Analysis of Ireland Property Prices

2022

By

**Ginu varghese**

Under the supervision of

Dr. Jack Mc Donnell

Logo

Description automatically generated with medium confidence

School of Informatics and Creative Arts

Department of Computing Science and Mathematics

# **Acknowledgements**

All the data used for this project was collected from the PSRA website. The data collected from this website is strictly used for research purposes without manipulating the data.

I would especially like to thank Dr. Jack Mc Donnell, the Program Director of Data Analytics program in Dundalk Institute of Technology. As my lecturer and supervisor, he has taught and helped me more than I could ever give him credit for here. He gave me suggestions and advice in each stage of my thesis and helped me to do the thesis more efficiently.

I am grateful to Dr. Fiona Lawless, the head of the Department of Computing Science and Mathematics and all the lectures from the Data Analytics program, Dr. Siobhan Connolly Kernan, Dr. Rajesh Jaiswal, Dr. Peadar Grant and Dr. Abhishek Kaushik, who provided me extensive professional knowledge and taught me a great deal about data science skills during lectures, laboratories, and projects.

This work is self-funded, and I would like to thank Dundalk Institute of Technology (DkIT) for facilitating the MSc in Data Analytics program that has helped me to work on this project.

Additionally, I am grateful to all the researchers who has done research in this area which helped to refer their works in times of doubts and the online materials which was available to learn all the technologies used in the thesis and to enhance my data science skills.

I would like to thank my parents, Varghese A.D and Lissy Varghese, whose love and guidance are with me in whatever I pursue. They are the ultimate role models.

I wish to thank my friends and siblings, who provide unending inspiration, support and love

# **Declaration**

I hereby declare that the work described in this project is, except where otherwise stated, entirely my own work and has not been submitted as part of any degree at this or any other Institute/University.

Signed:



Name: Ginu Varghese

Date: 30/08/2022

# **Abstract**

The house prices are increasing every year and it is demanding the analysis and prediction of property prices. The main objective of this study is to conduct both statistical and machine learning analysis on the property prices of Ireland from 2010 to 2022 collected from the PSRA website. Factors influencing the house prices county, year, province, location of property, property size and type are considered for analysis. Different machine learning algorithms like Linear Regression, Support Vector Machine regression, Decision Tree, XG Boost, and Random Forest regression were used for analysis to identify the better fit model. Statistical tests and visualizations were generated for the analyzing the effect of different attributes on property prices and the results implied that the province, year, property size, and location are statistically significant and are affecting the house prices significantly.

The code designed for this project is available on GitHub following the link: https://github.com/ginuvarghese16/Dissertation. And in addition to this, a screencast was recorded to

show the Python code’s functionality and structure, that is available on https://web.microsoftstream.

com/ video/ ead11ffc-78ea-4394-97c3-

# **Abbreviations**

|  |  |  |
| --- | --- | --- |
| **Abbreviation** | **Meaning** | **Page** |
| **SVM** | Support Vector Machine | **1,2,3,8,10,11, 13,14,22,27,53.** |
| **LR** | Linear Regression | **2,3,8,13,14,39,42,43,44,46,47,48,50,51,52,55,56,63** |
| **RF** | Random Forest | **1,2,3,5,6,7,8,9,10,11,13,14,15,17,20,23,24,27,28,31,39,56,57,58,60** |
| **DT** | Decision Tree | **1,3,8,13,14,17,23** |
| **XG-Boost** | eXtreme Gradient Boosting | **3,13,14,24,56,57,58,60,63** |
| **ANOVA** | Analysis of Variance | **3,26,39,40,41,42,63** |
| **COVID-19** | Coronavirus disease | **2** |
| **GBM** | Gradient Boosting Machine | **3,7,9,14** |
| **L1** | Lasso Regression | **3,14** |
| **L2** | Ridge Regression | **3,14** |
| **VS** | Visual Studio | **5,39,43,47,51,59,60** |
| **GAM** | Generalized additive models | **6,26** |
| **GWR** | Geographically Weighted Regression | **6** |
| **ML** | Machine learning | **6,9,13,25,27,31,39,42,43,44,46,50,51,52,55,56,57,58,63** |
| **K-NN** | K Nearest Neighbor | **8,13** |
| **MSE** | Mean Squared Error | **7,8,11,13,15,27,28,29,57,58,63** |
| **MAE** | Mean Absolute Error | **7,8,11,13,15,27,28,57,58** |
| **RMSE** | Root Mean Squared Error | **7,8,11,13,15,27,57,58,63** |
| **RMSLE** | Root Mean Square Logarithmic Error | **8** |
| **ANN** | Artificial Neural Network | **8,11,12,13,22,23,25,62** |
| **GA** | Genetic Algorithm | **8,10** |
| **HGA-ANN** | Hybrid genetic algorithm–artificial neural network | **8** |
| **BP** | Back Propagation | **8,10,11** |
| **LSSVM** | Least Squares Support Vector Machine | **8** |
| **PLS** | Partial Least Squares | **8** |
| **NN** | Neural network | **3,8,9,11,12,13,22,23,25,45,56,58,60,62** |
| **PCA** | Principal Component Analysis | **9,11** |
| **GM** | Grey Algorithm | **10** |
| **PSO** | Particle Swarm Optimization | **10** |
| **BPNN** | Back Propagation Neural Network | **11** |
| **RBF** | Radial Basis Function | **11** |
| **CRISP-DM** | Cross-Industry Standard Process for Data Mining | **15,16** |
| **PSRA** | Property Services Regulatory Authority | **4,17,63** |
| **CSO** | Central Statistics Office | **17** |
| **API** | Application Programming Interface | **17,61** |
| **ISM** | Interpretative structural model | **14,62** |
| **GDPR** | General Data Protection and Regulations | **61** |
| **HTTPS** | Hypertext Transfer Protocol Secure | **61** |
| **VAT** | Value Added Tax | **4,19** |
| **MLP** | Multilayer Perceptron | **25** |
| **KDE** | K Desktop Environment | **22** |

# **Table of Contents**

[**Acknowledgements** i](#_Toc112406374)

[**Declaration** ii](#_Toc112406375)

[**Abstract** iii](#_Toc112406376)

[**Abbreviations** iv](#_Toc112406377)

[**Table of Contents** vii](#_Toc112406378)

[**List of Tables** viii](#_Toc112406379)

[**List of Fgures** viii](#_Toc112406380)

[**Chapter 1** 1](#_Toc112406381)

[**Introduction** 1](#_Toc112406382)

[**Chapter 2** 6](#_Toc112406383)

[**Literature Review** 6](#_Toc112406384)

[**Chapter 3** 15](#_Toc112406385)

[**Life Cycle Of Project** 15](#_Toc112406386)

[**Chapter 4** 17](#_Toc112406387)

[**Data Exploration** 17](#_Toc112406388)

[**4.1** **Data Collection** 17](#_Toc112406389)

[**4.2** **Data Preparation** 17](#_Toc112406390)

[**4.3** **Data Description** 18](#_Toc112406391)

[**4.4** **Data Splitting and Scaling** 20](#_Toc112406392)

[**Chapter 5** 21](#_Toc112406393)

[**Methods** 21](#_Toc112406394)

[**5.1** **Scikit Learn** 21](#_Toc112406395)

[**5.2** **Keras libaray** 21](#_Toc112406396)

[**5.3** **Linear Regression** 22](#_Toc112406397)

[**5.4** **Support Vector Machine** 23](#_Toc112406398)

[**5.5** **Decision Tree** 23](#_Toc112406399)

[**5.6** **Random Forest** 24](#_Toc112406400)

[**5.7** **XG-Boost** 25](#_Toc112406401)

[**5.8** **Hedonic Regression** 25](#_Toc112406402)

[**5.9** **Neural Network Regression** 26](#_Toc112406403)

[**5.10** **ANOVA Test** 26](#_Toc112406404)

[**Chapter 6** 28](#_Toc112406405)

[**Evaluation Methods** 28](#_Toc112406406)

[**6.1** **Root Mean Squared Value** 28](#_Toc112406407)

[**6.2** **Mean Absolute Value** 28](#_Toc112406408)

[**6.3** **Mean Squared Error** 29](#_Toc112406409)

[**6.4** **R-Squared Error** 29](#_Toc112406410)

[**6.5** **Coefficient Std Error** 30](#_Toc112406411)

[**6.6** **Coefficient-t value** 30](#_Toc112406412)

[**6.7** **Coefficient Pr(>|t|)** 30](#_Toc112406413)

[**6.8** **Residual Standard Error** 31](#_Toc112406414)

[**6.9** **Multiple R-squared and Adjusted R-squared** 31](#_Toc112406415)

[**6.10** **F-statistic** 31](#_Toc112406416)

[**6.11** **p-value** 31](#_Toc112406417)

[**Chapter 7** 32](#_Toc112406418)

[**Results And Discussions** 32](#_Toc112406419)

[**7.1** **Visualizations** 32](#_Toc112406420)

[**7.2** **Statistical Analysis** 40](#_Toc112406421)

[**7.2.1** **ANOVA Test** 40](#_Toc112406422)

[**7.2.2** **Hedonic Pricing using MLR** 43](#_Toc112406423)

[**7.2.3** **SLR on Dublin Data** 47](#_Toc112406424)

[**7.2.4** **MLR on Dublin Data** 51](#_Toc112406425)

[**7.3** **ML Models** 56](#_Toc112406426)

[**Chapter 8** 63](#_Toc112406427)

[**Ethical Considerations** 63](#_Toc112406428)

[**8.1** **Data storage, security, and responsible data stewardship** 63](#_Toc112406429)

[**8.2** **Data hygiene and data relevance** 63](#_Toc112406430)

[**8.3** **Identifying and addressing ethically harmful data bias** 63](#_Toc112406431)

[**8.4** **Harms to Privacy and Security** 63](#_Toc112406432)

[**8.5** **Harms to Fairness and Justice** 64](#_Toc112406433)

[**8.6** **Harms to Transperancy and Autonomy** 64](#_Toc112406434)

[**Chapter 9** 65](#_Toc112406435)

[**Conclusion And Futurework** 65](#_Toc112406436)

# **List of Tables**

[Table 1. 1 Research questions 4](#_Toc112409725)

[Table 1. 2 Technology used for the research 5](#_Toc112409726)

[Table 4. 1 Variables in the dataset 19](#_Toc112408789)

[Table 7.2.1. 1 Two-way ANOVA interpretation 42](#_Toc112408833)

[Table 7.2.2. 1 MLR Results interpretation 47](#_Toc112408843)

[Table 7.2.3. 1 SLR Results interpretation 51](#_Toc112408849)

[Table 7.2.4. 1 MLR Results interpretation for Dublin 56](#_Toc112408859)

[Table 7.3. 1 Performance metrices of Linear Regression 58](#_Toc112408866)

[Table 7.3. 2 Performance metrices of Decision Tree Regression 58](#_Toc112408867)

[Table 7.3. 3 Performance metrices of Support Vector Regression 58](#_Toc112408868)

[Table 7.3. 4 Performance metrices of XG-Boost Regression 58](#_Toc112408869)

[Table 7.3. 5 Performance metrices of Random Forest Regression 59](#_Toc112408870)

[Table 7.3. 6 Performance metrices of NN Regression 59](#_Toc112408871)

# **List of Figures**

[Figure 3. 1 Life Cycle of Project based on CRISP-DM 17](#_Toc112416634)

[Figure 7.1. 1 Univariate for Price 34](#_Toc112416642)

[Figure 7.1. 2 Univariate for Province 34](#_Toc112416643)

[Figure 7.1. 3 Univariate for Location 34](#_Toc112416644)

[Figure 7.1. 4 Univariate for Month 34](#_Toc112416645)

[Figure 7.1. 5 Univariate for County 34](#_Toc112416646)

[Figure 7.1. 6 Univariate for VAT\_exclusive 34](#_Toc112416647)

[Figure 7.1. 7 Univariate for Year 34](#_Toc112416648)

[Figure 7.1. 8 Univariate for FMP 34](#_Toc112416649)

[Figure 7.1. 9 Univariate for property size 35](#_Toc112416650)

[Figure 7.1. 10 Univariate for property description 35](#_Toc112416651)

[Figure 7.1. 11 Year-price bivariate 35](#_Toc112416652)

[Figure 7.1. 12 Year-location bivariate 35](#_Toc112416653)

[Figure 7.1. 13 Year-property description bivariate 36](#_Toc112416654)

[Figure 7.1. 14 location-price bivariate 36](#_Toc112416655)

[Figure 7.1. 15 Province-price bivariate 36](#_Toc112416656)

[Figure 7.1. 16 Property description-price bivariate 36](#_Toc112416657)

[Figure 7.1. 17 Price-property size bivariate 36](#_Toc112416658)

[Figure 7.1. 18 location-property size bivariate 36](#_Toc112416659)

[Figure 7.1. 19 Multivariate1  37](#_Toc112416660)

[Figure 7.1. 20 Multivariate2 37](#_Toc112416661)

[Figure 7.1. 21 Multivariate3 37](#_Toc112416662)

[Figure 7.1. 22 Multivariate4 37](#_Toc112416663)

[Figure 7.1. 23 Multivariate5 37](#_Toc112416664)

[Figure 7.1. 24 Multivariate 37](#_Toc112416665)

[Figure 7.1. 25 Maximum property prices in counties. 38](#_Toc112416666)

[Figure 7.1. 26 Minimum property prices in counties 38](#_Toc112416667)

[Figure 7.1. 27 Property sizes and median property prices 39](#_Toc112416668)

[Figure 7.1. 28 Property types and median property prices 39](#_Toc112416669)

[Figure 7.1. 29 Property prices and median property prices in counties 40](#_Toc112416670)

[Figure 7.1. 30 Postal codes and median property prices 40](#_Toc112416671)

[Figure 7.1. 31 Residuals Vs Fitted values Plot 41](#_Toc112416672)

[Figure 7.1. 32 Probability Plot 41](#_Toc112416673)

[Figure 7.1. 33 ANOVA results for full model 42](#_Toc112416674)

[Figure 7.1. 34 Model summary for two-way ANOVA 44](#_Toc112416675)

[Figure 7.1. 35 Residuals Vs Fitted values Plot 45](#_Toc112416676)

[Figure 7.1. 36 Probability Plot 45](#_Toc112416677)

[Figure 7.1. 37 Histogram for MLR 45](#_Toc112416678)

[Figure 7.1. 38 Model summary for MLR 46](#_Toc112416679)

[Figure 7.1. 39 Residuals Vs Fitted values Plot 49](#_Toc112416680)

[Figure 7.1. 40 Probability Plot 49](#_Toc112416681)

[Figure 7.1. 41Histogram for SLR 49](#_Toc112416682)

[Figure 7.1. 42 Model summary for SLR 50](#_Toc112416683)

[Figure 7.1. 43 Residuals Vs Fitted values Plot 53](#_Toc112416684)

[Figure 7.1. 44 Probability Plot 53](#_Toc112416685)

[Figure 7.1. 45 *Histogram* 53](#_Toc112416686)

[Figure 7.1. 46 *Model Summary for MLR* 54](#_Toc112416687)

[Figure 7.3. 1 Correlation Matrix……………………………………………………58](#_Toc112416830)

[Figure 7.3. 2 Actual Vs Predicted Values in Linear Regression 61](#_Toc112416831)

[Figure 7.3. 3 Actual Vs Predicted Values in DC Regression 61](#_Toc112416832)

[Figure 7.3. 4 Actual Vs Predicted Values in SV Regression 61](#_Toc112416833)

[Figure 7.3. 5 Actual Vs Predicted Values in XG-Boost Regression 62](#_Toc112416834)

[Figure 7.3. 6 Actual Vs Predicted Values in RF Regression 62](#_Toc112416835)

[Figure 7.3. 7 Actual Vs Predicted Values in NN Regression 62](#_Toc112416836)

# **Chapter 1**

# **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 such as Support Vector Machine (SVM), Random Forest (RF), and Decision Tree (DT) 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 tiger years (FitzGerald 2007), 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). COVID-19 pandemic has shown a few factors that can drive the house prices 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. COVID-19 pandemic has provided the opportunity to work from home that caused less expenses in daily commute, coffee, and other expenditures that in turn added to the savings for buying houses. When people started working from home, proper workspace became a necessity. This has led to owning houses farther from cities but are with better facilities. 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 have more facilities (Sullivan 2021). These factors have affected an indirect increase in housing prices, especially in Ireland. During the COVID-19 pandemic people were unable to spend money on holidays, dining out, outdoor activities, entertainments like movies and concerts, and instore shopping 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 freightwill be affected and house prices will be highly influenced by any changes that will be seen in supply chain (Gazette Desk 2021). Along with these conflicts between countries 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 a key 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 analyse and 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, RF, NN, 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 extremely popular in commercial field for predicting the sales of the company (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 will be based on the property sales data of Ireland from 2010 – 2022 and it is collected from the Residential Property Price Register page of the PSRA website. The research scope and goals of the research are listed in the Table 1 and the core technologies used in the research are mentioned in Table 1.1.

|  |  |
| --- | --- |
| **Research questions** | **Project Goals** |
| 1. **What is the trend in the property prices over the years?** | * Find out the property prices in each year. * Find out the counties in which most and least houses are sold. * Find out the house prices in each county/province. * Find out the year in which most houses were sold. * To 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. * To estimate the different factors affecting the property prices. * To evaluate the effect of postcodes in determining the property prices. * To determine the impact of latitude and longitude on the property prices. |
| **2. Which County has the highest demand for properties?** |
| **3. Does the property price depend on which county the house is in? Is there any relationship between the counties and the prices?** |
| **4. Is there any relationship between the size of the properties, type of properties (like new or second hand) and the property prices?** |
| **5. What are the factors contributing to the price of the properties?** |
| **6. Is the post codes are impacting the house prices?** |
| **7.How does the latitude and longitude effect the property prices?** |

Table 1. Research questions

|  |
| --- |
| **Technologies Used** |
| Programming language: Python (Jupyter Notebook, VS, Spyder) |
| Python libraries:   * Pandas * Numpy * Matplotlib * Seaborn * scipy.stats * statsmodels * Plotly * Geopy |
| For visualization: Seaborn, Matplotlib, Plotly |
| For version control system: GitHub. |

Table 1. Technology used for the research

This paper is structured as follows; chapter 2 gives the related works, the life cycle of the project is described in chapter 3, data exploration is presented in chapter 4, and methods used in the research are mentioned in chapter 5. Chapter 6 describes the different evaluation methods used in the project and the chapter 7 focuses on the results and discussion which demonstrates the outcome of the initial and final analyses performed. The ethical considerations related to the research are mentioned in chapter 8, then conclusion and future work is given in chapter 9 and finally references are mentioned in chapter 10.

# **Chapter 2**

# **Literature Review**

Several studies have been performed in the past to analyze and predict the real estate and residential property prices to aid the customers as well as the developers. Hurley & Sweeney (2022) studied the impact of post codes and address in Irish property prices based on the data from January 2018 to November 2018. They focused on analyzing the property prices in Dublin with the dataset of 5028 properties for the development of geospatial statistical models where the post codes and addresses are being mislabeled. They also performed text mining to create additional variables that describes the features of the properties and its surroundings. A spatial hedonic regression model was used to separate the spatial and non-spatial contributions of property features to the resale value. Generalized additive models (GAM), regression kriging and Geographically Weighted Regression (GWR) have also been used since these methods provide greater interpretability with smaller data requirements. Three different Machine learning (ML) methods like Decision Tree, Random Forest and K-nearest neighbor algorithms were also applied on the data to evaluate how these ML models perform with smaller datasets.

Their models shown a reduction in median absolute percentage error with an increasing model complexity i.e., 12% for the hedonic model and 9.6% for linear model with spatial surface. And the authors also stated that although ML models are widely used for property prices prediction, they did not have the probability-based uncertainty intervals and the interpretability of the statistical spatial models. They also added that the random forest models may not fit well with the areas outside Dublin since the housing turnover is low outside Dublin and in that case statistical spatial modelling can work more efficiently. In Ireland, where property valuations are currently based on comparison to recently sold neighboring properties, the authors claim their model has higher applicability and will help to improve property tax computations and site value estimates because the model is not only based on the property value but uses the spatial location scaling (Hurley et al. 2022).

Machine learning models were used to predict house prices in Godavari district of Andhra Pradesh, India by Thamarai and Malarvizhi (2020). The models were built to help the people buy suitable houses for their needs. Decision tree regression, decision tree classification and linear regression models were performed on the data based on the attributes of the property like number of bedrooms, age of the house, availability of school near the house and shopping malls available nearby the house location. Attribute selection algorithms were used to remove the redundant features to reduce the impurity in the process before splitting the data for modelling.

To predict the availability of houses according to the requirement of the user, they used the decision tree classifier which gives responses like yes or no to show whether a house is available or not. Along with this, regression methods like decision tree regression and multiple linear regression were used to predict the prices of the houses. The dataset used for the modelling was a real-time data acquired from the Godavari district with all the attributes of the house and modelled using the scikit learn, a machine learning tool.

The main dataset was divided into train and test data and the decision tree classifier is performed using the training dataset. The accuracy of the model is then checked using the test data. Mean Squared Error (MSE), Mean Absolute Error (MAE) and root mean squared error (RMSE) were used to evaluate the performance of both the classification and regression models. The house price prediction with decision tree algorithm produced an output with some data record prices predicted with lesser deviations. From the multiple linear regression, it is found that the number of bedrooms is the feature having high influence on house price and age of the house is the feature having less influence on house price. From the performance metrics it is found that the prediction of house price using multiple linear regression has higher performance than the prediction using the decision tree regression. The authors also stated that the developed model can be used to predict the availability and prices of houses for any new attributes according to the users and the overall accuracy can be increased in the future with a large dataset and by identifying the best features.

Quang et al. (2020) conducted a study to compare housing prices prediction using traditional and advance machine learning methods. They used three different machine learning models: Random Forest, XGBoost and LightGBM. Further, two machine learning techniques are used, Hybrid Regression and Stacked Generalization Regression for prediction. For the analysis, the Beijing house price data from 2009 and 2018 was used and feature engineering was performed to select the appropriate features. In addition to this, exploratory data analysis was done to discover the patterns in the data. The machine learning models were applied on the training data which is split from the actual data and Root Mean Square Logarithmic Error (RMSLE) was used for evaluating the performance of the models. Results from the study shows that the Random Forest method is prone to overfitting even though it has lowest errors. The Hybrid regression method is better performing among all three methods. They also stated that although the Stacked regression method has the worst time complexity, it is the best choice when accuracy is considered(Quang et al. 2020).

To understand the factors influencing the property prices, Decision Tree (DT), Random Forest (RF) and linear regression (LR) methods were used by (Yee et al. 2021) using the Malaysian dataset from April 2017 to December 2019, to help the buyers and sellers who need to finance in the property market. The accuracy of the models is evaluated based on the R squared, RMSE and MAE values. Natural Language Processing was used to transform the data so that it is readable by the machine. The outcome of the study explained that the RF is the better model with higher accuracy compared to the DT and LR.

Alen Ihre (2019) discussed the machine learning algorithms K-Nearest Neighbour (K-NN) and Random Forest (RF) regression to predict the house prices using the Ames dataset of 3000 observations. Five-fold cross validation was performed on the dataset to minimize the bias and grid search algorithm was utilized to select the best number of hyperparameters for the prediction. Finally, the RF was found to be the best performing model based on the MAE values. However, there is small differences in the actual price and the predicted price, and the author suggested that the results could be improved by using a larger and less biased dataset (Ihre 2019).

Artificial Neural Network (ANN) is used for analysing the real estate price by (Shi and Li 2009) to evaluate the house price determinants. An improved Genetic Algorithm (GA) was used to optimize the weights of the neural network. The results of the study shown that the GA-ANN is more capable of determining the house price determinants more time efficiently and the errors of the HGA-ANN model was found to be lower than the back propagation (BP) and the genetic algorithm (GA) models.

Support Vector Machine (SVM) has been employed for predicting and forecasting house prices by Mu et al. (2014), Phan (2018), Ho et al. (2020), Chen et al. (2017), Gu et al. (2011) and Wang et al. (2014).

House value forecasting-based machine learning methods by Mu et al. (2014) aimed at helping the developers and government to take decisions regarding developing real estate in the Boston area. The authors collected the data from the UCI data sets and support vector machine (SVM), least squares support vector machine (LSSVM), and partial least squares (PLS) algorithms were applied on the training data to predict the housing value. From the prediction results it is found that the SVM and LSSVM has better efficiency with the nonlinear data. PLS algorithm is better for linear data due to the simplicity of the algorithm. They also added that to achieve best forecasting effect and an optimal solution, SVM can be used (Mu et al. 2014).

Phan (2018) utilized the SVM technique to predict the house prices in the Melbourne city of Australia to help the house buyers and sellers. Neural network (NN), Polynomial regression, Linear regression and Regression Tree models were also developed along with SVM to identify the best fit. The data used in the study was downloaded from the Kaggle website and it has the house sold houses transaction from 2016 to 2018. Data imputation and descriptive analysis techniques were performed on the data prior to modelling. Along with this, principal component analysis (PCA) was also performed to select the desired features and stepwise method was utilized for subset selection. The results shown that the SVM with the subset selection method gives the best efficiency with lower errors. When regression tree and linear regression delivered almost equal prediction result, the polynomial regression gave better accuracy with lower errors. The neural network seemed to be not working well with the available dataset. The authors also stated that regression tree and neural network worked more faster than the SVM, where PCA with SVM took more time than SVM with stepwise (Phan 2018).

SVM methods were employed by Ho et al. (2020) for predicting the property prices in Hong Kong. Random forest (RF) and Gradient Boosting Machine (GBM) were also utilized along with SVM to compare the performance of algorithms. The dataset used was a sample data with over 40000 housing transactions in a time of 18 years. Correlation matrix was used to determine the features which should be included in the models. The results from the performance metrices revealed that the RF and GBM were able to estimate the house prices better than the SVM with smaller errors. They also found that the ML algorithms need more computation time than the traditional Hedonic pricing model and among the three models used, SVM is the better choice for forecasting when speed is the priority and RF and GBM should be used if the accuracy is considered (Ho et al. 2020).

To predict the housing prices in the Taipei city of Hong Kong, SVM models were implemented by Chen et al. (2017). By using stepwise multi regression, the support vectors were found, and a SVM hedonic price model was built using the support vectors, the structural and the spatial variables to predict the house prices in the Taipei city. The SVM model is then developed based on the identified support vectors to forecast the future housing prices for the data from 2007 to 2010. One of the advantages in using SVM is that it is not depending on the probability distribution assumption and hence it could plot the input variables into a high dimensional feature space. To compensate the bias variance trade-off, five-fold cross validation has been used for testing and training in the analysis. The outcome from the study points out that the SVM can be considered as a superior approach which legitimise the issues in the multiple regression analysis and combining the hedonic approach with SVM is feasible for non-linear modelling (Chen et al. 2017).

Gu et al. (2011) used the SVM methods along with hybrid genetic algorithm methods to forecast the house prices in China. SVM has proven to be one of the best algorithms in both classification and regression in lots of applications. In the study, genetic algorithm (GA) has been used instead of grid algorithm to optimize the parameters of the SVM since, GA is more time efficient and the G-SVM is developed. The results of the study revealed that the G-SVM method giving more accuracy than the Grey Algorithm (GM) which is used in the past to predict the house prices (Gu et al. 2011).

Machine learning algorithms has used to forecast the real estate prices by many researchers and Wang et al. (2014) used SVM to forecast the real estate price with particle swarm optimization (PSO) in the Chongqing city in China. One of the reasons to choose SVM is its ability to conquer the ‘Curse of dimensionality’. To identify the parameters of SVM, PSO method is used instead of GA and grid algorithm since it is easy to enforce. The actual data was divided into train and test data for the modelling. The study shown that the PSO-SVM model is performing better than the BP neural network methods used by other researchers.

PSO has also used by Alfiyatin et al. (2017) for predicting the house prices to help the builders to decide the selling price of the house and to help the buyers to set the right time to buy the house. PSO and regression analysis was implemented on the houses data in the Malang city of Indonesia within 2014-2017. Hedonic regression was chosen as the regression prediction model and PSO to select the appropriate features. The error prediction values are found to be higher for the regression model compared to the PSO-regression model. Hence the study proved that the combination of PSO with regression can give the minimum prediction error.

Support Vector Machine Regression (SVR) has used by several researchers to predict the real estate and housing prices due to its efficiency and application. A SVR model was used by Li et al. (2009) to understand the possibility of predicting real estate prices in China from 1998-2008. The results of the model are compared with the Back Propagation Neural Network (BPNN) model to analyze the performance of the model. Several indicators like CPI, loan interest rate, real estate price, real estate investment, income etc. were used to forecast the real estate prices. The analysis demonstrated that the SVR model works better than the BPNN model for real estate forecasting based on the MAE, MAPE and RMSE. The study also proved that the SVR method is an efficient approach to forecast the real estate price (Li et al. 2009).

SVM regression with Gaussian kernel was utilized by Miao et al. (2021) to predict the property prices in Boston area. The aim of the study was to help the people to estimate the price of the house based on the properties of the house. They selected the most important features using the decision tree with ID3 algorithm, divided the data to train and test data and SVR is employed on the data with the gaussian kernel to predict the prices and compared the model with different regression methods to analyze the performance. The study proved that the SVR with gaussian kernel has more efficiency than the KNN, decision tree and SVR with linear kernel. But the model still has some drawbacks which are mentioned by the authors and the important one is that some of the factors that impact the house prices cannot be measured such as the cultural tolerance of the neighbors and so on.

Jiao Yang Wu (2017) also performed a study to analyse the house price prediction using SVR based on the housing sales data of Kings County, USA with an aim to help the buyers and sellers. Feature selection methods like Random Forest selector, Lasso ridge and Recursive Feature Extraction and the feature extraction method PCA is done prior to building the model. Further to this, parameter tuning, and transformation techniques have been used to improve the accuracy of the model. But the results from the experiments shown that there is not much difference in the performance of the model with feature extraction and feature selection. The Radial Basis Function (RBF) kernel with SVR was found to be the best one among the performed combinations. The author also pointed out that in future other machine learning models like XGBoost and other feature engineering methods can be applied on the data (Wu 2017).

Hedonic Regression is one of the estimation and prediction method preferred by the researchers when it comes to prices. It is used in most of the scenarios when a price variable is to be considered. Property prices in Croatia is studied by Kunovac and Zagreb (2019) using Hedonic regression based on the data collected from different sources. One of the goals of the research was to propose how the hedonic models can be used for the evaluation of residential property. The hedonic model built allows the evaluation of some attributes of the property such as age, location and so on. The results of the analysis shown that the micro location of the property should also be considered to the hedonic models and to commonly used other models to improve the prediction of residential house prices.

Abdulai and Owusu-Ansah (2011) used hedonic regression to determine the house price determinants in Liverpool, The United Kingdom over a period of 18 years from 1990-2008. They also analysed how the past and present buyers valued the property features. The regression model was able to explain almost 75% of variation in the housing prices and all the variables in the dataset was found to be statistically significant. Another thing they found was that the price of the new properties is almost double than the old properties and the detached houses are more expensive than the flats. When the past buyers before 1999 focused more on the number of bedrooms, bathrooms and detached houses, the buyers after 2000 more value the number of floors, gardens, and showers (Abdulai and Owusu-Ansah 2011).

Selim (2008) identified the house price determinants in Turkey using hedonic regression using the 2004 household survey data. Environmental factors were not included in the data for analysis and natural logarithm of the house price is treated as the dependent variable. To estimate the hedonic model, ordinary least square method is utilized. The heteroscedasticity present in the model was eliminated using the White’s heteroscedasticity consistent coefficient covariance matrix. And the results of the analysis shown most of the variables to be significant and the house prices seems to be higher for houses with more rooms. The variables such as water system, pool, type of the house, number of rooms, size of house, locational attributes and building structure are the found to be the most significant variables that influence the house prices.

Limsombunc et al. (2004) compared the hedonic regression with Artificial Neural Network (ANN) model on house price prediction based on a sample dataset of 200 New Zealand houses. The age of the house, size of house, bedrooms, bathrooms, and other features were considered for the experiment. The heteroscedasticity consistent coefficient covariance matrix and weighted least squares method were used instead of the ordinary least square method to eliminate the heteroscedasticity issues. However, the heteroscedasticity problem was not completely removed. The results from the study revealed that even though the hedonic model was able to explain almost 70% of variance in the model, it did not outperform the neural network model. The authors mentioned that the small dataset and the lack of environmental features may be some reasons for the poor performance of the hedonic model.

Following researchers have used supervised and semi-supervised ML models for predicting property and real estate prices (Ihre 2019; Phan 2019; Quang et al. 2020; Mu et al. 2014; Ho et al. 2020; Chen et al. 2017; Gu et al. 2011; Wang et al. 2014; Pow and Janulewicz 1995; Alfiyatin et al. 2017; Li et al. 2009) and have employed the RMSE, MAE, MSE, MAPE, and R Squared performance metrices for evaluating the models. Either regression or classification supervised learning techniques are utilized by most scholars for prediction. There are certain algorithms which are preferred by several scholars such as the RF, DT, LR, K-NN, XG-Boost, SVM and ANN algorithm (Ja’afar et al. 2021).

Factors affecting the house prices have been identified by several researchers and (Rahman et al. 2012) explored the factors affecting the house prices in the Chinese city, Hangzhou. The paper used the Ordinary Least Square (OLS) method by taking time series data of several variables from 1990-2009. It is noticed that household’s income, lagged sum of annual real estate development investment, the urbanization rate, and the foreign direct investment in Hangzhou positively and considerably affect Hangzhou’s housing price. The saving deposits of urban residents in Hangzhou have significant negative effect on the housing price. These results were in line with the theoretical predictions. They also mentioned that they could not include some important variables in the model due to lack of data such as the interest rate, expected return on housing investment, housing tax and capital gain tax, land release data, etc. Hence, they suggested that future research will be helpful by including these variables and with a long time period, to obtain better research results.

(Antonina Mavrodiy 2005) analysed the determinants of real estate prices in the Kiev city of Ukraine using regression methods and confirmed that different micro and macro factors are affecting the property prices. From the analysis the found out that the GDP, income level, population rate is having a direct impact on the property prices. On the other hand, the interest rate is having a negative impact on the real estate market. He also found out that different geographical factors such as the location of the property, different facilities surrounding the property such as schools, metro station, hospital are having higher effect on the property prices along with the property size and flooring details.

Different factors responsible for the housing prices in China are evaluated by (Gao et al. 2018) by using the theory of interpretative structural model (ISM). They selected 50 real estate residential price factors to build housing price factors binary relation on set and quantitative analysis matrix based on the document analysis of the residential real estate prices. ISM model for residential real estate prices were build using standardized methods and the results showed that, property management companies, brand value, population structure, education welfare state, political stability, disposable income growth, economic growth, land policy, real estate tax policy, living facilities, convenient living and other factors, building properties, location are the fundamental factors affecting the real estate prices and they have significant effect on determining the prices.

From the previous studies the most preferred algorithm is the RF. It is one of the ensemble methods used in machine learning and it is generally utilized for predicting the real estate prices due to its less error compared to other algorithms (Shinde and Gawande 2018). Another most chosen algorithm is SVM, and it is applied in both regression and classification problems. Since SVM can better deal with bias and can use high dimensional data, it is used as recognition in classification problems (Ja’afar et al. 2021). DT, XG-Boost and LR are also preferred by the researchers because, 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). From the previous studies (Gao et al. 2018); (Antonina Mavrodiy 2005; Rahman et al. 2012) it is evident that the location, the property size, and its other facilities are having a great impact on the property prices.

# **Chapter 3**

# **Life Cycle Of Project**

The project is based on the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology to indicate Machine Learning aspects and ensure a successful execution of the project. The whole life cycle of the project consists of six stages: data collection, data preparation, data exploration, data modelling, evaluation, and deployment.

* Data Collection: The data was collected from the Residential Property Price Register website, which is available to the public for download. The data has the property selling details of Ireland from 2010 - 2022 with the address, post codes, county, prices, and several other attributes.
* Data Preparation: Data preparation involves the cleaning of the data including removing the data recorded in the Irish language, dealing with the missing values, and adding extra data such as province and latitude/longitude details to the available data.
* Data Exploration: Exploratory data analysis and summary statistics were performed to explore the patterns in the data. Data visualizations were performed on the data using different plots to understand the trend of the house pricing.
* Modelling: Statistical and machine learning techniques were employed on the data to predict the house prices. Dependent variable was selected from the data prior to modelling.
* Evaluation: To evaluate the models, performance metrices like MSE, RMSE and MAE were used and along with this, graphs were used to identify the patterns and outliers in the data.
* Deployment: In deployment stage, the review of the project is performed, and future improvements are suggested.

Diagram

Description automatically generated

Figure 3. Life Cycle of Project based on CRISP-DM

# **Chapter 4**

# **Data Exploration**

## **Data Collection**

The data is collected from the Residential Property Price Register page of PSRA website which provides the residential property sales data of Ireland since January 2010. The data is collected as csv files for the ease of the analysis. Prior to this, an attempt was made to collect data from daft.ie, which is a property sales application of Ireland. Unfortunately, it failed since they were not willing to provide their data to a third person. The data from Kaggle and CSO was found to be insufficient due its size and incomplete information. Finally, the data is collected from the <https://propertypriceregister.ie>. website which has sufficient size and information for the analysis.

## **Data Preparation**

Data wrangling is performed on the collected data and outliers and missing values are treated for further analysis. To expand the actual data, province, latitude & longitude, and location columns were added. The province details were collected from the ‘IrelandTownList’ website as a csv file, and it is then merged to the actual dataset. To access the latitude and longitude, an API is needed, and it was created through the <https://opencagedata.com/> which is a free and secured website for creating free and paid API keys. The first option for creating API key was Google API but it is paid and not affordable so, I couldn’t use it. Python Geopy was also tried for getting the latitude and longitudes however, the Jupiter Notebook was not able to process all the data and hence that attempt was also in vain. Finally, API is created through ‘opencagedata’ website, and the latitude and longitude of each county are accessed and merged to the data. This is because, only 2000 requests can be processed in a single a day using the free API and the dataset has more than 500000 rows which makes it impossible to process the entire data. Another column ‘location’ is added to the data which contains the details of the place of sale either as ‘Dublin’ or ‘Outside Dublin’ for the analysis. Unemployment, income, and homelessness data were supposed to be added to the actual data for further analysis unfortunately, these data were not available for all the years since 2010 and as per county and thus those details were not considered for analysis.

The combined data was then cleaned using Python pandas library and missing values are treated. There are many missing values in the data, but they are kept same since removing them will affect the entire analysis and modelling. The duplicates were checked, datatypes were corrected and Irish text in the data were converted to English using Python. The address values were changed to string title as some of the values were in uppercase. Additional columns were created for date variable with month and year for further analysis and visualization.

## **Data Description**

The dataset contains the details of the residential property sales of Ireland from 2010 - 2022. It was collected from the Residential Property Prices Register website and has 516586 rows × 9 columns. The data has the date of the sale, the price of the property, county, and other related information about the sale. After removing duplicates and adding the additional columns there were 515792 rows × 15 columns. Table. 3 describe the variables in the dataset. The variables month, year, province, location, latitude, and longitude were additionally created for the analysis.

|  |  |  |  |
| --- | --- | --- | --- |
| **Type of Variables** | | | |
| **Variable Name** | **Category** | **Type** | **Description** |
| **date\_of\_sale** | Date | Datetime | The date of the property sale |
| **address** | Nominal Categorical | String | The address of the sold property |
| **postal\_code** | Nominal Categorical | String | The postal code of the sold property |
| **county** | Nominal Categorical | String | The county name of the sold property |
| **price** | Continuous Numerical | Float | The price of the sold property |
| **FMP** | Nominal Categorical | String | The information about whether the sold property price is full market price or not. |
| **VAT\_exclusive** | Nominal Categorical | String | The information about whether the sold property price is VAT exclusive or not. |
| **property\_description** | Nominal Categorical | String | The type of property. |
| **property\_size\_description** | Nominal Categorical | String | The size of the property. |
| **province** | Nominal Categorical | String | The province name of the sold property |
| **lat** | Continuous Numerical | Float | Latitude of the property. |
| **lon** | Continuous Numerical | Float | Longitude of the property. |
| **location** | Nominal Categorical | String | The information about whether the sold property is in Dublin or outside Dublin. |
| **year** | Discrete Numerical | Integer | The year in which property is sold. |
| **month** | Discrete Numerical | Integer | The month in which property is sold. |

Table 4. Variables in the dataset

## **Data Splitting and Scaling**

To analyse the property prices in Dublin County, the cleaned dataset was divided into a new one with the details of county Dublin. This was performed by Python techniques and there were 164027 observations and 15 variables in the Dublin dataset. The data splitting was done to mainly analyse the effect of Dublin postcodes in the property prices. Data scaling was performed using standard scaler method since it works well for regression problems. Unfortunately, scaled data seemed to give less accurate results than non-scaled data for some of the algorithms and hence scaling was performed only for SVM regression.

# **Chapter 5**

# **Methods**

To analyze and predict the property prices different machine learning as well as statistical techniques were utilized in the research and this section will give an overview of those methods. The proposed work is implemented using scikit learn, a machine learning and statistical modeling tool in Python is utilized and for building NN model, Keras library has been used.

## **Scikit Learn**

Scikit learn is a useful Python library for machine learning and it contains different tools for machine learning and statistical modelling including classification, regression, and clustering. It is developed upon the Scientific Python (SciPy) which must be installed before using scikit learn (Brownlee 2014).

The stepwise execution of a model using scikit learn is as follows (Andrade 2021).

1. Import the required libraries.
2. Load the required dataset.
3. Split the data into train and test data.
4. Fit the model into the data.
5. Predict the dependent variable for the test data.

## **Keras libaray**

**Keras** is a powerful and easy-to-use free open-source Python library which is used for developing and evaluating [deep learning](https://machinelearningmastery.com/what-is-deep-learning/)**models**. It is part of the [TensorFlow](https://machinelearningmastery.com/tensorflow-tutorial-deep-learning-with-tf-keras/) library and it helps to define and train neural network models in just a few lines of code (Javatpoint 2010).

Some of the advantages of Keras are:

* It is very easy to understand and integrate the faster deployment of network models.
* Most of the AI companies are focussed on using Keras and this gives Keras vast community support.
* Keras supports multi backend which means TensorFlow, CNTK or Theano can be used as backend according to the user requirement.
* Keras can be deployed in many devices such as, Raspberry pi, Web browser with .js support and Cloud engine etc.
* It supports data parallelism. So, it can be trained on multiple GPUs at the same time for speeding up the process (Javatpoint 2010).

## **Linear Regression**

Linear regression is a supervised machine learning algorithm which identifies the best fit linear line between the dependent and independent variable. In other words, it identifies the linear relationship between the dependent and the independent variable. There are two types of linear regression, simple linear regression, where there is only one independent variable and multiple linear regression, where there are multiple independent variables (Deepanshi 2021).

Simple linear regression equation is given by, , where is the intercept, b1 is slope, x is the independent variable and y is the dependent variable.

Multiple linear regression equation is given by, , where  is the intercept, b1,b2,b3…,bn are slopes of the independent variables x1,x2,x3,x4…,xn and y is the dependent variable.

The basic assumptions for linear regression are as follows,

1. There should be a linear relationship between the independent and dependent variable. This can be verified using scatter plot between the variables.
2. The independent and dependent variables should be normally distributed. This can be identified using the KDE plots and histograms.
3. The spread of the residuals should be constant for all values of independent variable, which can be verified using the residual plot.
4. There should be no correlation between the independent variables and the error terms should be normally distributed.
5. There should be no autocorrelation.

When the assumptions are violated, it leads to the decrease in accuracy which in turn affects the predictions and makes the high errors (Deepanshi 2021).

## **Support Vector Machine**

Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for both classification and regression problems. The main goal of the SVM is to find a hyperplane in an n-dimensional space where n is the number of features that classifies the data points with maximum distance. SVM is built using the support vectors, which are the data points that are close to the hyperplane and are used to maximize the distance (Gandhi 2018; Ray 2017).

SVM can be of two types, Linear SVM and Non-linear SVM.

***Linear SVM*:** Linear SVM is used for linear data. If a dataset can be divided into two classes using a single straight line, them such data is known as linearly separable data and SVM linear classification is used for such data (SVM Algorithm - Javatpoint 2010).

***Non-linear SVM*:** Non-linear SVM is used for non-linear data. If a data cannot be divided into classes using a straight line, that data is a non-linear data and SVM non-linear classifier is used for such data (SVM Algorithm - Javatpoint 2010.).

***Hyperplane in SVM:*** There are different lines to separate the classes in n-dimensional space and the best line or boundary that separate the classes is the hyperplane of SVM. The dimension of the hyperplane depends on the features in the dataset i.e., if there are two features the hyperplane will be a straight line and if there are three features then the hyperplane will be two-dimension plane (SVM Algorithm - Javatpoint 2010) .

***Support Vectors in SVM:*** The datapoints that are nearest to the hyperplane and that influence the position of the hyperplane are known as support vectors and hence the name support vector machine.

## **Decision Tree**

Decision Tree (DT) is a supervised machine learning algorithm which can be utilized for both classification and regression problems. It is a tree like structure with root node and then branches into solutions as like in a tree. There are root nodes, decision nodes and leaf nodes in a DT where, root nodes are the beginning node of the DT and the decision nodes are the nodes which we get after splitting the root nodes and the leaf nodes, where further splitting is not possible. In DT, the overfitting can be reduced by pruning, the cutting down of some nodes (Saini 2021; Amrutha 2022).

***Root Node:*** It is the starting point of the decision trees. It represents the dataset, and it is then divided into two or more homogenous sets. It is also known as the parent node.

***Leaf Node:*** It is the final output node, and the tree cannot be separated further after getting the leaf node.

***Splitting:*** It is the process of dividing the root node into sub-nodes based on the given criteria.

***Branch/Sub Tree:*** It is the tree formed by dividing the tree.

***Pruning:*** It is the process of getting rid of the unnecessary branches from the tree.

***Child Node:*** The nodes other than the root node are called the child nodes (javatpoint 2010).

## **Random Forest**

It is a supervised learning algorithm used for both classification and regression problems. It has number of decision trees on subsets and takes the average of it to predict the accuracy. The higher the number of trees in the model, the lower the chance of overfitting (Yiu 2019).

***Assumptions of RF:*** The RF joins multiple trees to predict the class of the data. This makes some decision trees to predict the correct output and others to not. But when combined all the trees predict the correct output. Hence, there are some assumptions for the RF to predict the better output.

* There should be some actual values in the feature variables of the dataset for the classifier to predict the correct results.
* There should be low correlations between the predictions from each tree.

RF takes less time than other algorithms and it can predict outputs with higher accuracy for larger datasets and even with large number of missing values (javatpoint 2010).

## **XG-Boost**

Extreme Gradient Boosting or XG-Boost is an ensemble learning method. It is an extension of gradient boosted decision trees and are designed to improve the performance of the model. XG-Boost has the regularized learning feature, which helps to reduce the overfitting (analyticsvidhya 2018).

The boosting ensemble technique has mainly 3 steps,

1. A primary model F0 is defines for predicting the target dependent y. This will be related to the residual (y-F0)
2. New model h1 is fit to the residuals from the step1.
3. The boosted version (F1) of F0 is obtained by combining F0 and h1. Similarly, new models are created after each residual to improve the performance. The mean squared error of each new model will be lower than previous model. The residuals can be minimized through m iterations (analyticsvidhya 2018).

## **Hedonic Regression**

Hedonic regression is a regression method which analyses the impact of different attributes on the price of a good. In hedonic regression, the dependent variable will be the price of the good and the independent variables will be the attributes of the good that influence the price. It uses the ordinary least squares and other techniques to estimate how the attributes affect the price of real estate like house.

Hedonic regression function is defined as, pi = j (ci), where p is the price of good i, and ci is the vector of attributes related to the good. The attributes can be location, structure of the property, environmental properties, and accessibility to the property. The hedonic regression is mostly used to estimate the property prices in real estate industry (investopedia 2021; CFI 2022).

The main uses of hedonic models are:

* The hedonic regression method can be used to calculate property and other asset values based on actual choice of the customer.
* It is probably the most efficient method for making use of the available data.
* The use of hedonic regression methods in real estate is effective, as the property sales data is easily available, and other secondary data to get descriptive variables can be accessed or generated readily.
* Hedonic pricing model is versatile, it can be adapted to other market goods and services and environmental qualities (CFITeam 2022).

## **Neural Network Regression**

Regression ANNs are used to predict an output variable as a function of the inputs. In ANN regression problems the dependent variable should be numeric. Neural networks consist of nodes and each node has corresponding activation function that defines the output of the node with the set of inputs. By changing the last activation function the NN classification problem can be changed to a regression model and the appropriate library to build the NN ML is the ‘Keras’ library (Ajay Ohri 2022).

To ensure the nonlinearity, the output of neuron is mapped to different values in the NN regression and single or a range of parameters can be selected for predicting output using NN regression. The outputs of neural network related to each other, and these neurons help in predicting future values as well as mapping a relationship between dependent and independent variables. A multilayer perceptron (MLP) is a class of a feedforward artificial neural network (ANN). MLP models are the most basic deep neural network and is composed of a series of fully connected layers (Ajay Ohri 2022).

## **ANOVA Test**

Analysis of Variance (ANOVA) is a statistical test which helps to find out whether the difference between different groups of data is statistically significant. It also helps us to understand the relationship between the dependent and independent variables (Kenton 2022). One-way ANOVA is the basic form of ANOVA and there are other variations which can be used in different situations.

* Two-way ANOVA
* Factorial ANOVA
* Welch’s F-test ANOVA
* Ranked ANOVA
* Games-Howell pairwise tests.

ANOVA works by evaluating the levels of variance within the groups of data through samples taken from each group. If there is lot of difference in the variance of the data groups, then there is chance for the mean of the selected group to be different from the population. ANOVA also deals with the sample size and the difference between the sample means. All these components are linked into a F value, which can then be analyzed to give a probability (p-value) of whether the differences between sample groups are statistically significant or not. One-way ANOVA compares the effects of independent variable on dependent variables and two-way ANOVA does the same with multiple independent variables (Qualitrics 2022b).

# **Chapter 6**

# **Evaluation Methods**

This section explains the different evaluation methods used in the project to analyze the performance of the statistical as well as ML models. The performance metrics used to evaluate the ML models were:

## **Root Mean Squared Value**

Root Mean Squared Error is the square root of the average of the squared difference between the actual value and the predicted value of the regression model (Bajaj Aayush 2022).

RMSE =

Some of the key advantages of RMSE are:

* It preserves the differentiable property of MSE.
* It manages the penalization of smaller errors made by MSE by square rooting the value.
* Error interpretation can be done easily since the scale is same as the random variable.
* It is less prone to outliers (Bajaj Aayush 2022).

## **Mean Absolute Value**

Mean Absolute Error is the average of the difference between the ground truth and the predicted values (Bajaj Aayush 2022).

MAE =

Key information related to MAE are:

* MAE is more robust towards outliers than MAE, since it doesn’t amplify errors.
* It gives a measure of how far the predictions were from the actual output.
* Since MAE uses absolute value of the residual, it doesn’t give us an idea of the direction of the error, i.e., whether the model is under-predicting or over-predicting the data.
* MAE is non-differentiable as opposed to MSE, which is differentiable.
* Error interpretation perfectly aligns with the original degree of the variable (Bajaj Aayush 2022).

## **Mean Squared Error**

Mean squared error is the most popular metric used for regression problems. It calculates the average of the squared difference between the actual value and the predicted value of the regression model (Bajaj Aayush 2022).

MSE =

Some key information related to MSE are:

* MSE can be optimized better.
* It penalizes even small errors by squaring them, which leads to an overestimation of how bad the model is.
* Error interpretation should be done by considering the squaring factor(scale).
* It is more prone to outliers than other metrics.
* The value of MSE is always positive.
* MSE value close to zero will represent better quality of the regression model (Bajaj Aayush 2022).

## **R-Squared Error**

R-Squarederror works as a post metric, which means it is a metric that is calculated using other metrics. It explains how much variation in actual variable is being explained by the variation regression line. It is calculated using the sum of squared errors (Bajaj Aayush 2022).

Variance of y, SE (y

Percentage of variation described the regression line:

The percentage of variation described the regression line: 1-

R- Squared error = 1-

Key points related to R squared error are:

* If the sum of Squared Error of the regression line is small then R² will be close to 1 which is the ideal value. This means that the regression line was able to capture 100% of the variance in the target variable.
* On the contrary, if the sum of squared error of the regression line is high then the R² will be close to 0. This means that the regression line was not able to capture any variance in the target variable.
* The range of R² is (-∞,1) because the ratio of squared errors of the regression line and mean can exceed the value 1 if the squared error of regression line is too high i.e., if greater than squared error of the mean (Bajaj Aayush 2022).
* R-Squared is also known as the standardized version of MSE(Bajaj Aayush 2022).

The evaluation metrices used for the statistical models were:

## **Coefficient Std Error**

The standard deviation of an estimate is called the standard error. The standard error of the coefficient determines how accurately the model estimates the coefficient’s unknown value. The coefficient standard error is always positive (Vineet Jaiswal 2018).

## **Coefficient-t value**

t value = estimate/std error

The higher the t value will be helpful in analysis as this would indicate if we could reject the null hypothesis and it is used to determine the p value (Vineet Jaiswal 2018).

## **Coefficient Pr(>|t|)**

Individual p value for each variable to accept or reject the null hypothesis. Lower the p value allow us to reject null hypothesis (Vineet Jaiswal 2018).

## **Residual Standard Error**

It is the average error of the model, how well the model is predicting the data on average (Vineet Jaiswal 2018).

## **Multiple R-squared and Adjusted R-squared**

It is always between 0 to 1, higher values indicate better percentage of variation in the dependent variable that is explained by variation in the explanatory variables. This is used to estimate how good the model can explain the variance and when the number of variables is increased, then the values of R-squared will also increase (Vineet Jaiswal 2018).

## **F-statistic**

F-statistic shows the relationship between predictor and response variables. Higher the value will give more reasons to reject null hypothesis (Vineet Jaiswal 2018).

## **p-value**

The overall p-value based on the F-statistic. If the value is less than 0.05, then the overall model is significant (Vineet Jaiswal 2018).

# **Chapter 7**

# **Results And Discussions**

Statistical analyses, ML models and visualizations were implemented to explore the patterns in the data and to analyze the attributes contributing to the property prices. This section describes the outcome of the analyses performed.

## **Visualizations**

Matplotlib, Seaborn and Plotly libraries were used to analyze the underlying patterns of the data. Univariate, bivariate, and multivariate visualizations were generated for better understanding and to explore how the variables are related to each other.

Figure.2 - 11 shows the information about the distribution of observations on each variable in the dataset. Count plot was used to analyze the frequency of observations in the variable and log of the price variable is chosen for better visualization and understanding. This is because the price variable is having higher values, and this makes the plot difficult to interpret. From the figures it is evident that, there are outliers in some of the variables. Figure. 2 and 3 reveals that, some of the properties were sold at higher prices and others at very lower prices and most of the sold properties are at Leinster province and least number of properties are sold at Ulster. Figure. 4 and 6 shows that almost half of the observations account for Dublin. So, most of the houses were sold in Dublin than other counties. From Figure. 5 and 8, it is evident that the property selling was higher in the year 2019 and least in 2012 and most of the houses are being sold in the month of December. Another observation from the figures is that most sold prices were not full market price and was VAT inclusive. The univariate figures also provide information about the size and type of the sold properties. As of Figure 10 and 11 most of the sold properties were second hand than new houses. And the highest sold houses were with size greater than or equal to 38 Sqm and least sold houses were with size less than 38 Sqm.

Figure 7.1. 1 Univariate for Price *Chart, histogram

Description automatically generated*

Chart, bar chart

Description automatically generated

Figure 7.1. Univariate for Province

A picture containing logo

Description automatically generated

Figure 7.1. Univariate for Location

Chart, bar chart

Description automatically generated

Figure 7.1. Univariate for Month

Chart, histogram

Description automatically generated Figure 7.1. 5 Univariate for County

A picture containing chart

Description automatically generated

Figure 7.1. Univariate for VAT\_exclusive

Chart, bar chart

Description automatically generated

Figure 7.1. Univariate for Year

Shape

Description automatically generated with low confidence

Figure 7.1. Univariate for FMP

**Chart, bar chart

Description automatically generated**

Figure 7.1. Univariate for property size

**A picture containing chart

Description automatically generated** Figure 7.1. Univariate for property description

Figure.12 - 19 displays the relationship between two variables in the dataset. Line plots, count plots and box plots were generated to analyze the relationship between the variables. Figure 12 shows that the house price shows a decrease in the year 2013 and then shows significant increase. This indicates that the house prices have been increased considerably from about 2015 to 2022. The count plots reveals that higher number of properties were sold outside of Dublin but, when the whole observations are considered, sold properties are higher in Dublin than other counties and the secondhand properties are sold higher than the new properties. This may be because of the higher living conditions and job opportunities in Dublin. Additionally, one important data revealed from the count plots was that the number of sold houses are decreasing each year. This may be a result of the higher house prices. The box plots show the outliers in each variable, the IQR of each data and the spread of data points. Figures 15, 16, 17 and 19 displays the same information as the count plot and it shows that most of the sold properties in County Dublin and most of them were second hand. Likewise, most of the sold houses were with size greater than or equal to 38 Sqm and least sold houses were with size less than 38 Sqm.

Chart, line chart

Description automatically generated

Figure 7.1. Year-price bivariate

Chart, bar chart

Description automatically generated

Figure 7.1. Year-location bivariate

Chart, bar chart

Description automatically generated

Figure 7.1. Year-property description bivariate

**Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated**Figure 7.1. location-price bivariate

Figure 7.1. Province-price bivariate

**Chart, box and whisker chart

Description automatically generated**

Figure 7.1. Property description-price bivariate

**Chart, bar chart

Description automatically generated**

**Chart, box and whisker chart

Description automatically generated**Figure 7.1. Price-property size bivariate

Figure 7.1. location-property size bivariate

Multivariate plots were generated to understand the relationship among several variables in the data. Line plots and box plots were implemented for this purpose. Line plots were used to analyze the data with respect to the year frame and box plots to analyze the spread, IQR and outliers in the data. Figures 20, 21 and 22 shows that the house price has been increased significantly over the years in each county and province. It also shows that the increase in house price is very higher in Dublin than other counties. The box plots in figures 23, 24 and 25 reveals that there are outliers in the data and the highest sold houses are second hand and had a size of greater than or equal to 38 Sqm.

**Chart, line chart

Description automatically generated**

Figure 7.1. Multivariate1 **Chart, line chart

Description automatically generated**

Figure 7.1. Multivariate2

Chart, line chart

Description automatically generated

Figure 7.1. Multivariate3

**Chart, box and whisker chart

Description automatically generated**

Figure 7.1. Multivariate4

**Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated** Figure 7.1. Multivariate5

Figure 7.1. Multivariate

Further plots were made to understand the data for later analysis. The maximum price of sold property in each county were plotted. It is revealed that, Dublin has the sold properties with highest prices around 182 million Euro and the least sold price is in county Offaly with 1.4 million Euro.

Shape, rectangle

Description automatically generated

Figure 7.1. Maximum property prices in counties.

Similarly, the minimum sold prices in each county is plotted and it shows that, county Cork has the minimum property price of 5000 Euro and county Carlow has the minimum property price of 7000 Euro.

Chart

Description automatically generated

Figure 7.1. Minimum property prices in counties

To understand the relation between the size and type of the property and price, a bar chart is plotted between the ‘property\_size\_description’ variable and the median of the prices and ‘property\_description’ variable and median of the prices. From that, it is understood that the houses with greater than 125 square meters are having higher prices than the others and new houses are having higher prices than the second-hand houses.

Logo

Description automatically generated

Figure 7.1. Property sizes and median property prices

Icon

Description automatically generated

Figure 7.1. Property types and median property prices

Map

Description automatically generated

Figure 7.1. Property prices and median property prices in counties

From the above choropleth, it is evident that property prices in Dublin are far higher than other counties. The counties Roscommon, Longford, Offlay and Leitrim are having considerably lower property prices when compared to Dublin. So, to analyze the price of properties inside Dublin a bar chart and choropleth is plot between the Dublin post codes and price. It shown that, within Dublin, the post code area Dublin 6 has the highest property prices and Dublin 10 has the lowest property prices. This implies that the post codes are also having an impact in property prices.

Diagram

Description automatically generated

Figure 7.1. Postal codes and median property prices

## **Statistical Analysis**

For statistical analyses, a random sample of 500 observations are selected from the total population to get an accurate plots and results for ANOVA. The number of samples chosen was based on an online calculation that considered 95% confidence level and 5% margin of error (Qualitrics 2022a). For MLR and SLR, the entire dataset is used and even though the assumptions are not fully satisfied, I have generated the MLR and SLR models to interpret the coefficients and to estimate the significant variables. ANOVA tests were performed as a part of initial analysis to understand the relationship between the dependent and independent variables. Price variable is chosen as the dependent variable and log of the price variable was applied to meet the assumptions of the tests.

### **ANOVA Test**

The assumptions to be met for two-way ANOVA are as follows:

1. Observations should be sampled independently.
2. The variance of data in the different groups should be the same.
3. Each sample should be normally distributed.

A two-way model ANOVA is generated excluding the property size variable since it has many missing values.

Chart, scatter chart

Description automatically generatedChart, line chart

Description automatically generated

Figure 7.1. Residuals Vs Fitted values Plot

Figure 7.1. Probability Plot

From the above figures it is implied that there is constant variance for the data points, and more than half of the data points are following the normality line in the probability plot. So, the model is generated to predict the significance of the variables. From figure 34, it is revealed that that all the p values are less than 0.05 which shows that the variables are statistically significant.

Table

Description automatically generated

Figure 7.1. ANOVA results for full model

Interpretation of the model as per *Figure. 34.* and *Figure. 35.*

|  |  |
| --- | --- |
| First Hypothesis:  H0: μ2010 = μ2011 = μ2012 = μ2013 = μ2014= μ2015 = μ2016 = μ2017 = μ2018 = μ2019= μ2020 = μ2021 = μ2022 (All means are equal)  Where μi is the sample mean for year i.  H1: Not all the means of the sample population is same. (at least on mean is different so, there is an effect with year on price)  P-value = 6.6x< 0.05, hence, reject H0.  Inference: Year is statistically significant; it is indicated that a change in year will affect the price or log price significantly. | Second Hypothesis:  H0: μD = μOD  H1: μD ≠ μOD  Where:  μD is the sample population mean for the log price when the property location is Dublin,  μOD is the sample population mean for the log price when the property location is outside Dublin.  P-value = 3.2x< 0.05, hence, reject H0.  Inference: Location is statistically significant. So,  changing the location will impact the price or log price significantly. |
| Third Hypothesis:  H0: μnew = μused  H1: μnew ≠ μused  Where:  μnew is the sample population mean for the log price when the property is new,  μused is the sample population mean for the log price when the property second-hand.  P-value = 1.07x < 0.05, hence, reject H0.  Inference: Property type is statistically significant; it shows that changing the type will impact the price or log price significantly. | Fourth Hypothesis:  H0: = = =  H1: Not all the means of the sample population is same. (at least on mean is different so, there is an effect with province on price)  P-value = = 1.0x <0.05. hence reject H0.  Inference: Province is statistically significant; it shows that changing the province will impact the price or log price significantly. |

Table 7.2.1. Two-way ANOVA interpretation

Graphical user interface, application, table

Description automatically generated

Figure 7.1. Model summary for two-way ANOVA

### **Hedonic Pricing using MLR**

The assumptions to be met for MLR are as follows:

1. There should be linear relationship between the dependent and independent variables.
2. The residuals should be normally distributed.
3. There should be no collinearity.

A MLR model is generated excluding the property size variable since it has many missing values. Square root transform is applied on the log value of price variable to deal with the outliers and make it symmetrical. However, the log price without transform is fed to the model to get accurate results. A MLR model will be considered as Hedonic pricing model when the dependent variable is price or log value of price. So here the MLR model is acting as a Hedonic model since our dependent variable is the log value A picture containing line chart

Description automatically generatedof the price of the property.

Figure 7.1. Residuals Vs Fitted values Plot

Chart, histogram

Description automatically generated Figure 7.1. Probability Plot

Chart, histogram

Description automatically generated

Figure 7.1. Histogram for MLR

From the above figures it is implied that there is constant variance for the data points, and more than half of the data points are following the normality line in the probability plot. In addition to this, the histogram is almost symmetrical, so the model is generated to predict the significance of the variables. Figure 39 shows the model summary and from that only 29 percentage of the variance in the data is being explained by the model. The standard error coefficients are positive with lower values. The p value of all variables except one variable is less than 0.05 which means the variables are significant.

Table

Description automatically generated

Figure 7.1. Model summary for MLR

Interpretation of the model coefficients as per Figure. 39.

|  |
| --- |
| y = 24.5 - 0.15 year\_2011 – 0.33 year\_2012 - 0.38 year\_2013 - 0.27 year\_2014 - 0.11 year\_2015 - 0.01 year\_2016 + 0.005 year\_2017 + 0.08 year\_2018 +0.13 year\_2019 + 0.17 year\_2020 + 0.26 year\_2021 + 0.33 year\_2022 + 0.08 province\_leinster - 0.13 province\_munster - 0.23 province\_ulster - 0.60 outside\_dublin - 0.15 second\_hand – 0.21 lat + 0.07 lon.   * 24.5 is the estimated intercept i.e., the mean average selling price of the house when all predictor variables are zero. This doesn’t make sense since it is not possible to make all predictors zero. However, we still need to keep the intercept term in the model to use it to make predictions. The intercept just doesn’t have any meaningful interpretation for this model. This can also be interpreted as, the expected price (log price) when we have 2010 as the year, Connacht as province, with type new and location Dublin. * For 1 year increase in year\_2011, the average increase in property price will be -0.15, keeping all other variables constant. * For 1 year increase in year\_2012, the average increase in property price will be -0.33, keeping all other variables constant. * For 1 year increase in year\_2013, the average increase in property price will be -0.38, keeping all other variables constant. * For 1 year increase in year\_2014, the average increase in property price will be -0.27, keeping all other variables constant. * For 1 year increase in year\_2015, the average increase in property price will be -0.11, keeping all other variables constant. * For 1 year increase in year\_2016, the average increase in property price will be -0.01, keeping all other variables constant. * For 1 year increase in year\_2017, the average increase in property price will be 0.005, keeping all other variables constant. * For 1 year increase in year\_2018, the average increase in property price will be 0.08, keeping all other variables constant. * For 1 year increase in year\_2019, the average increase in property price will be 0.13, keeping all other variables constant. * For 1 year increase in year\_2020, the average increase in property price will be 0.17, keeping all other variables constant. * For 1 year increase in year\_2021, the average increase in property price will be 0.26, keeping all other variables constant. * For 1 year increase in year\_2022, the average increase in property price will be 0.33, keeping all other variables constant. * For 1 unit increase in province\_leinster, the average increase in property price will be 0.08, keeping all other variables constant. * For 1 unit increase in province\_munster, the average increase in property price will be 0.13, keeping all other variables constant. * For 1 unit increase in province\_ulster, the average increase in property price will be -0.23, keeping all other variables constant. * For 1 unit increase in outside\_dublin, the average increase in property price will be -0.60, keeping all other variables constant. * For 1 unit increase in second\_hand, the average increase in property price will be -0.15, keeping all other variables constant. * For 1 unit increase in latitude, the average increase in property price will be -0.21, keeping all other variables constant * For 1 unit increase in longitude, the average increase in property price will be 0.07, keeping all other variables constant * The property price is having positive increase after year 2017, it shows that the price has been increased significantly after 2017.   All p-values are < 0.05, except for year\_2017 which 0.08 this implies that the variables year, location, province, and property\_description are statistically significant. |

Table 7.2.2. MLR Results interpretation

### **SLR on Dublin Data**

To analyze the impact of Post codes on property prices, a new dataset was created from the original dataset with the details of County Dublin. The assumptions to be met for SLR are as follows:

1. There should be linear relationship between the dependent and independent variable.
2. The residuals should be normally distributed.
3. There should be no relationship between the residuals and the independent variable.

A SLR model is generated using the price and the post codes. Square root transform is applied on the log value of price variable to deal with the outliers and make it symmetrical. However, the log price without transform is fed to the model to get accurate results.

Chart, line chart

Description automatically generatedChart, scatter chart

Description automatically generated

Figure 7.1. Residuals Vs Fitted values Plot

Figure 7.1. Probability Plot

Chart, histogram

Description automatically generated

Figure 7.1. Histogram for SLR

From the Figures 40 and 41 it is indicated that there is constant variance for the data points, and more than half of the data points are following the normality line in the probability plot. In addition to this, the histogram is almost symmetrical, so the model is generated to predict the significance of the variables. Figure 43 shows the model summary and from that only 19 percentage of the variance in the data is being explained by the model. The standard error coefficients are positive with lower values. The p value of all variables is less than 0.05 which means the variables are statistically significant.

Table

Description automatically generated

Figure 7.1. Model summary for SLR

Interpretation of the model coefficients as per Figure. 43.

|  |
| --- |
| y = 12.36 - 0.42 Dublin10 – 0.17 Dublin11 + 0.12 Dublin12 + 0.33 Dublin13 + 0.72 Dublin14 + 0.13 Dublin15 + 0.56 Dublin16 – 0.21 Dublin17 + 0.54 Dublin18 + 0.40 Dublin2 + 0.08 Dublin20 - 0.09 Dublin22 + 0.03 Dublin24 + 0.41Dublin3 + 0.81 Dublin4 + 0.32 Dublin5 + 0.88 Dublin6 + 0.65 Dublin6w + 0.16 Dublin7 + 0.09 Dublin8 + 0.3 Dublin9.   * 12.36 is the estimated intercept i.e., the mean average selling price of the house when all predictor variables are zero. This doesn’t make sense since it is not possible to make all predictors zero. However, we still need to keep the intercept term in the model to use it to make predictions. The intercept just doesn’t have any meaningful interpretation for this model. So, this can also be interpreted as, the expected price (log price) when we have Dublin 1 as the postcode. * For 1 unit increase in Dublin10, the average increase in property price will be -0.42, keeping all other variables constant. * For 1 unit increase in Dublin11, the average increase in property price will be -0.17, keeping all other variables constant. * For 1 unit increase in Dublin12, the average increase in property price will be 0.12, keeping all other variables constant. * For 1 unit increase in Dublin13, the average increase in property price will be 0.33, keeping all other variables constant. * For 1 unit increase in Dublin14, the average increase in property price will be 0.72, keeping all other variables constant. * For 1 unit increase in Dublin15, the average increase in property price will be 0.13, keeping all other variables constant. * For 1 unit increase in Dublin16, the average increase in property price will be 0.56, keeping all other variables constant. * For 1 unit increase in Dublin17, the average increase in property price will be -0.21, keeping all other variables constant. * For 1 unit increase in Dublin18, the average increase in property price will be 0.54, keeping all other variables constant. * For 1 unit increase in Dublin2, the average increase in property price will be 0.40, keeping all other variables constant. * For 1 unit increase in Dublin20, the average increase in property price will be 0.08, keeping all other variables constant. * For 1 unit increase in Dublin22, the average increase in property price will be -0.09, keeping all other variables constant. * For 1 unit increase in Dublin24, the average increase in property price will be 0.04, keeping all other variables constant. * For 1 unit increase in Dublin3, the average increase in property price will be 0.41, keeping all other variables constant. * For 1 unit increase in Dublin4, the average increase in property price will be 0.81, keeping all other variables constant. * For 1 unit increase in Dublin5, the average increase in property price will be 0.32, keeping all other variables constant. * For 1 unit increase in Dublin6, the average increase in property price will be 0.88, keeping all other variables constant. * For 1 unit increase in Dublin6w, the average increase in property price will be 0.65, keeping all other variables constant. * For 1 unit increase in Dublin7, the average increase in property price will be 0.17, keeping all other variables constant. * For 1 unit increase in Dublin8, the average increase in property price will be 0.09, keeping all other variables constant. * For 1 unit increase in Dublin9, the average increase in property price will be 0.03, keeping all other variables constant.   All p-values are < 0.05 this implies that the post codes statistically significant and influences the property prices. |

Table 7.2.3. SLR Results interpretation

### **MLR on Dublin Data**

The assumptions to be met for MLR are as follows:

1. There should be linear relationship between the dependent and independent variables.
2. The residuals should be normally distributed.
3. There should be no collinearity.

A MLR model is generated excluding the property size variable since it has many missing values. Square root transform is applied on the log value of price variable to deal with the outliers and make it symmetrical. However, the log price without transform is fed to the model to get accurate results.

Chart, scatter chart

Description automatically generatedChart, line chart

Description automatically generated

Figure 7.1. 43 Residuals Vs Fitted values Plot

Figure 7.1. 44 Probability Plot

Chart, histogram

Description automatically generated

Figure 7.1. Histogram

The figures 44 and 45 shows that there is constant variance for the data points, and more than half of the data points are following the normality line in the probability plot. In addition to this, the histogram is almost symmetrical, so the model is generated to predict the significance of the variables. Figure 39 shows the model summary and from that only 29 percentage of the variance in the data is being explained by the model. The standard error coefficients are positive with lower values. The p value of all variables except one variable is less than 0.05 which means the variables are statistically significant.

Table

Description automatically generated

Graphical user interface, application, table, Excel

Description automatically generated

Figure 7.1. Model Summary for MLR

Interpretation of the model coefficients as per Figure. 47.

|  |
| --- |
| y = 0.0042 - 0.17 year\_2011 – 0.36 year\_2012 - 0.31year\_2013 + 0.015 year\_2014 + 0.04 year\_2015 + 0.15 year\_2016 + 0.24year\_2017 + 0.31 year\_2018 +0.34 year\_2019 + 0.35 year\_2020 + 0.42 year\_2021 + 0.43 year\_2022 - 0.44 Dublin10 – 0.19 Dublin11 + 0.09 Dublin12 + 0.28 Dublin13 + 0.69 Dublin14 + 0.08 Dublin15 + 0.52 Dublin16 – 0.22 Dublin17 + 0.51 Dublin18 + 0.4 Dublin2 + 0.05 Dublin20 - 0.15 Dublin22 - 0.03 Dublin24 + 0.39 Dublin3 + 0.8 Dublin4 + 0.3 Dublin5 + 0.86 Dublin6 + 0.77 Dublin6w + 0.14 Dublin7 + 0.06 Dublin8 + 0.27 Dublin9 + 0.02 second\_hand + 0.22 lat – 0.03 lon.   * 0. 0042 is the estimated intercept i.e., the mean average selling price of the house when all predictor variables are zero. This doesn’t make sense since it is not possible to make all predictors zero. However, we still need to keep the intercept term in the model to use it to make predictions. The intercept just doesn’t have any meaningful interpretation for this model. So, this can also be interpreted as, the expected price (log price) when we have 2010 as the year, Dublin 1 as postcode, with type new property. * For 1 year increase in year\_2011, the average increase in property price will be -0.17, keeping all other variables constant. * For 1 year increase in year\_2012, the average increase in property price will be -0.36, keeping all other variables constant. * For 1 year increase in year\_2013, the average increase in property price will be -0.31, keeping all other variables constant. * For 1 year increase in year\_2014, the average increase in property price will be 0.015, keeping all other variables constant. * For 1 year increase in year\_2015, the average increase in property price will be 0.04, keeping all other variables constant. * For 1 year increase in year\_2016, the average increase in property price will be 0.15, keeping all other variables constant. * For 1 year increase in year\_2017, the average increase in property price will be 0.25, keeping all other variables constant. * For 1 year increase in year\_2018, the average increase in property price will be 0.31, keeping all other variables constant. * For 1 year increase in year\_2019, the average increase in property price will be 0.34, keeping all other variables constant. * For 1 year increase in year\_2020, the average increase in property price will be 0.35, keeping all other variables constant. * For 1 year increase in year\_2021, the average increase in property price will be 0.42, keeping all other variables constant. * For 1 year increase in year\_2022, the average increase in property price will be 0.43, keeping all other variables constant. * For 1 unit increase in Dublin10, the average increase in property price will be -0.45, keeping all other variables constant. * For 1 unit increase in Dublin11, the average increase in property price will be -0.2, keeping all other variables constant. * For 1 unit increase in Dublin12, the average increase in property price will be 0.09, keeping all other variables constant. * For 1 unit increase in Dublin13, the average increase in property price will be 0.28, keeping all other variables constant. * For 1 unit increase in Dublin14, the average increase in property price will be 0.69, keeping all other variables constant. * For 1 unit increase in Dublin15, the average increase in property price will be 0.13, keeping all other variables constant. * For 1 unit increase in Dublin16, the average increase in property price will be 0.08, keeping all other variables constant. * For 1 unit increase in Dublin17, the average increase in property price will be -0.23, keeping all other variables constant. * For 1 unit increase in Dublin18, the average increase in property price will be 0.51, keeping all other variables constant. * For 1 unit increase in Dublin2, the average increase in property price will be 0.4, keeping all other variables constant. * For 1 unit increase in Dublin20, the average increase in property price will be 0.05, keeping all other variables constant. * For 1 unit increase in Dublin22, the average increase in property price will be -0.16, keeping all other variables constant. * For 1 unit increase in Dublin24, the average increase in property price will be -0.03, keeping all other variables constant. * For 1 unit increase in Dublin3, the average increase in property price will be 0.39, keeping all other variables constant. * For 1 unit increase in Dublin4, the average increase in property price will be 0.8, keeping all other variables constant. * For 1 unit increase in Dublin5, the average increase in property price will be 0.3, keeping all other variables constant. * For 1 unit increase in Dublin6, the average increase in property price will be 0.86, keeping all other variables constant. * For 1 unit increase in Dublin6w, the average increase in property price will be 0.77, keeping all other variables constant. * For 1 unit increase in Dublin7, the average increase in property price will be 0.14, keeping all other variables constant. * For 1 unit increase in Dublin8, the average increase in property price will be 0.06, keeping all other variables constant. * For 1 unit increase in Dublin9, the average increase in property price will be 0.27, keeping all other variables constant. * For 1 unit increase in second-hand, the average increase in property price will be 0.02, keeping all other variables constant. * For 1 unit increase in latitude, the average increase in property price will be 0.22, keeping all other variables constant * For 1 unit increase in longitude, the average increase in property price will be -0.03, keeping all other variables constant   Almost all the p-values are < 0.05 this implies that the post codes statistically significant and influences the property prices. |

Table 7.2.4. MLR Results interpretation for Dublin

## **ML Models**

Machine learning algorithms were used to analyze and predict the property prices. The dataset had 515792 rows × 15 columns. The dataset was divided into training and test data with 80 percentage training and 20 percentage of test data. The training data consists of 412633 observations and the test data has 103159 observations. The ML models are mainly classified as three supervised, semi supervised, and reinforced models. Here for predicting the house prices supervised ML algorithms are considered since the data has labels.

Python and Jupyter notebook were used for implementing the models using the scikit and keras library. The ML modelling consists of data ingestion, data cleaning, exploratory data analysis, feature engineering and finally machine learning. The data is cleaned prior to the modelling and the NaN values and insignificant columns are removed to perform the modelling. Since the dependent variable is numeric, ML regression algorithms were used to analyze the data. Dummy variables were created for the categorical variables in the data to make it appropriate to fed to the model. A correlation matrix was plotted to understand the relationship of the variables and to identify the correlations in the data. There were no strong correlations in the data and hence it is acceptable to continue with modelling without altering any variables.

Chart, histogram, treemap chart

Description automatically generated

Figure 7.3. Correlation Matrix

The standard development cycle for ML is followed in the project to develop and evaluate the models. Standard scaler was chosen to preprocess and scale the data prior to modelling. LR, RF, DCT, XG-Boost, SVR and NN algorithms were utilized for building the regression models. At first the models were implemented using the scaled data and the performance metrices were evaluated and after that data without scaling is fed to the models to analyze the performance. It is found that the scaled data gives less performance for some models compared to unscaled data and hence some models were implemented without data scaling. Hence, SVM regressor was performed with scaling and other algorithms were implemented without scaling data. RMSE, MSE, MAE and r squared error were utilized for evaluating the performance of the models to identify the best fit model. The RMSE, MSE and MAE values were found to be large values, this is because we are dealing with real and economic data and there is chance for the performance metrics of economic data to be high values. Hence r squared value for each model was compared to identify the best model. The evaluation metrices for the ML models were as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| **Linear Regression Model** | | | |
| **RMSE** | **MSE** | **MAE** | **R Squared** |
| 803454.106 | 645538500662.903 | 135139.365 | 0.023 |

Table 7.3. Performance metrices of Linear Regression

|  |  |  |  |
| --- | --- | --- | --- |
| **Decision Tree****Regression Model** | | | |
| **RMSE** | **MSE** | **MAE** | **R Squared** |
| 871447.338 | 759420462634.51 | 137785.41 | 0.00 |

Table 7.3. Performance metrices of Decision Tree Regression

|  |  |  |  |
| --- | --- | --- | --- |
| **Support Vector Regression Model** | | | |
| **RMSE** | **MSE** | **MAE** | **R Squared** |
| 0.786 | 0.617 | 0.629 | 0.452 |

Table 7.3. Performance metrices of Support Vector Regression

|  |  |  |  |
| --- | --- | --- | --- |
| **XG-Boost****Regression Model** | | | |
| **RMSE** | **MSE** | **MAE** | **R Squared** |
| 653997.82 | 427713148859.75 | 134917.819 | 0.010158 |

Table 7.3. Performance metrices of XG-Boost Regression

|  |  |  |  |
| --- | --- | --- | --- |
| **Random Forest Regression Model** | | | |
| **RMSE** | **MSE** | **MAE** | **R Squared** |
| 1094282.69 | 1197454605589.61 | 140099.601 | 0.004 |

Table 7.3. Performance metrices of Random Forest Regression

|  |  |  |  |
| --- | --- | --- | --- |
| **NN Regression Model** | | | |
| **RMSE** | **MSE** | **MAE** | **R Squared** |
| 728057.406 | 530067586324.427 | 153320.551 | 0.022 |

Table 7.3. Performance metrices of NN Regression

The Tables 7.3.1 to 7.3.6 shows the performance metrices of the different ML algorithms employed and R squared and RMSE was chosen as the criteria to decide the better model. This is also because, the R squared can tell how well a model can predict the response variable in percentage and it captures the fraction of variance of actual values captured by the regression model. It also tends to give a better picture of the quality of the regression model. On the other hand, RMSE can be dealt with high error values and can tell the average deviation between the predicted house price made by the model and the actual house price. By comparing the metrices, it is revealed that the SV regression has the best R squared and the DCT has the lowest R squared value. Likewise, SV regression and XG-Boost has the lowest RMSE values. Hence, SV regression and XG-Boost algorithms seems to be accurate for property price analysis.

The actual and the predicted values of each model has been plotted to understand how well the model predicts apart from the results of the performance metrices. Only 50 data points were chosen from the data for plotting to understand the data more precisely. And it shows that the XG-Boost algorithm’s prediction is better with the available data compared to other models. The DC regressor seems to be the poorly predicting model among the six algorithms used.

Chart, line chart, histogram

Description automatically generated

Figure 7.3. Actual Vs Predicted Values in Linear Regression

Chart

Description automatically generated

Figure 7.3. Actual Vs Predicted Values in DC Regression

Chart, line chart, histogram

Description automatically generated

Figure 7.3. Actual Vs Predicted Values in SV Regression

Chart, line chart, histogram

Description automatically generated

Figure 7.3. Actual Vs Predicted Values in XG-Boost Regression

Chart, line chart, histogram

Description automatically generated

Figure 7.3. Actual Vs Predicted Values in RF Regression

Chart, line chart, histogram

Description automatically generated

Figure 7.3. Actual Vs Predicted Values in NN Regression

K-Fold cross validation is applied on the best models by dividing the data into different groups. This is performed as the train and test data does not always have the same variation as the original data and hence it will affect the performance of the model when using the train test split method. In cross validation this problem is solved by dividing the data into multiple groups other than just two. K-fold cross validation is utilized here, which will shuffle the data and divide the data into the desired number of folds. In this method one fold will be used for testing and the other folds for training which will reduce the chances of underfitting as we are using most of the data for training, and also significantly reduces overfitting as most of the data is also being used in validation set. K-fold cross validation on the SVM regression model produced a r-squared value of 0.4 in the fourth fold, which shows the model performance can be improved by this method. The same technique was applied to XG-Boost regression also and it produced 93100 as the average RMSE value for all the folds. This also implies that by K-fold cross validation the overall performance can be improved.

# **Chapter 8**

# **Ethical Considerations**

This section describes the ethical challenges faced in the different stages of the research as a data scientist and the ethical concerns with the data. To avoid the ethical issues, the data was collected from a website which is accessible to the public for download and use and no personal information were used for analysis. The ethical challenges faced are as follows.

## **Data storage, security, and responsible data stewardship**

Even though the data does not contain any personal information, to restrict the loss of the data from the system and to avoid the misuse of it especially the API key, a folder is created in Google drive and one drive to save the details related to the dissertation. The folder is secured with passcode to avoid the misuse of it. Anti-virus packages are installed in machines to reduce the risk of hacking and virus.

## **Data hygiene and data relevance**

To add additional data to the dataset, details are obtained from genuine websites with proper referencing. The API key to get the latitude is taken from a website which follows GDPR and sends requests as HTTPS.

## **Identifying and addressing ethically harmful data bias**

The dataset has a column with postal code, and it contains the postal code of Dublin only. So, if I use that column in my modelling there may be a chance of bias towards Dublin than other places. Even though it doesn’t affect any individuals and doesn’t create any harms, but for the better modelling and to reduce bias, I may drop that column.

The ethical issues related to data are also considered in the research and they are,

## **Harms to Privacy and Security**

To avoid the issues with privacy and security, no personal information is using in the research. The data is collected from an open website, and the data is free to public for downloading and using. Reuse and download of the data are permitted with acknowledging the website and the data cannot be used in a misleading way. The data does not contain any personal information and hence there is no ethical harms related to privacy and security (Vallor et al.).

## **Harms to Fairness and Justice**

Since no personal data is being used, there is no issue related to discrimination and bias related to gender or age and hence there is no issue related to sexism, racism, and ableism. Since the data is considering the whole Ireland and the properties sold and not about people, there is no other ethical issues related to fairness and justice also (Vallor et al.).

## **Harms to Transperancy and Autonomy**

To avoid the transparency issues, the data is collected from the free website, and anyone can visit the data without any issues. The research is not using any humans for the analysis or any personal data and hence there is no issues related to transparency and autonomy and no human rights have been violated (Vallor et al.).

# **Chapter 9**

# **Conclusion And Futurework**

Property prices in Ireland were analyzed using different statistical methods, ML algorithms and visualizations to identify the factors effecting the property prices, to predict the property prices and to analyze the relationship between the different parameters. Different plots were created using seaborn and matplotlib for the entire dataset which is downloaded from the PSRA website and ANOVA test was generated for the random sample size of 500 samples from the entire data. MLR model was utilized to generate the hedonic pricing model to determine the influence of different factors. The results from the ANOVA test and MLR model implies that the attributes ‘year’, ‘province’, ‘property\_description’, ‘latitude’, ‘longitude’ and ‘location’ has significant effect on the property prices. From the visualizations it is revealed that the property prices are higher in Dublin compared to other counties and more houses are sold in Dublin also. The post codes of the county Dublin are also having significant effect on the property prices where Dublin 14 is having higher property prices and Dublin 10 is having lower prices. The SLR model implemented for the Dublin data confirmed this and reveled that the post codes are having a positive effect on the property prices. Different ML algorithms were used to identify the accurate model for the house price prediction and the SV regression and XG-Boost was found to the accurate models based on the R squared and RMSE values while XG-Boost model was found to be predicting the prices more precisely.

As the data was limited and the access to other environmental and geographical data was not possible, more accurate results and analysis could not be done. In future, if the data is available, the prediction could be done based on that data which makes the prediction more valuable and helpful to the customers. In addition to this the proposed model has high scalability which can be used to analyze the house prices of other countries with similar population and landscapes and properties.