# Introduction

An accurate house price prediction is important for prospective homeowners, real estate developers, investors, banks, governments, tax assessors, insurers, and mortgage lenders (Frew and Jud 2020; Ihre 2019). House price in Ireland has been highly volatile due to several factors such as availability, material cost, population density, and rising rents (Reddan 2018). Several prospective homebuyers are struggling to secure enough money for their first home (Coughlan 2022). Having an accurate prediction regarding the house price will help these buyers to plan for their first home. Different machine learning models can predict house prices considering several factors that are not considered while using traditional methods for house price predictions.

While traditional methods use factors such as type, size, quality of finish, location, number of floors, area of the property, availability of facilities such as schools, hospitals, and grocery stores; modern machine learning algorithms can factor in several more such as inflation, salary, environmental factors, and geography (Hurley et al. 2022). Furthermore, machine learning will help to identify house price determinants that are selectively applicable for the stakeholder who will be using these predictions. However, individual contributors that influence the house price will be different for each state and country. So, it is always a benefit for the stakeholders to know the changes that are predicted to happen soon to the property market to act accordingly.

The house prices in Ireland have been increasing every year and the prospective homeowners are finding it difficult to find a home within their budget. Ireland faced a construction boom with wage growth, bank credit and rapid increase in property prices in the early 2000s (Jose Doval Tedin et al. 2020). However, from 2007 to 2013, during the crisis time the house prices decreased sharply by almost 50%. As the economy recovered, the prices increased from 2013 onwards and by 2020 the rents have reached 32% higher than the previous years (Jose Doval Tedin et al. 2020). Even though the house prices have hit a new record, beyond the records during the Celtic era, several investors are considering properties as a long-term investment, which is further driving the house prices. Most of the model that predicts housing prices are not defining the investment aspect of owning a house as one of the variables that can affect the housing prices. Traditional methods have totally ignored the trends that are observed in major cities like Dublin, Manchester, Barcelona where vulture investors are investing in properties instead of companies, which is one of the highly influential factors in driving the house prices (Petrov 2009). Pandemic has shown a few less considered factors that can drive the house prices almost by 50% (Sullivan 2021). Disruptions in supply chain and unavailability of the work force has also been a factor that has to be considered while using different methods for predicting house prices. Few more factors that are revealed with pandemic that was never observed to be a factor that drives the house prices are increase in savings and the opportunity to work from home. Working from home has enabled the people to buy houses even far from cities. Different countries have reported this to be a factor to drive house prices in rural areas of the countries. Also at least a few people have considered the opportunity to work from home as a factor to consider buying houses that are far from cities but are having more facilities (Sullivan 2021). These factors have affected an indirect increase in housing prices especially in Ireland. During pandemic people were unable to spend money for holidays which accumulated into the savings and thereby increasing the buying power. Thus, those savings were converted to long term investments and a major portion of that investment was in real estate which drove the property prices even further. Furthermore, a country like Ireland which highly depends on external sources for building materials that are being brought to Ireland by using freightswill be affected and the house prices will be highly influenced by any changes that will be seen in supply chain (Gazette Desk 2021). Along with these human actions like war can drive the material cost and fuel cost that will affect the house prices and these events are highly unpredictable (MacFarlane 2022). In this scenario, it is important to know the trends in the property market and to understand the factors influencing the property prices.

Machine learning algorithms can assist us in predicting the housing prices while considering all these factors. These models can be trained using data that has all the mentioned attributes. Generally, scholars have used hedonic model, Linear Regression (LR), Random Forest (RF), and Support Vector Machine (SVM) algorithms for predicting house prices (Ja’afar et al. 2021). The data used for training machine learning models will be an important factor in determining the accuracy of the model. Historically, Ireland has large database in relation to price and traditional factors that affect the price of property. Having a model that can be trained with this data will result in better accurate predictions for the future. Along with this the same model can be expanded to be used in different countries that are having similar structures.

This research will be based on both statistical and machine learning methods to predict the property prices of Ireland from 2010- 2022. For statistical analysis I will be using hedonic regression, linear regression, and ANOVA methods. Machine learning methods such as SVM, DT, XG-Boost, and Linear regression (LR) algorithms will be employed to predict the property prices. Hedonic model is one of the most popular methods used for house price prediction and it considers the house as a combination of many attributes (Limsombunc et al. 2004). The main goal of hedonic model is to estimate the contribution of different attributes to the price of property (Montero and Fernández-Avilés 2014). SVM has been considered by many researchers since it works well with high dimensional data, unstructured and semi-structured data. Also, the outliers have less influence on the SVM, and larger amount of data can be modelled using SVM (Advantages of Support Vector Machines (SVM) n.d.). In addition to this SVM has been found very popular in commercial filed for predicting the company’s sales (Ho et al. 2020). When it comes to RF, it is been used for prediction in many applications due its ability to reduce the over fitting (Ho et al. 2020). RF also provides higher accuracy compared to other models and it can also deal with the larger datasets (Mbaabu 2020). Compared to other algorithms, DT requires less time for data preparation while pre processing and data normalization and scaling is not required for it. Both classification and regression problems can be solved using the DT (K 2019). In XG-Boost algorithm, it has the in-built ability to deal with the missing values and is faster than the gradient boosting machine (GBM). It is also referred as the regularized form of GBM because it has in-built Lasso regression (L1) and Ridge regression (L2) which reduces the overfitting (Kumar 2019). Linear regression is considered as one of the simplest algorithms and the over fitting problem in modelling can be avoided by using the regularization and cross validation techniques. It also works well with systems that has less computational power and has a noticeably lower time complexity (Waseem 2022). The research scope and goals of the research are listed in the Table 1 and the core technologies using in the research are mentioned in the Table 2.

|  |  |
| --- | --- |
| Research questions | Project Goals |
| 1. Has the property prices increased over the years? | * Find out the property prices in each year. * Find out the counties in which most of the houses are sold. * Find out the counties in which the least number of houses sold. * Find out the house prices in each county/province. * Find out the year in which most houses were sold. * Find out the trend in the property pricing over the years. * Find out either the New or secondhand dwelling has the highest prices and is it inclusive of VAT. * To analyze how the house prices depends on the type of house and its size. |
| 1. Which county has the highest prices for properties? |
| 1. Does the property price depend on which county the house is in? Is there any relationship between the counties and the prices? |
| 1. Is there any relationship between the size of the properties, type of properties (like new or second hand) and the property prices? |

Table. 1. Research questions

|  |
| --- |
| Technology |
| Programming language: Python (Jupyter Notebook, VS, Spyder) |
| Python libraries:   * Pandas * Numpy * Matplotlib * Seaborn * scipy.stats * statsmodels |
| For visualization: Seaborn, Matplotlib, Plotly |
| For version control system: GitHub. |
| Table 2. Technology used for the research |