out of 5 rating score ([AirBnb help centre: super host](https://www.airbnb.co.uk/help/article/829" \l "section-heading-2-0)).
   11. host\_thumbnail\_url: host thumbnail.
   12. host\_picture\_url: picture of the host in the platform.
   13. host\_neighbourhood: neighbourhood of the host.
   14. host\_listings\_count: number of listings of the host.
   15. host\_verifications: means of verification used by the host (email, phone).
   16. host\_has\_profile\_pic: if host has profile picture or not.
   17. host\_identity\_verified: is host verified?
2. Accommodation related information:   
   1. Neighbourhood: in which neighbourhood the listing is located.
   2. latitude: latitude of the listing.
   3. longitude: longitude of the listing.
   4. property\_type: categorical type of the listing such as for example full apartment, Condo, cabin, room in shared apartment.
   5. room\_type: type of room such as double or single.
   6. accommodates: number of people the listing can accommodate.
   7. bathrooms: number of bathrooms in the property
   8. bedrooms: number of bedrooms in the property
   9. beds: number of beds in the property which can be different from number of bedrooms in the case of lettings by bed and not by entire room.
   10. price:
   11. minimum\_nights: minimum nights to be selected for the listing.
   12. maximum\_nights: maximum nights to be selected for the listing.
   13. calendar\_updated: last date the listing was updated.
   14. has\_availability is the property available on the date 15/12/2023
   15. Availability: number of times the listing was available in the past 30, 60, 90 and 365 days
   16. calendar\_last\_scraped: the date the listing was extracted (all listing in this research have the same date 15/12/2023)
   17. instant\_bookable: is the listing instantly bookable? Instant book is a feature in Airbnb which allows customer to book the room instantly without waiting for an answer from the host
3. Reviews related information:   
   1. number\_of\_reviews: total number of reviews by listing.
   2. first\_review: date of the first review for the listing.
   3. last\_review: date of the last review for the listing.
   4. review\_scores\_rating: total review score.
   5. review\_scores\_accuracy: review score per theme “accuracy”.
   6. review\_scores\_cleanliness: review score per theme “cleanliness”.
   7. review\_scores\_checkin: review score per theme “check in”.
   8. review\_scores\_communication: review score per theme “communication”.
   9. review\_scores\_location: review score per theme “location”.
   10. review\_scores\_value: review score per theme “value”.

## Data preparation

An important part of the research has been to extract relevant features from data to be modelled later, it has been the case for environmental noise and air quality data; there’s evidence that noise can have an impact on property prices, especially for areas with high and sustained levels; and based on previous literature the following features have been extracted:

Both for air quality and environmental noise basic statistical aggregates (Mean, maximum, minimum and variance) have been extracted; especially for noise data mean and maximum values are considered to be relevant as in one case sustained levels of noise (due to construction or other reasons) that can be captured as mean values can derive a less attractive property, paradoxically in some areas high level of noise can also be correlated to night life and therefore being a factor of attractivity.

Due to the nature of the topic, it was of high importance to extract features which could differentiate between working days and weekends; a listing in which weekend environmental noise can represent quiet a different experience from a quieter area and therefore having two different prices, this is why the same statistical measures have been extracted for weekend days only, hoping to capture the variation between weekdays and weekends.

Also, for environmental noise It has been extracted the number of readings above certain thresholds, respectively 60, 70 and 80 decibels; the number of readings above these thresholds can be quiet telling as for comparison a decibel level of 60 can be compared to a normal conversation while a decibel level of 80 is considered very loud and it can be compared to the sound of a truck passing by; this features can highlight the peaks of noise in certain area and differentiate even further between noisy and quit areas.

In terms of combining together all different data sources, each listing has been associated with the closest sensor and their respectively measurements (mean, minimum, maximum, weekend mean etc…) by using coordinates; the function for putting together sensor data and listings have been implemented by calculating the distance by using a formula derived from the Pythagorean theorem as in figure.



The formula has been placed in a series of loops where the distance from a listing to every sensor has been compared and then closest selected, after that the attributes (mean, minimum etc…) from the closest sensor has been attached to the selected listing and continuing with the next listing in line, full function in figure.

A computer screen shot of text

Description automatically generated

The same formula has been used for calculating the distance of each listing from the selected location of tourism point of interest: The Spire, Trinity college, Dublin Zoo, Guinness storehouse, National gallery of Ireland, Irish museum of modern art, Saint Patrick cathedral).

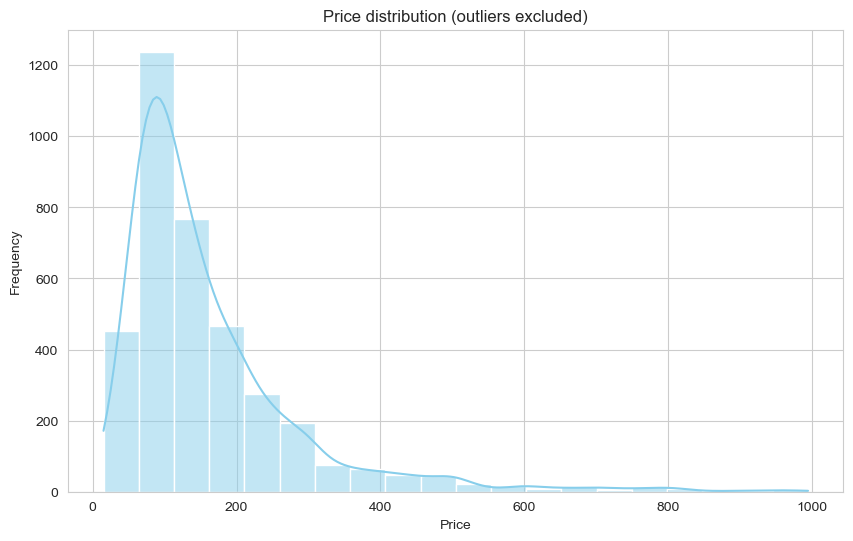
## Exploratory Data analysis

As part of the research extensively exploratory data analysis has been conducted to validate previous assumptions and to find insights for better modelling and gaining insights for the many features present in the dataset.

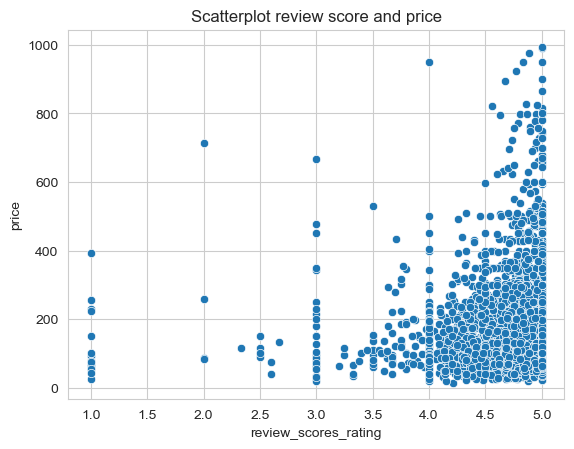
After data cleaning the dataset was composed of 71 features and 3754 rows in total.

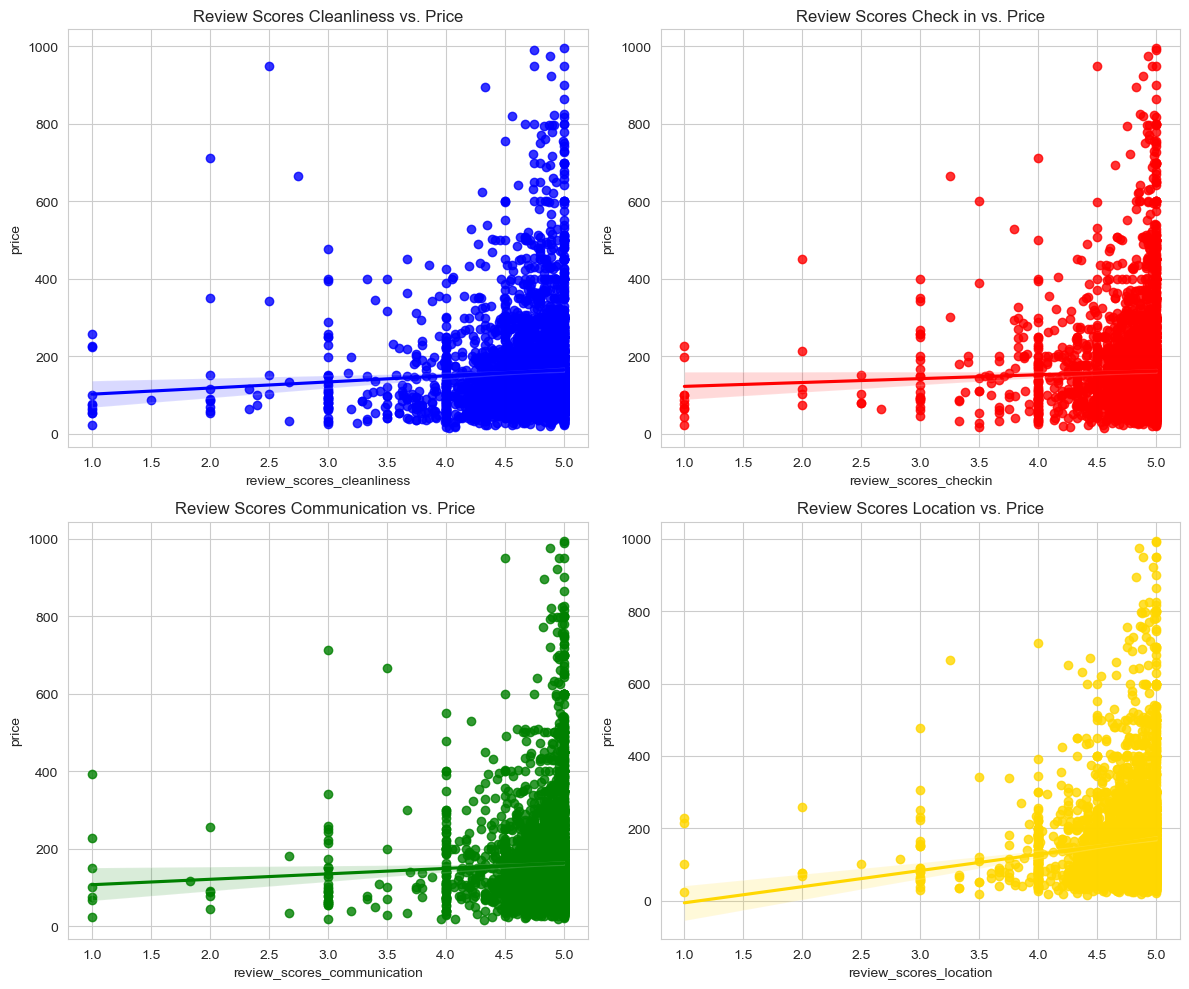
Firstly, as this research relates to understanding and modelling for prices of Airbnb listings, a closer look at the distribution of prices for Dublin city at the end of 2023; as shown in figure the price attributes of listings in Dublin approximate to normal with a cutoff at around 50 euro per night with a mean of euro per night.

All visualization regarding price are done on a dataset without outliers for better suiting the visual aspect, however outliers have been kept for modelling as there was no reason for getting rid of them from a exploratory point of view, even further it would have possibly compromised the research by introducing bias.



As expected, the higher the overall reviews the higher the price of the room, the dataset also contained reviews by theme (cleanliness, check in, communication and location) which also highlighted a positive correlation between rating scores and price.



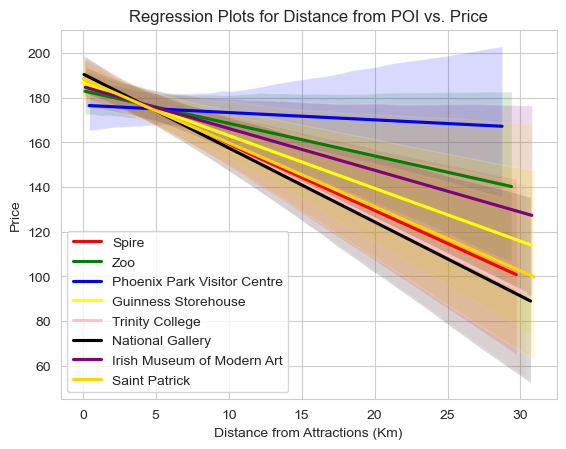


Other insights were gained during exploratory data analysis, is has been discover how the price was higher for property listed as entire house rather than room in shared apartments, as shown in fig where 0 represents a room in shared accommodation while 1 represents an entire house.

A graph of a number of objects

Description automatically generated with medium confidence

Regarding the features extracted from other sources (distance from Point of interest, Air quality and environmental Noise levels), a steep negative correlation has been highlighted between distance from most of the attractions taken in consideration as highlighted in the regression plot in fig

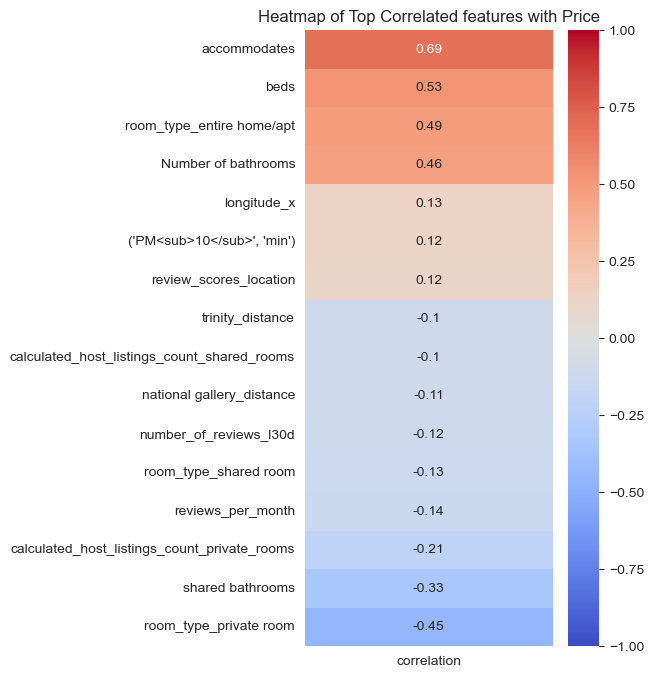


Regarding the impact of both noise and air quality on prices, very few factors highlighted a possible correlation or evidence of impact on prices, expressed in figure n for example, price seems to be independent from both the average Noise level and the average PM 2.5 level.

A red and green lines

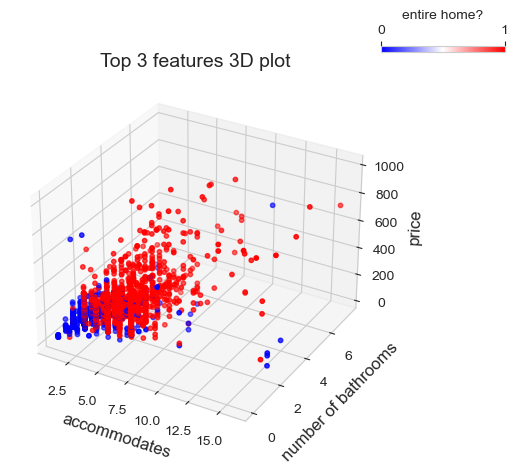
Description automatically generated

For valuate a first approach to which features had the most impact on price, correlations for each of the 70 features with price have been created, the result is visible in fig showing the most correlated features (both negatively and positively) with price.



It’s clear how the number of people a listing can accommodate as well as number of beds have a higher degree of correlations among the features, features related to noise and air quality however do not seem to correlate much either positively or negatively with the price factor, it is true however that distances from many touristic points of interest beat in terms of correlations even with relatively low scores.

A representation of the most correlated features has been created in fig by using a 3D scatterplot using the features: price, number of people a listing can accommodate, number of bathrooms in listing and as colour is the listing is an entire apartment or room in shared house.



The positive correlation seems apparent between number of people a listing can accommodate and number of bathrooms, as well as being clear how room in shared house are worth less than entire apartments.

# Model Results

## Time series analysis

Nature of this scope is to evaluate the use of sensor data for predicting prices of Airbnb listings; as part of this analysis, it’s clear how in order to have effective predictions is vital that the attributes for sensor data can be reliably predicted therefore a time series analysis has been deemed essential for the progression of the study.

As explained before sensor data collection has included data from a month prior to the extracted date of listings (15th December 2023) therefore the total number of readings gather were only spanning across one month, thus creating limitation for sure but since readings where aggregated on an hourly basis, sensor data was deemed to be effective enough to proceed.

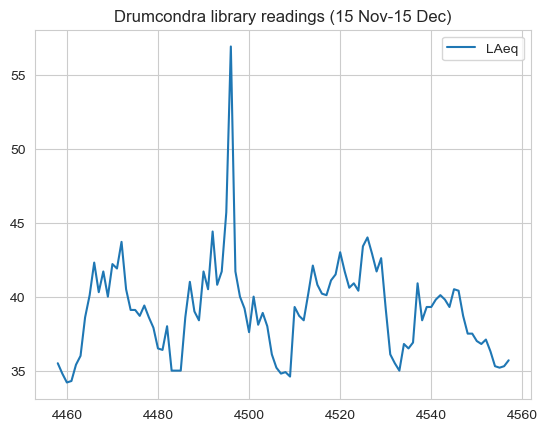
Time series analysis has been conducted on one sensor only picked up randomly, the overall scope of the research was to develop a process and an understanding of modelling features, therefore developing an hyper parameter tuning models for each sensor would have been too much labour intensive and outside of the scope of the analysis: the sensor selected as example was the one named “Drumcondra library”.

First step of the time series analysis was to check for stationarity of the time series, a Dickey-Fuller test has been, Dickey-Fuller test is a statistical test used to determine whether a given time series is stationary or not based on the present of a Unit-root, the test assumptions are as follow:  
  
H0 = the time series has a unit-root therefore it is Non-stationary

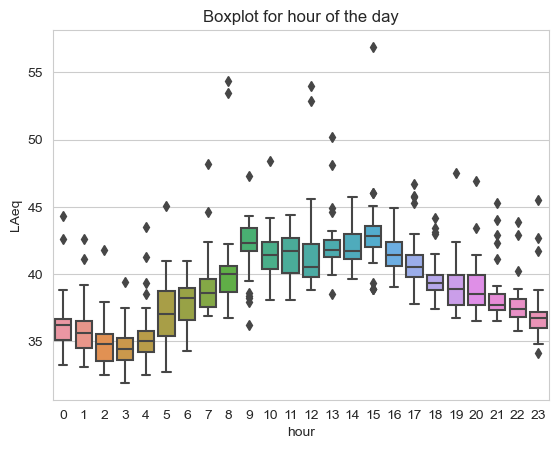
Ha = the time series has NOT a unit-root therefore it is stationary

The p value selected for the analysis was 0.05, in case of a non-stationary series, additional processing has to be run in order to simplify the modelling phase by utilizing techniques such as trend-removal and seasonal adjustments.

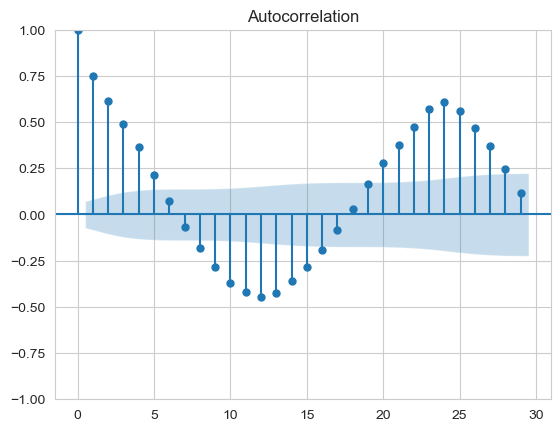
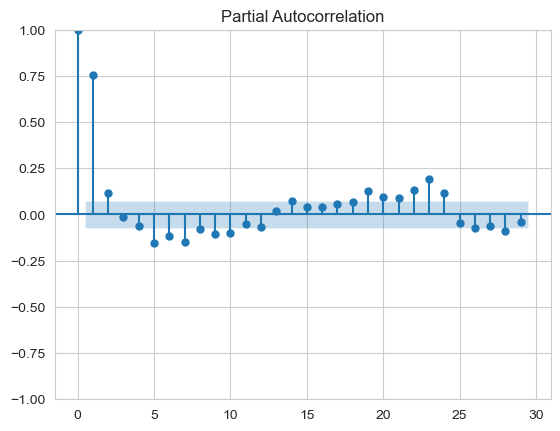
The p value for the test was 0.0011, being smaller than 0.05 is meant that the series was stationary therefore it was appropriate to continue with the analysis.



By exploring the data it was apparent that a level of weekly and hourly pattern was present, the pattern shown in fig is expected as noise in an urban setting is closely related to the presence of cars in the street and night time hours are expected to be quitter than day time hours.



In terms of modelling, the clear pattern revealed deemed necessary to consider hour as exogenous feature for better modelling, also, partial and auto correlation has been run in order to best fit the model.

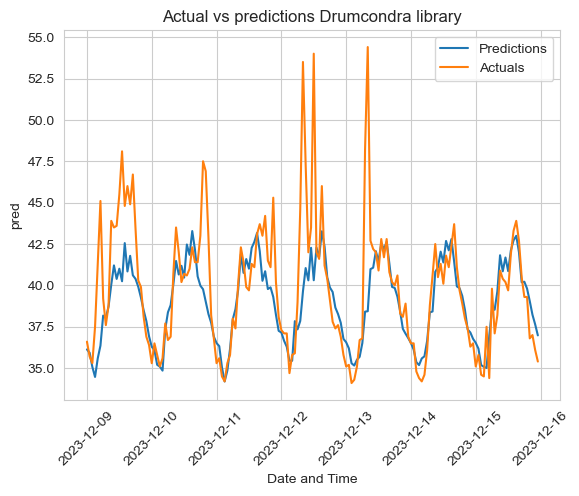


The model chosen for autoregression was a random forest with multiple hyper parameters:

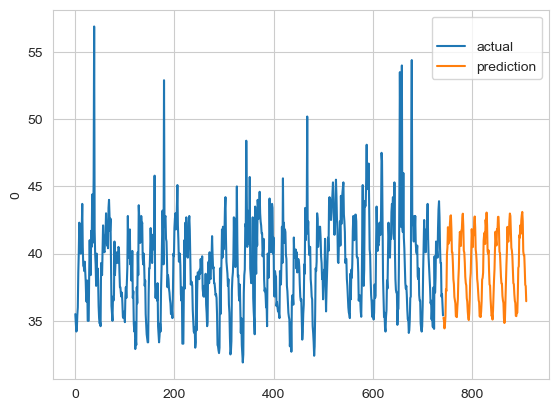
* n\_estimators:
  + Values: [10, 50, 100]
  + Definition: The number of trees used in the forest.
* max\_depth:
  + Values: [None, 10, 20]
  + Definition: The maximum depth of each decision tree in the forest.
* min\_samples\_split:
  + Values: [2, 5, 10]
  + Definition: The minimum number of samples required to split an internal node.
* min\_samples\_leaf:
  + Values: [1, 2, 4]
  + Definition: The minimum number of samples required to be at a leaf node. It controls the minimum size of terminal nodes.
* bootstrap:
  + Values: [True, False]
  + Definition: Whether bootstrap samples are used when building trees. If True, each tree is built on a random subset of the training data with replacement.
* max\_features:
  + Values: ['auto', 'sqrt', 'log2', None]
  + Definition: The number of features to consider when looking for the best split.

In total the number of models compared was 1944 taking in total approximately 27 minutes, the selected evaluation matric was mean squared error and at the end of the Gridsearch parameter tuning, the best model was selected to have bootstrapping, a max depth of 20, the max feature strategy set on “log2”, a minimal sample leaf of 2 and a minimum sample split of 5, the number of estimators was 10.

Predictions and actual values have been plotted to closer evaluate the accuracy of the predictions in fig, the testing window was set as the last seven days in the dataset: overall, it is clear how with one month worth of data the model does a decent job in minimizing the error, however it comes short in capturing higher values.



Once evaluated the model was then fitted for all data (1 month long) and predictions for the following week plotted in fig.



## Regression Results

Two main measurement tools have been used to evaluate the accuracy of the regression models: Mean absolute percentage error and R2:

* MAPE calculates the average difference between actuals and predicted values as percentage difference.
* R2 is a very common evaluation metrics that represent the proportion of the variance that is explained by the independent variables in the model.

As expected, ANN and random forest were the best performing models, while linear regression and SVM failed to capture much of the variance, decision tree performed very poorly as expected.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **MAPE** |  | **R squared** |
| Random Forest | 0.337887 |  | 0.703256 |
| Linear regression | 0.485357 |  | 0.585342 |
| SVM | 0.332732 |  | 0.593810 |
| Decision tree | 0.358011 |  | 0.508484 |
| Neural network | 0.376989 |  | 0.790208 |

### Linear regression (R2 0.58)

Linear regression is a simple model which is not able to extract non-linearity from a dataset; in this research the model was used to set the baseline benchmark before experimenting with more complex models. By using all available features and being tuned by using two parameters available using GridsearchCV and Scikit-learn:

* Fit\_intercept:
  + Values: [True, False]
  + Definition: when False the intercept is forced to be 0, otherwise the best fitted line will calculate the intercept value.

Overall, the model performs poorly with a R2 of 0.58 and a MAPE of 58%, this may be due to the fact that the many features in the dataset might present non-linearity attributes which would not be captured by the linear model.

### Support vector regressor (R2 0.59)

Differently from linear regression, Support vector regression can capture non-linear attributes, however SV models work best for datasets when the number of features is higher compared to the number of samples, the model was tuned using multiple parameters:

* Kernel:
  + Values: [Linear, Poly, rbf]
  + Definition: Kernel is the mathematical function which is used to transform the training set into a high number of dimensions, linear is a linear function, rbf stands for gaussian radial basis function, poly stands for polynomial function.
* C:
  + Values: [0.1,1,10,100]
  + Definition: C is the regularization parameter and it determines the penalty for misclassification, a high C value narrows the margins of the SVM.
* epsilon:
  + Values: [0.01, 0.1, 0.2,0.5]
  + Definition: epsilon specifies the epsilon-tube within which no penalty is associated in the training
* gamma:
  + Values: [scale, auto]
  + Definition gamma parameter defines how far the influence of a single training example reaches.

SVR performed poorly with scores very similar to linear regression with a R2 of 0.59 and however a smaller MAPE value of 33%.

### Decision tree (R2 0.51)

Decision tree was the first tree-based model to be used in the analysis, the scope of the model was to be used as benchmark for random forest, as explained in the methodology DT don’t perform as good in regression: because DT models use a step-based approach where each node is split based on a categorical rule (either < or >) they don’t work well for regression as for classification, however the model has been developed and tuned with the following parameters:

* Criterion:
  + Values: [absolute\_error, poisson, friedman\_mse, squared\_error]
  + Definition: this parameter determines which measure is to be use for calculating the impurity of a split.
* Max\_depth:
  + Values: [100,10]
  + Definition: the maximum depth of the decision tree.
* splitter:
  + Values: [best, random]
  + Definition: this parameter is used to select which features will be used for the split, random selects the features randomly while best select the feature for the best split.

The model performance was the worst in terms of R2 0.50 with a MAPE of 35%.

### Random Forest (R2 0.70)

The random forest model was the most tuned out of the five models, in total the tuning and training took 14 minutes with a GridsearchCV tuning strategy as follow:

* n\_estimators:
  + Values: [50, 100]
  + Definition: Number of trees in the forest.
* max\_depth:
  + Values: [10, 20]
  + Definition: Maximum depth of each tree.
* min\_samples\_split:
  + Values: [2, 5]
  + Definition: The minimum number of samples required to split a node.
* min\_samples\_leaf:
  + Values: [1, 4]
  + Definition: The minimum number of samples required to be present in a leaf node.
* bootstrap:
  + Values: [True, False]
  + Definition: if bootstrap samples is used when building trees.
* max\_features:
  + Values: ['auto', 'sqrt', 'log2']
  + Definition: The number of features to consider when looking for the best split.

Random forest was expected to be one of the best performing models as its ability to be robust to outliers and ability to combine multiple regressor into a coherent prediction, in fact the model performed better than decision tree, linear regression and SVR with a R2 of 0.70 and a MAPE of 34%.

### Neural network (R2 0.79)

To compare more advanced models, An artificial neural network has been developed; different architectures have been tested and the best one developed with the following parameters:

In total the best model was formed with three layers: 78 neurons in the first one as per the number of features with activation function “Leaky Relu”, then into 156 neurons for the hidden layer with “Leaky Relu” activation function into a single neuron in the last layer as the model is meant to be a regression and not a classification model.

As expected the feedforward neural network was the best performing model in terms of R2 with a value of 0.79 and a MAPE of 37%, this meant that the model was capturing non-linear relationships better than any other model.

## Relevance of Noise, Air quality and Distance for predicting price

In order to understand which features have been the most influential in determining the variance in Airbnb Prices: feature importance has been run on the Random Forest model; Random Forest was the good balance between performances and model explainability having a R2 score very close to the neural network and being able to explain how much of the variance each of the 78 features was explaining.

The results are shown in Annex 1, as shown in figure the main factors for determining the price of the Airbnb were features present in the Airbnb Dataset, mainly the number of people a listing can accommodate with a score of 16%, the number of rooms a host is renting on the App (10%) and the number of beds in the listing (5%); distances from various points of views have a score of around 1% with the best being the distance from the Irish museum of Modern Art (1.05%); all features related to Noise and air quality sensors were among the worst in terms of feature importance with scores below 1%.

# Conclusions

## Limitations

## Future steps

# References

# Annex

# 

VIP